

Robust Cell-Load Learning with a Small Sample Set

Daniyal Amir Awan, *Student Member, IEEE*, Renato L.G. Cavalcante, *Member, IEEE*,
and Slawomir Stanczak, *Senior Member, IEEE*.

Abstract—Learning of the cell-load in radio access networks (RANs) has to be performed within a short time period. Therefore, we propose a learning framework that is robust against uncertainties resulting from the need for learning based on a relatively small training sample set. To this end, we incorporate prior knowledge about the cell-load in the learning framework. For example, an inherent property of the cell-load is that it is monotonic in downlink (data) rates. To obtain additional prior knowledge we first study the feasible rate region, i.e., the set of all vectors of user rates that can be supported by the network. We prove that the feasible rate region is compact. Moreover, we show the existence of a Lipschitz function that maps feasible rate vectors to cell-load vectors. With these results in hand, we present a learning technique that guarantees a minimum approximation error in the worst-case scenario by using prior knowledge and a small training sample set. Simulations in the network simulator NS3 demonstrate that the proposed method exhibits better robustness and accuracy than standard multivariate learning techniques, especially for small training sample sets.

Index Terms—machine learning, 5G, robust learning, optimal approximation

I. INTRODUCTION

The fifth-generation (5G) networks will be based on orthogonal frequency-division multiple access (OFDMA). Due to inter-cell interference, radio resource management (RRM) and performance optimization in these networks are challenging. In fact, many RRM problems in OFDMA-based networks, such as small-scale optimal assignment of time-frequency resource blocks and powers to users, have been shown to be NP-hard [1]. Recent research has therefore focused on the development of frameworks that capture the essence of OFDMA-based networks, while leading to a tractable problem formulation. An example of such a framework is the *non-linear load-coupling model* proposed in [2], [3], [4]. In this framework the *cell-load* at a base station is the fraction of time-frequency resource blocks that are used to support downlink data rates (henceforth simply rates). With this model, and given some power budget that can be used for transmission, one can estimate the cell-load required at each base station to support given rates.

The study in [5] shows the intuitive result that the cell-load is monotonic in rates. The interference coupling between cells implies that increasing the rates in an arbitrary cell increases the cell-load at each base station, which also increases the inter-cell interference.¹ So, it is important for a base station to have a reliable forecast of the cell-load before serving higher rate demands from its associated users. Therefore, cell-load learning can be used to make radio resource management and self-organizing-network (SON) algorithms more reliable and efficient.

Cell-load learning is also a vital part of energy saving mechanisms in radio access networks (RANs). For instance in [6], the value of the cell-load is used as an input to a simple heuristic algorithm that switches off base station antennas when the cell-load is low. Large gains in energy savings are reported with minimal effect on the cell sum throughput. The same concept can be used in the case of virtual base station formations in *cloud RANs* [7]. In these virtual systems some power-hungry components of a RAN (digital signal processors, line cards, fronthaul, etc.) are virtualized in a central location, and these components can be allocated on-demand to cells according to the cell-load. Therefore, given RAN data traffic (or rates) predictions, the corresponding cell-load forecasts can enable us to proactively manage network components for energy savings.

A. The Need for Robust Cell-Load Learning

Note that even though the load-coupling model has been shown to work sufficiently well in predicting the cell-load in some scenarios [3], [8], [9], models are only idealizations and in general they do not capture all the intricacies of dynamic wireless environments. Therefore, our objective is to directly learn the underlying function that maps user rates to cell-load values given a training sample set consisting of rate vectors and the corresponding measured cell-load vectors. To improve the learning process, we use the load-coupling model to study some salient aspects of the relationship between rates and the cell-load. We use these aspects as prior knowledge in the learning process.

Compared to the core network, the RAN data traffic is volatile and it shows irregular patterns throughout a day because of the unpredictable nature of user activity and relatively fast changes in the network topology [10]. Therefore, the underlying statistics (i.e., the joint probability distribution) of rates and the corresponding cell-load values, which are part of the so-called *environment*, can be assumed to remain constant for only a short time. This implies that a training sample set must be acquired during this short time before the environment changes, since otherwise the sample set can be rendered useless for predicting future cell-load values. However, in general, the smaller the sample set, the larger the uncertainty about the underlying phenomenon, which makes large prediction errors on unseen rates more probable.

In uncertain situations we need “robust” learning methods that provide a guaranteed worst-case performance under uncertainty. The objective of this study is to develop such a robust learning framework. Our method is optimal in the sense that it minimizes the worst-case or maximum error of approximation which is a classical robust optimization problem (see, e.g.,

¹For brevity, we assume that cells are not mutually orthogonal.

[11], [12], [13], [14]). This means that no matter how small the training sample set is, we are guaranteed the best worst-case error. Our method involves only low-complexity and stable mathematical operations and its theoretical properties are very well understood. The above mentioned optimization problem is solved by explicitly incorporating prior knowledge regarding the Lipschitz continuity of the function to be approximated. By incorporating additional prior knowledge concerning monotonicity of the function, we further reduce the worst-case error.

We point out that our framework is different to many modern conventional machine learning frameworks that target mean or average performance rather than the worst-case performance we consider in this study. The performance of many current complex learning methods, such as deep neural networks (DNNs), is often dependent on the availability of a large training (or pre-training) sample set. Including prior knowledge in these frameworks to reduce the reliance on large training sets is not easy, and it is often discouraged [15]. Even if some prior knowledge could be enforced in neural networks (as in [16]), it is theoretically unclear whether (or how) this enables neural networks to learn better. This makes DNNs ill-suited to our setting because we consider learning with very small training sample sets.

B. Related Work

The load-coupling model [2], [3], [4] is commonly used when designing networks according to the long-term evolution (LTE) standard. Recently it has also attracted attention in the context of 5G networks [17]. More specifically, the load-coupling model has been used in various optimization frameworks dealing with different aspects of network design including data offloading [5], proportional fairness [18], energy optimization [19], [20], [21], and load balancing [22]. In the context of energy savings, and by using the theory of *implicit functions* [23], the study in [21] shows that there exists a continuously differentiable function relating user associations with the base stations to the cell-load. In contrast to [21], the user association is assumed to be fixed in this study; we study the relationship between downlink rates and the cell-load and we incorporate this prior knowledge in our learning framework. Previous studies dealing with cell-load estimation, for instance, in the context of data offloading [5] and maximizing the scaling-up factor of traffic demand [24], have used load coupling model driven methods that require information about channel gains, powers, etc.. Most of these methods employ iterative algorithms to estimate the cell-load for given downlink rates and other parameters by exploiting the fact that the cell-load is the fixed point of the *standard interference mapping* [25] that is constructed using the network information. In contrast, we directly learn the underlying function that maps feasible rates, i.e., downlink data rates that can be supported by the network, to the observed cell-load in the network using a sample training set and prior knowledge. Our framework, therefore, does not require information about powers, channels, etc..

Inclusion of prior knowledge in the form of constraints, known properties, and logic has also been widely used in

other areas, such as optimal control [26], [27], to deal with uncertainty. However, incorporating prior knowledge in machine learning algorithms for multivariate data² with arbitrary dimensions is difficult, and most of the well-known algorithms either do not preserve the “shape” (i.e., known properties such as monotonicity, continuity, etc.) of the underlying function or they become too complex for high-dimensional data [28]. An inherent property of the cell-load is that it is monotonic in rates. The study in [29] shows that monotonicity is difficult to incorporate in popular online learning methods even in the case of univariate data. In [28] the author proposes a shape preserving multivariate approximation of scalar monotonic functions that are also *Lipschitz*. The author shows that Lipschitz continuity of the function to be approximated allows for computing tight upper and lower bounds on the function values. Using these bounds one can obtain an optimal solution in the sense that this solution minimizes a worst-case error of approximation [11], [12], [13]. Furthermore, the approximation preserves both the monotonicity and the Lipschitz continuity of the underlying function.

C. Our Contribution

This study deals with the problem of learning cell-load in RANs as a function of downlink rates given a relatively small training sample set. The assumption of small training sample sets is crucial because modern RAN networks do not permit a long observation and sample acquisition period (see Section I-A). To cope with this limitation, we propose a robust learning framework that guarantees a minimum worst-case error of approximation. To achieve robustness, we incorporate prior knowledge about the cell-load and its relationship with rates. We show that the incorporation of prior knowledge enables us to provide explicit tight bounds that cannot be achieved by using a sample set alone, no matter how large the sample set is.

In the following we summarize the main contributions of this study.

- 1) We study the feasible rate region which is defined as the set of all rates that can be supported by the network. In the conference version of this study [30] we stated without proof that the feasible rate region is compact. In this work we provide a formal proof for this assertion along with some other related results.
- 2) In particular, we show that there exists a function that maps rates to the cell-load and that this function is monotonic and Lipschitz continuous over the feasible rate region.
- 3) We use the prior knowledge developed in 1) and 2) to perform robust learning of the cell-load by using the framework of *minimax approximation* [11], [12], [13]. Note that, this technique cannot be directly used without the prior knowledge above.
- 4) In contrast to [28], where the main concern is to preserve the monotonicity, we show theoretically and by

²Multivariate data in this context means that the input argument (or domain) of the function to be approximated has an arbitrary dimension.

experiments that including the prior knowledge regarding monotonicity results in reduced uncertainty.

- 5) Our machine learning framework does not require network information such as powers and channel gains in contrast to traditional cell-load approximation methods. The guaranteed performance of our framework with small sample sets makes it suitable in such scenarios where other learning frameworks such as DNNs cannot be applied.
- 6) In contrast to the conference version, we perform simulations in the network simulator NS3 to demonstrate the performance of the algorithm in a realistic cellular wireless network. We compare our framework with standard multivariate learning techniques and show that our method outperforms these standard techniques for small sample sizes.

D. Overview

The remainder of this study is organized as follows. Section II provides the mathematical background and results that are used throughout the study. Section III presents the non-linear load coupling model. In Section IV we provide our results on the feasible rate region. In Section V we discuss the robust optimization problem for cell-load learning along with some more related results. Section VI deals with the implementation of the cell-load learning framework developed in this study in a wireless network. Finally, in Section VII, empirical analysis is performed by simulations in the network simulator (NS3).

II. MATHEMATICAL BACKGROUND

Throughout this study \mathbb{R} , $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{> 0}$ denote the sets of reals, non-negative reals, and positive reals, respectively. We denote by $\|\cdot\|$ and $\|\cdot\|_{\infty}$ the usual Euclidean norm and l_{∞} norm in \mathbb{R}^m , respectively. The sets of non-negative integers and natural numbers are denoted by $\mathbb{Z}_{\geq 0}$ and $\mathbb{N} := \mathbb{Z}_{\geq 0} \setminus \{0\}$, respectively. We define $\bar{N}_1, \bar{N}_2 := \{N_1, N_1 + 1, N_1 + 2, \dots, N_2\}$, $N_1, N_2 \in \mathbb{Z}_{\geq 0}$ with $N_1 \leq N_2$. We denote by $(\mathbf{x})_+$ the operation $\max\{\mathbf{x}, \mathbf{0}\}$ for a vector $\mathbf{x} \in \mathbb{R}^N$, where the max is taken component-wise and $\mathbf{0}$ is the all-zero vector. For two vectors \mathbf{x} and \mathbf{y} , the inequality $\mathbf{x} \leq \mathbf{y}$ should be understood component-wise.

Let \mathcal{S} be a normed vector space equipped with a norm $\|\cdot\|_{\mathcal{S}}$ and its induced metric $d_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}_{\geq 0} : (\mathbf{s}_o, \mathbf{s}) \mapsto \|\mathbf{s}_o - \mathbf{s}\|_{\mathcal{S}}$. We denote by $\mathcal{B}_{\mathcal{S}}(\mathbf{s}_o, \delta) := \{\mathbf{s} \in \mathcal{S} \mid \|\mathbf{s} - \mathbf{s}_o\|_{\mathcal{S}} < \delta\}$ the open-ball of radius $\delta > 0$ centered at $\mathbf{s}_o \in \mathcal{S}$. A sequence $(\mathbf{s}_n)_{n \in \mathbb{N}} \subset \mathcal{S}$ is said to converge (in norm) to $\mathbf{s} \in \mathcal{S}$ if $\|\mathbf{s}_n - \mathbf{s}\|_{\mathcal{S}} \rightarrow 0$ [31, Page 26].

We now define the concepts of *boundedness*, *closedness*, and *compactness* that we use throughout this study.

Definition 1 (*Boundedness, Closedness, and Compactness*). [31, Chapter 2] Consider a set \mathcal{K} in the normed space $(\mathcal{S}, \|\cdot\|_{\mathcal{S}})$.

- a). *Boundedness*: \mathcal{K} is bounded if $(\exists L \geq 0) (\forall \mathbf{k} \in \mathcal{K}) \|\mathbf{k}\|_{\mathcal{S}} \leq L$.
- b). *Closedness*: \mathcal{K} is closed if and only if every convergent sequence $(\mathbf{k}_n)_{n \in \mathbb{N}} \subset \mathcal{K}$ has a limit in \mathcal{K} .

- c). *Compactness*: \mathcal{K} is compact if every sequence $(\mathbf{k}_n)_{n \in \mathbb{N}} \subset \mathcal{K}$ has a convergent subsequence with a limit in \mathcal{K} .

In this study we consider the space $C(\mathcal{X}, \mathcal{Y})$ of vector-valued continuous functions mapping $\mathcal{X} \subset \mathbb{R}_{> 0}^N$ to $\mathcal{Y} \subset \mathbb{R}_{\geq 0}^M$. For a function $\mathbf{g} \in C(\mathcal{X}, \mathcal{Y})$ its i th component ($i \in \overline{1, M}$) $g_i : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ is a scalar continuous function. We equip $C(\mathcal{X}, \mathcal{Y})$ with the uniform norm [31, Page 23]

$$\|\mathbf{g}\|_{C(\mathcal{X})} = \sup_{\mathbf{x} \in \mathcal{X}} \max_{1 \leq i \leq M} g_i(\mathbf{x}). \quad (1)$$

If \mathcal{X} is compact, then the sup is attained according to the *extreme value theorem* [32] because the max operation³ preserves continuity.

We now present some important concepts to keep the study as self-contained as possible. These concepts are essential to understanding our results in Section IV and in Section V.

Definition 2 (*Monotonic Function*). Let $\mathcal{X} \subset \mathbb{R}_{> 0}^N$ and $\mathcal{Y} \subset \mathbb{R}_{\geq 0}^M$. A function $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{Y}$ is said to be *monotonic* if $(\forall \mathbf{x} \in \mathcal{X}) (\forall \mathbf{y} \in \mathcal{X}) \mathbf{x} \leq \mathbf{y} \Rightarrow \mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{y})$.

Definition 3 (*L-Lipschitz function*). Consider $\mathbf{f} \in C(\mathcal{X}, \mathcal{Y})$ and a vector $\mathbf{L} := [L_1, L_2, \dots, L_M]^T \in \mathbb{R}_{\geq 0}^M$. We say that \mathbf{f} is *L-Lipschitz* on \mathcal{X} if $(\forall i \in \overline{1, M}) (\forall \mathbf{x} \in \mathcal{X}) (\forall \mathbf{y} \in \mathcal{X}) |f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq L_i \|\mathbf{x} - \mathbf{y}\|$.

Definition 4 (*L-Lipschitz-Monotonic Function*). We say that $\mathbf{f} \in C(\mathcal{X}, \mathcal{Y})$ belongs to the class of *L-Lipschitz-Monotonic Functions (LIMF)* if \mathbf{f} is monotonic and there exists $\mathbf{L} \in \mathbb{R}_{\geq 0}^M$ such that \mathbf{f} is L-Lipschitz.

Note that a function $\mathbf{f} \in C(\mathcal{X}, \mathcal{Y})$ is continuous at $\mathbf{x}_o \in \mathcal{X}$ if given $\epsilon > 0$, there exists $\delta_{\mathbf{x}_o} > 0$ such that $(\forall \mathbf{x} \in \mathcal{B}_{\mathcal{X}}(\mathbf{x}_o, \delta_{\mathbf{x}_o})) \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_o)\| < \epsilon$. The following concept of *equicontinuity* extends the concept of continuity to a collection/set $\mathcal{F} \subset C(\mathcal{X}, \mathcal{Y})$ of functions.

Definition 5 (*Equicontinuity of a Set*). [32, Chapter 7] A function set $\mathcal{F} \subset C(\mathcal{X}, \mathcal{Y})$ is called *equicontinuous* at $\mathbf{x}_o \in \mathcal{X}$ if for every $\epsilon > 0$ there exists $\delta_{\mathbf{x}_o} > 0$ such that $(\forall \mathbf{x} \in \mathcal{B}_{\mathcal{X}}(\mathbf{x}_o, \delta_{\mathbf{x}_o})) (\forall \mathbf{f} \in \mathcal{F}) \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_o)\| < \epsilon$. Furthermore, if for every $\epsilon > 0$ there exists $\delta > 0$ such that $(\forall \mathbf{x}_o \in \mathcal{X}) (\forall \mathbf{x} \in \mathcal{B}_{\mathcal{X}}(\mathbf{x}_o, \delta)) (\forall \mathbf{f} \in \mathcal{F}) \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_o)\| < \epsilon$, then \mathcal{F} is said to be (uniformly) *equicontinuous*.

Remark 1 (*Set of L-Lipschitz Functions*). An example of a (uniformly) equicontinuous subset of $C(\mathcal{X}, \mathcal{Y})$ is the set of L-Lipschitz functions, i.e., Lipschitz functions with the Lipschitz constant determined by $\mathbf{L} \in \mathbb{R}_{\geq 0}^M$ (see Definition 3). For completeness, a proof is shown in Appendix A.

The general concept of compactness in normed vector spaces has been introduced in Definition 1. The following Fact, along with Remark 2, characterizes compact subsets of $C(\mathcal{X}, \mathcal{Y})$.

Fact 1 (*Compact subsets of $C(\mathcal{X}, \mathcal{Y})$*). [33][32, Corollary 45.5] Let \mathcal{X} be compact. Then,

³The usage of max in (1) is different to the component-wise max in $\max\{\mathbf{x}, \mathbf{0}\}$. The distinction between the two usages shall be clear by the context in which they are used.

- a). *Arzelà-Ascoli's Theorem: Every bounded and equicontinuous sequence $(\mathbf{f}_n)_{n \in \mathbb{N}} \subset C(X, \mathcal{Y})$ has a convergent subsequence.*
- b). *A set $\mathcal{F} \subset C(X, \mathcal{Y})$ is compact if it is bounded, equicontinuous, and closed.*

Remark 2 (Compactness in \mathbb{R}^m and in $C(X, \mathcal{Y})$). *A subset of a finite dimensional Euclidean space is compact if and only if it is bounded and closed (see Heine-Borel Theorem [32, Theorem 27.3]). However, in $C(X, \mathcal{Y})$, equicontinuity is required in addition to boundedness and closedness for compactness.*

Finally, we present the concept of *implicit functions*, which plays an important role in our study.

Fact 2 (*Implicit function theorem*). [23] *Consider sets $X \subset \mathbb{R}^N$, $\mathcal{Y} \subset \mathbb{R}^M$, and $\mathcal{Z} \subset \mathbb{R}^M$, and a vector-valued continuous function $\mathbf{g} : \mathcal{Y} \times X \rightarrow \mathcal{Z}$. Denote by $(i \in \overline{1, M})$ $g_i : \mathcal{Y} \times X \rightarrow \mathbb{R}$ the i th component of \mathbf{g} . Now, assume that \mathbf{g} is continuously differentiable in a neighborhood $(\exists \delta_{\bar{\mathbf{x}}}, \delta_{\bar{\mathbf{y}}} > 0)$ $\mathcal{B}_{\mathcal{Y}}(\bar{\mathbf{y}}, \delta_{\bar{\mathbf{y}}}) \times \mathcal{B}_X(\bar{\mathbf{x}}, \delta_{\bar{\mathbf{x}}})$ of a point $(\bar{\mathbf{y}}, \bar{\mathbf{x}}) \in \mathcal{Y} \times X$, and that $\mathbf{g}(\bar{\mathbf{y}}, \bar{\mathbf{x}}) = \mathbf{0}$. Let the Jacobian of \mathbf{g} with respect to variables \mathbf{y} (i.e., the first argument), denoted by $\nabla_{\mathbf{y}}^{\mathbf{g}} : \mathcal{Y} \times X \rightarrow \mathbb{R}^{M \times M}$ and defined as*

$$\nabla_{\mathbf{y}}^{\mathbf{g}} := \begin{pmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdots & \frac{\partial g_1}{\partial y_M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_M}{\partial y_1} & \frac{\partial g_M}{\partial y_2} & \cdots & \frac{\partial g_M}{\partial y_M} \end{pmatrix},$$

be invertible at $(\bar{\mathbf{y}}, \bar{\mathbf{x}})$. Then, there exists a (unique and continuous) “implicit” function $\mathbf{f} : \mathcal{B}_X(\bar{\mathbf{x}}, \delta_{\bar{\mathbf{x}}}) \rightarrow \mathcal{B}_{\mathcal{Y}}(\bar{\mathbf{y}}, \delta_{\bar{\mathbf{y}}})$ such that $(\forall \mathbf{x} \in \mathcal{B}_X(\bar{\mathbf{x}}, \delta_{\bar{\mathbf{x}}}))$ $\mathbf{g}(\mathbf{f}(\mathbf{x}), \mathbf{x}) = \mathbf{0}$. Furthermore, \mathbf{f} is continuously differentiable on $\mathcal{B}_X(\bar{\mathbf{x}}, \delta_{\bar{\mathbf{x}}})$. The value of the Jacobian of \mathbf{f} is given by

$$(\forall \mathbf{x} \in \mathcal{B}_X(\bar{\mathbf{x}}, \delta_{\bar{\mathbf{x}}})) \quad \nabla_{\mathbf{x}}^{\mathbf{f}}(\mathbf{x}) = -(\nabla_{\mathbf{y}}^{\mathbf{g}}(\mathbf{f}(\mathbf{x}), \mathbf{x}))^{-1} \nabla_{\mathbf{x}}^{\mathbf{g}}(\mathbf{f}(\mathbf{x}), \mathbf{x}), \quad (2)$$

where $\nabla_{\mathbf{x}}^{\mathbf{g}} : \mathcal{Y} \times X \rightarrow \mathbb{R}^{M \times N}$ is the Jacobian of \mathbf{g} with respect to variables \mathbf{x} (i.e., the second argument) given by

$$\nabla_{\mathbf{x}}^{\mathbf{g}} := \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_M}{\partial x_1} & \frac{\partial g_M}{\partial x_2} & \cdots & \frac{\partial g_M}{\partial x_N} \end{pmatrix}.$$

III. SYSTEM MODEL

In this study we consider an urban cellular base station deployment consisting of $M \in \mathbb{N}$ base stations and $N \in \mathbb{N}$ users. We consider the downlink and we denote by $r_j \in \mathbb{R}_{>0}$ the rate of user $j \in \overline{1, N}$ per unit time. We collect the rates of all users in a vector $\mathbf{r} := [r_1, r_2, \dots, r_N]^T \in \mathbb{R}_{>0}^N$.

A. Load Coupling Model and the Feasible Rate Region

We now present the load-coupling model proposed in [2], [5], which has been shown to be sufficiently accurate in certain scenarios in practice [3], [8], [9]. This model is based on the fact that time-frequency resources available at a base station are divided into physical resource blocks to facilitate resource allocation. The cell-load (at a base station) is defined to be the fraction of available resource blocks that are allocated to

TABLE I
LIST OF VARIABLES

Description	Symbol
Number of base stations	M
Number of users	N
Set of base stations	$\mathcal{M} = \{1, 2, \dots, M\}$
Set of users	$\mathcal{N} = \{1, 2, \dots, N\}$
Set of users for base station i	$\mathcal{N}(i)$
Rate of user j	$r_j \in \mathbb{R}_{>0}$
Minimum user rate vector	$\mathbf{r}_{\min} \in \mathbb{R}_{>0}^N$
Device SNR between base station i and user j	γ_{ij}
Number of resource blocks	$R \in \mathbb{N}$
Bandwidth of each resource block	$B \in \mathbb{R}_{>0}$
Cell-load	$\rho \in \mathbb{R}_{\geq 0}^M$
Load mapping	$\mathbf{q} : \mathbb{R}_{\geq 0}^M \times \mathbb{R}_{>0}^N \rightarrow \mathbb{R}_{\geq 0}^M$
Base station transmit power	$\mathbf{p} \in \mathbb{R}_{>0}^M$
Path-loss between base station i and user j	$G_{i,j} \in \mathbb{R}_{>0}$
Space of continuous functions from X to \mathcal{Y}	$C(X, \mathcal{Y})$
Lipschitz constant	$L \in \mathbb{R}_{\geq 0}^M$
Euclidean open-ball centered at $\mathbf{x} \in X$	$\mathcal{B}_X(\mathbf{x}, \delta)$
Network coherence time	$T_{\text{net}} \in \mathbb{R}_{>0}$
Sample acquisition time	$T_{\text{obv}} \in \mathbb{R}_{>0}$
Sample average time	$T_{\text{avg}} \in \mathbb{R}_{>0}$
Sample set size	$K \in \mathbb{N}$

support the rates of the users associated with the base station. Resource blocks are allocated to users based on their rates and channel qualities given in terms of their average signal-to-interference-plus-noise ratios (SINRs). In the following we denote by $\mathcal{M} := \{1, 2, \dots, M\}$ and $\mathcal{N} := \{1, 2, \dots, N\}$ the set of base stations and users, respectively, and we denote by $\mathcal{N}(i)$ the set of users associated with base station $i \in \mathcal{M}$.

Consider the case where base station $i \in \mathcal{M}$ is serving user $j \in \mathcal{N}(i)$ and denote by $G_{i,j}$ the path-loss between base station i and user j . The load-based SINR model represents the inter-cell interference from base station $k \in \mathcal{M} \setminus i$ as the product $p_k G_{k,j} \rho_k \geq 0$, where p_k is the fixed transmit power of base station k per resource block, and where $0 < \rho_k \leq 1$ denotes the cell-load at base station k [3]. With this model in hand, the network layer (averaged) SINR of the wireless link between base station i and user j is expressed as [2], [5]

$$\gamma_{ij}(\boldsymbol{\rho}) = \frac{p_i G_{i,j}}{\sum_{k \in \mathcal{M} \setminus i} p_k G_{k,j} \rho_k + \sigma^2}, \quad (3)$$

where $\boldsymbol{\rho} := [\rho_1, \rho_2, \dots, \rho_M]^T \in \mathbb{R}_{>0}^M$ is the vector of cell-load values at all base stations in the network and where σ^2 denotes noise power. Note that the denominator in (3) provides an interpretation of the cell-load as the probability of inter-cell interference from base station k [2]. For further details of the model including its strengths and weaknesses see [2], [5]. Let $R \in \mathbb{N}$ be the total number of resource blocks available at the base station, each with bandwidth $B \in \mathbb{R}_{>0}$. Given SINR $\gamma_{ij}(\boldsymbol{\rho})$, we assume that base station i can reliably transmit at a rate $r_{ij}^s = B \log(1 + \gamma_{ij}(\boldsymbol{\rho}))$ per resource block to user j . Thus, to “support” the rate r_j , base station i has to allocate $\rho_{ij} = \frac{r_j}{r_{ij}^s}$ resource blocks to user j . Summing the resource block consumption over all $\mathcal{N}(i)$, we obtain the “cell-load” (in terms of total resource consumption) of base station $i \in \overline{1, M}$

$$\rho_i = \frac{1}{RB} \sum_{j \in \mathcal{N}(i)} \frac{r_j}{\log(1 + \gamma_{ij}(\boldsymbol{\rho}))}. \quad (4)$$

Note that, we can express the right-hand side of (4) for the entire network as a vector-valued mapping

$$\mathbf{q} : \mathbb{R}_{\geq 0}^M \times \mathbb{R}_{> 0}^N \rightarrow \mathbb{R}_{> 0}^M$$

$$(\boldsymbol{\rho}, \mathbf{r}) \mapsto \begin{bmatrix} \frac{1}{RB} \sum_{j \in \mathcal{N}(1)} \frac{r_j}{\log(1 + \gamma_{ij}(\boldsymbol{\rho}))} \\ \vdots \\ \frac{1}{RB} \sum_{j \in \mathcal{N}(M)} \frac{r_j}{\log(1 + \gamma_{ij}(\boldsymbol{\rho}))} \end{bmatrix},$$

which we refer to as the *load mapping*. Given $\bar{\mathbf{r}} \in \mathbb{R}_{> 0}^N$, it follows from (4) that the cell-load vector is the solution (if it exists) to the fixed point problem: Find $\boldsymbol{\rho}^* = [\rho_1^*, \rho_2^*, \dots, \rho_M^*]^\top \in \mathbb{R}_{\geq 0}^M$ such that:

$$\boldsymbol{\rho}^* = \mathbf{q}(\boldsymbol{\rho}^*, \bar{\mathbf{r}}). \quad (5)$$

Since the cell-load is defined as a fraction of the available resources at the base station, a rate vector is *feasible* (i.e., there are sufficient resource blocks available at all base stations to support rate of every user) if the solution (if it exists) to (5) satisfies $\boldsymbol{\rho}^* \leq \mathbf{1}$. For a given supported $\bar{\mathbf{r}} \in \mathbb{R}_{> 0}^N$, the solution to (5) can be obtained by iterative fixed point algorithms as long as the network information (path-losses, powers, user association, etc. in (4)) required by these algorithms is available. In more detail, given $\mathbf{r} \in \mathbb{R}_{> 0}^N$, the mapping $\Gamma_{\mathbf{r}} : \mathbb{R}_{\geq 0}^M \rightarrow \mathbb{R}_{> 0}^M : \boldsymbol{\rho} \mapsto \mathbf{q}(\boldsymbol{\rho}, \mathbf{r})$ is a *positive concave mapping*, so it also belongs to the class of *standard interference functions* [34], [25]. Therefore, the following holds:

Fact 3 (The unique fixed point solution). [25] Suppose the rate vector $\bar{\mathbf{r}} \in \mathbb{R}_{> 0}^N$ is feasible, then the solution set of (5) given by

$$\text{Fix}(\Gamma_{\bar{\mathbf{r}}}) := \{\boldsymbol{\rho}^* \in \mathbb{R}_{\geq 0}^M \mid \mathbf{0} < \Gamma_{\bar{\mathbf{r}}}(\boldsymbol{\rho}^*) = \boldsymbol{\rho}^* \leq \mathbf{1}\}$$

contains at most one fixed point.

As mentioned previously in Section I-C, we incorporate prior knowledge about the cell-load in our learning framework presented in Section V to ensure robust learning. To this end, Fact 4 presents an important property of the cell-load, namely its monotonicity in the rate vector:

Fact 4. [5, Theorem 2] Consider any two feasible rate vectors $\mathbf{r}^k, \mathbf{r}^j \in \mathcal{R}$ and the corresponding fixed points $\boldsymbol{\rho}^j \in \text{Fix}(\Gamma_{\mathbf{r}^j}) \neq \emptyset$ and $\boldsymbol{\rho}^k \in \text{Fix}(\Gamma_{\mathbf{r}^k}) \neq \emptyset$. Then $\mathbf{r}^j \geq \mathbf{r}^k \implies \boldsymbol{\rho}^j \geq \boldsymbol{\rho}^k$.

In the next section we define and study the *feasible* rate region, which is the set of all rates supported by the network.

IV. PROPERTIES OF THE FEASIBLE RATE REGION

In light of Fact 3 and Fact 4, and given the minimum feasible rate vector $\mathbf{r}_{\min} \in \mathbb{R}_{> 0}^N$ (e.g., corresponding to the lowest order *modulation and coding scheme* in the network) that induces the cell-load $\boldsymbol{\rho}_{\min} \in \mathbb{R}_{> 0}^M$, we are now in a position to define the feasible rate region and the set of cell-load vectors over this set.

Definition 6 (Feasible Rate Region and the Cell Load Set). The feasible rate region is defined as

$$\mathcal{R} := \{\mathbf{r} \geq \mathbf{r}_{\min} \in \mathbb{R}_{> 0}^N \mid (\exists \boldsymbol{\rho}^* \in \text{Fix}(\Gamma_{\mathbf{r}})) , \boldsymbol{\rho}_{\min} \leq \boldsymbol{\rho}^* \leq \mathbf{1}\}. \quad (6)$$

Similarly, the feasible cell-load set is given by the set of fixed points (see Fact 3)

$$\mathcal{L} := \{\boldsymbol{\rho} \in \mathbb{R}_{> 0}^M \mid (\exists \mathbf{r}^* \in \mathcal{R}) , \boldsymbol{\rho}_{\min} \leq \Gamma_{\mathbf{r}^*}(\boldsymbol{\rho}) = \boldsymbol{\rho} \leq \mathbf{1}\}. \quad (7)$$

In the following we extend the prior knowledge in our learning framework by studying the feasible rate region $\mathcal{R} \in \mathbb{R}_{> 0}^N$ in Definition 6. In particular, we show in Theorem 1 that \mathcal{R} is compact. The compactness of \mathcal{R} is also required for our results in Section V.

Note that \mathcal{R} is bounded from below by $\mathbf{r}_{\min} \in \mathbb{R}_{> 0}^N$. Since power, bandwidth, and the total number of resource blocks are fixed in (3) and (4), and because the cell-load is monotonic in the user rate vector by Fact 4, arbitrarily large user rates cannot be supported. We state this fact formally in Lemma 1, which we use to prove compactness of \mathcal{R} in Theorem 1.

Lemma 1. The feasible rate region is bounded.

We now present the main result of this section.

Theorem 1. The feasible rate region is compact.

Proof. Recall from Definition 1(b) that a subset of a normed space is closed *if and only if* it contains all of its limit points. We denote by $\text{clo}(\mathcal{R})$ the *closure* of \mathcal{R} in Definition 6, which is the smallest closed set in $\mathbb{R}_{> 0}^N$ containing \mathcal{R} . Similarly, denote by $\text{clo}(\mathcal{L})$ the closure of \mathcal{L} in Definition 6. Consider an arbitrary sequence $(\mathbf{r}_n, \boldsymbol{\rho}_n)_{n \in \mathbb{N}} \subset \mathcal{R} \times \mathcal{L}$, of tuples consisting of feasible rate vectors and the corresponding cell-load vectors. Suppose $(\mathbf{r}_n, \boldsymbol{\rho}_n) \rightarrow (\bar{\mathbf{r}}, \bar{\boldsymbol{\rho}}) \in \text{clo}(\mathcal{R}) \times \text{clo}(\mathcal{L})$. From (5) it follows that, given $\mathbf{r}_n, \boldsymbol{\rho}_n$ must be the solution to the fixed point problem with the load mapping \mathbf{q} . Therefore, we have

$$(\forall n \in \mathbb{N}) \quad \boldsymbol{\rho}_{\min} \leq \boldsymbol{\rho}_n = \mathbf{q}(\boldsymbol{\rho}_n, \mathbf{r}_n) \leq \mathbf{1}. \quad (8)$$

Now, since \mathbf{q} is continuous, we have

$$\boldsymbol{\rho}_{\min} \leq \lim_{n \in \mathbb{N}} \boldsymbol{\rho}_n = \lim_{n \in \mathbb{N}} \mathbf{q}(\boldsymbol{\rho}_n, \mathbf{r}_n) \leq \mathbf{1}$$

$$\boldsymbol{\rho}_{\min} \leq \bar{\boldsymbol{\rho}} = \mathbf{q}(\bar{\boldsymbol{\rho}}, \bar{\mathbf{r}}) \leq \mathbf{1}$$

which implies that $(\bar{\mathbf{r}}, \bar{\boldsymbol{\rho}}) \in \mathcal{R} \times \mathcal{L}$. Thus, every convergent sequence in \mathcal{R} has its limit in \mathcal{R} which implies that \mathcal{R} is closed. Now, according to Lemma 1, \mathcal{R} is bounded and recall from Remark 2 that every bounded and closed subset of a finite dimensional Euclidean space is compact. \square

V. ROBUST LEARNING OF CELL-LOAD

Building upon the results from the previous section we formulate the robust learning of cell-load. Note that the cell-load is modeled by the load-coupling model in (4). This means that given the network information required by the model, we can calculate the value of the “modeled” cell load. However, as mentioned in Section I-A, dynamic wireless networks are in general difficult to model accurately. Therefore, in the following we present a framework to directly approximate the cell-load values in networks that may not follow the cell-load model accurately. We use the cell-load model in this study only to extract some useful prior knowledge. In addition to the monotonicity of the cell-load and the compactness of the feasible rate region \mathcal{R} established in Theorem 1, we show in Theorem 2 that the function that maps rates to cell-load is

continuously differentiable and therefore Lipschitz continuous on \mathcal{R} . The Lipschitz continuity is then used to solve our robust optimization problem formulated in the following.

Let $\mathcal{D} = \{(\mathbf{r}^k, \boldsymbol{\rho}^k := \mathbf{f}^*(\mathbf{r}^k)) \in \mathcal{R} \times \mathcal{L}, k \in \overline{1, K}\}$ be a sample set of rates and their corresponding cell-load values, where $\mathbf{f}^* : \mathcal{R} \rightarrow \mathcal{L}$ is assumed to be a continuous but unknown function, and where \mathcal{R} and \mathcal{L} are defined in Definition 6. We denote by $C(\mathcal{R}, \mathcal{L})$ the space of vector-valued continuous functions mapping \mathcal{R} to \mathcal{L} , equipped with the norm defined in (1). Our objective is to learn a function \mathbf{g}^* that approximates $\mathbf{f}^*(\mathbf{r})$ for any $\mathbf{r} \in \mathcal{R}$ which is a classical problem considered in, for example, [13], [11], [12]. As mentioned in Section I-C we are interested in a robust approximation of \mathbf{f}^* . To this end, we consider the minimax optimization problem that leads to robust solutions under uncertainties:

Problem 1. [11], [35] Given $\mathcal{D} = \{(\mathbf{r}^k, \boldsymbol{\rho}^k) \in \mathcal{R} \times \mathcal{L}, k \in \overline{1, K}\}$, find $\mathbf{g}^* \in C(\mathcal{R}, \mathbb{R}_{\geq 0}^M)$ such that the worst-case error

$$E_w(\mathbf{g}) = \sup_{\mathbf{f} \in C(\mathcal{R}, \mathcal{L})} \|\mathbf{f} - \mathbf{g}\|_{C(\mathcal{R})}, \quad (9)$$

attains its minimum (if it exists) subject to: $(\forall k \in \overline{1, K}) \mathbf{g}(\mathbf{r}^k) = \mathbf{f}(\mathbf{r}^k) = \boldsymbol{\rho}^k$.

It is known that Problem 1 can be solved by restricting \mathbf{f}^* to a compact subset of $C(\mathcal{R}, \mathcal{L})$ and by computing finite tight upper and lower bounds on the values $(\forall \mathbf{r} \in \mathcal{R}) \mathbf{f}^*(\mathbf{r})$ [13], [36], [28]. If the only information available about \mathbf{f}^* is that it satisfies the interpolation constraints in Problem 1, then computing tight bounds on unseen function values $\mathbf{f}^*(\mathbf{r})$ is not possible, no matter how large the sample set \mathcal{D} is. However, if we impose an additional restriction on \mathbf{f}^* that satisfies certain properties [13], then we can obtain tight bounds $\sigma_l(\mathbf{r})$ and $\sigma_u(\mathbf{r})$ such that $\sigma_l(\mathbf{r}) \leq \mathbf{f}^*(\mathbf{r}) \leq \sigma_u(\mathbf{r})$, where $\sigma_l(\mathbf{r})$ and $\sigma_u(\mathbf{r})$ can be computed explicitly. The optimal approximation $\mathbf{g}^*(\mathbf{r})$ of $\mathbf{f}^*(\mathbf{r})$ is simply given by $\mathbf{g}^*(\mathbf{r}) = \frac{\sigma_l(\mathbf{r}) + \sigma_u(\mathbf{r})}{2}$ and the magnitude of uncertainty $\frac{|\sigma_u(\mathbf{r}) - \sigma_l(\mathbf{r})|}{2}$ is minimal [14]. Therefore, no matter how small the sample set \mathcal{D} is we are guaranteed the minimum worst-case error (9). It is in this sense that we refer to the learning as being robust (see Section I-A).

In [36], [28] the analysis is restricted to Lipschitz functions in which case the above mentioned additional restriction results from the Lipschitz continuity. Following this approach, and by considering the cell-load model, we show in Theorem 2 that \mathbf{f}^* belongs to the class of \mathbf{L} -Lipschitz-Monotone Functions (LIMF) (see Definition 4). Moreover, Proposition 1 shows that this class is a compact subset of $C(\mathcal{R}, \mathcal{L})$. The computation of the bounds $\sigma_l(\mathbf{r})$ and $\sigma_u(\mathbf{r})$ is presented in Fact 5.

In the following we denote by $\tilde{\mathcal{R}} \subset \mathbb{R}_{>0}^N$ the set of all rate vectors (not necessarily feasible/supported) for which there exists a fixed point solution of (5), i.e., $\tilde{\mathcal{R}} := \{\bar{\mathbf{r}} \in \mathbb{R}_{>0}^N \mid (\exists \bar{\boldsymbol{\rho}} \in \mathbb{R}_{>0}^M) \bar{\boldsymbol{\rho}} = \mathbf{q}(\bar{\boldsymbol{\rho}}, \bar{\mathbf{r}})\}$. So we have $\mathcal{R} \subset \tilde{\mathcal{R}}$.

Theorem 2. Consider the load mapping $\mathbf{q} : \mathbb{R}_{\geq 0}^M \times \mathbb{R}_{>0}^N \rightarrow \mathbb{R}_{>0}^M$ in (5).

- There exists a continuously differentiable function $\mathbf{f}^{\text{imp}} : \tilde{\mathcal{R}} \rightarrow \mathbb{R}_{>0}^M$ such that $(\forall \bar{\mathbf{r}} \in \tilde{\mathcal{R}}) \mathbf{f}^{\text{imp}}(\bar{\mathbf{r}}) = \bar{\boldsymbol{\rho}} = \mathbf{q}(\bar{\boldsymbol{\rho}}, \bar{\mathbf{r}})$.
- The restriction of \mathbf{f}^{imp} to the feasible rate region $\mathcal{R} \subset \tilde{\mathcal{R}}$ is a LIMF function.

Proof. a). From the uniqueness of the fixed point solution of (5) it follows that, for two solution pairs $(\bar{\boldsymbol{\rho}}_1, \bar{\mathbf{r}}_1)$ and $(\bar{\boldsymbol{\rho}}_2, \bar{\mathbf{r}}_2)$, if $\bar{\boldsymbol{\rho}}_1 \neq \bar{\boldsymbol{\rho}}_2$, then we must have $\bar{\mathbf{r}}_1 \neq \bar{\mathbf{r}}_2$. Thus, there exists a function $\mathbf{f}^{\text{imp}} : \tilde{\mathcal{R}} \rightarrow \mathbb{R}_{>0}^M : \bar{\mathbf{r}} \mapsto \mathbf{f}^{\text{imp}}(\bar{\mathbf{r}}) = \mathbf{q}(\mathbf{f}^{\text{imp}}(\bar{\mathbf{r}}), \bar{\mathbf{r}})$ that maps every feasible rate vector to a unique fixed point. We now show that \mathbf{f}^{imp} is continuously differentiable on $\tilde{\mathcal{R}}$.

Consider the function $\mathbf{g} : \mathbb{R}_{>0}^N \times \mathbb{R}_{>0}^M \rightarrow \mathbb{R}^M$ defined as $\mathbf{g}(\mathbf{r}, \boldsymbol{\rho}) := \boldsymbol{\rho} - \mathbf{q}(\boldsymbol{\rho}, \mathbf{r})$, where \mathbf{q} is the load mapping in (5), and note that $(\forall \bar{\mathbf{r}} \in \tilde{\mathcal{R}}) (\bar{\boldsymbol{\rho}} = \mathbf{f}^{\text{imp}}(\bar{\mathbf{r}})) \mathbf{g}(\bar{\mathbf{r}}, \bar{\boldsymbol{\rho}}) = \mathbf{0}$. We now show that \mathbf{g} is continuously differentiable, and the Jacobian matrix $\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}(\bar{\mathbf{r}}, \bar{\boldsymbol{\rho}})$ is non-singular (invertible), on $\tilde{\mathcal{R}} \times \mathbb{R}_{>0}^M$ (see Fact 2). To show that \mathbf{g} is continuously differentiable, we show that the Jacobians $\nabla_{\mathbf{r}}^{\mathbf{g}}$ and $\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}$ are continuous. The two Jacobians are given in Appendix B and Appendix C, respectively, and it can be verified that they are continuous. The invertibility of the $M \times M$ matrix $\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}(\bar{\mathbf{r}}, \bar{\boldsymbol{\rho}})$ is shown in Appendix D. Therefore, according to Fact 2, \mathbf{f}^{imp} is continuously differentiable.

- According to part (a) and Fact 2, the Jacobian $\nabla_{\mathbf{r}}^{\mathbf{f}^{\text{imp}}}$ is continuous on $\tilde{\mathcal{R}}$. Denote by $\mathbf{f} : \mathcal{R} \rightarrow \mathcal{L}$ and $\nabla_{\mathbf{r}}^{\mathbf{f}}$, the restriction of \mathbf{f}^{imp} and $\nabla_{\mathbf{r}}^{\mathbf{f}^{\text{imp}}}$, respectively, to the set of feasible rate vectors $\mathcal{R} \subset \tilde{\mathcal{R}}$. Since \mathcal{R} is compact according to Theorem 1, $\nabla_{\mathbf{r}}^{\mathbf{f}}$ is bounded on \mathcal{R} according to the *extreme value theorem* [32] which implies that $\exists \mathbf{L} \in \mathbb{R}_{>0}^M$ such that \mathbf{f} is \mathbf{L} -Lipschitz on \mathcal{R} . Moreover, by Fact 4, \mathbf{f} is monotonic on \mathcal{R} , so \mathbf{f} is a LIMF function (see Definition 4). □

In the following we denote by $\mathcal{F} \subset C(\mathcal{R}, \mathcal{L})$ the class of LIMF functions $\mathbf{f} : \mathcal{R} \rightarrow \mathcal{L}$ with a given $\mathbf{L} \in \mathbb{R}_{\geq 0}^M$ (see Definition 4). Before we proceed further, we obtain the following important result whose proof is shown in Appendix E.

Proposition 1. The class $\mathcal{F} \subset C(\mathcal{R}, \mathcal{L})$ of LIMF functions, with a given $\mathbf{L} = [L_1, L_2, \dots, L_M]^T \in \mathbb{R}_{\geq 0}^M$, is compact.

A. Minimax Optimal Approximation

We are now in a position to incorporate the prior information obtained in previous sections into Problem 1. Moreover, we formally state the robust learning problem considered in this study as an optimization problem.

Definition 7 (Minimax Optimal Approximation). Let $\mathcal{D} = \{(\mathbf{r}^k, \boldsymbol{\rho}^k) \in \mathcal{R} \times \mathcal{L}\}_{k=1}^K$ be a sample set and assume that $(\forall k \in \overline{1, K}) \boldsymbol{\rho}^k := \mathbf{f}^*(\mathbf{r}^k)$ are values generated by an unknown function $(\mathcal{F} \ni) \mathbf{f}^* : \mathcal{R} \rightarrow \mathcal{L}$, where $\mathcal{F} \subset C(\mathcal{R}, \mathcal{L})$ is a set of LIMF functions with a given $\mathbf{L} \in \mathbb{R}_{\geq 0}^M$. The minimax optimal approximation problem can be then stated as follows:

Problem 2. [11], [28], [35] Find \mathbf{g}^* such that

$$\mathbf{g}^* \in \arg \min_{\mathbf{g} \in \mathcal{S}} E_{\max}(\mathbf{g}) \quad (10)$$

where $\mathcal{S} := \{\mathbf{g} \in C(\mathcal{R}, \mathbb{R}_{>0}^M) \mid (\forall k \in \overline{1, K}) \mathbf{g}(\mathbf{r}^k) = \boldsymbol{\rho}^k\}$, and $E_{\max}(\mathbf{g}) := \max_{\mathbf{f} \in \mathcal{F}} \|\mathbf{f} - \mathbf{g}\|_{C(\mathcal{R})}$ is the worst-case error (9) computed over the set \mathcal{F} .

The study [28] proposes a framework for interpolation of scalar Lipschitz functions defined over a compact set by using a *central algorithm* [11], [12]. This framework can be used to obtain a solution to Problem 2. Furthermore, this method is also “shape preserving”, i.e., the approximation preserves the Lipschitz continuity and monotonicity of the underlying original function. The following fact summarizes the important properties of an optimal solution obtained based on this framework.

Fact 5. [28] Let $\mathcal{D} = \{(\mathbf{r}^k, \rho^k) \in \mathcal{R} \times \mathcal{L}\}_{k=1}^K$ be a dataset generated by an unknown function $\mathbf{f}^* \in \mathcal{F}$, where \mathcal{F} is the set of LIMF functions with the same $\mathbf{L} := [L_1, L_2, \dots, L_M]^\top \in \mathbb{R}_{\geq 0}^M$. Then, the following holds:

- a). A minimax optimal approximation \mathbf{g}^* of $\mathbf{f}^* \in \mathcal{F}$ can be constructed component-wise by

$$(\forall i \in \overline{1, M}) (\forall \mathbf{r} \in \mathcal{R}) \quad g_i^*(\mathbf{r}) = \frac{\sigma_l^i(\mathbf{r}) + \sigma_u^i(\mathbf{r})}{2}, \quad (11)$$

where $\sigma_l^i(\mathbf{r}) = \max_k \{\rho_i^k - L_i \|(\mathbf{r}^k - \mathbf{r})_+\|\}$, $\sigma_u^i(\mathbf{r}) = \min_k \{\rho_i^k + L_i \|(\mathbf{r} - \mathbf{r}^k)_+\|\}$, and $L_i \in \mathbb{R}_{\geq 0}$ is the Lipschitz constant of the i th component f_i^* of \mathbf{f}^* .

- b). The approximation preserves the \mathbf{L} -Lipschitz continuity and monotonicity, i.e., \mathbf{g}^* is \mathbf{L} -Lipschitz and monotonic.
c). \mathbf{g}^* interpolates the sample set \mathcal{D} .

B. Complexity

The complexity of the closed-form computation (11) is linear in the sample size K , i.e., the complexity is $O(K)$. Since we consider very small sample sizes, the complexity is not of a practical concern. Moreover, (11) can be computed independently for each base station. Therefore, the complexity is independent of the number of base stations M .

Remark 3 (Prior Knowledge Decreases Uncertainty). Note that the study [28] is concerned with shape preserving approximation and it does not consider learning from a small sample set. However, we show in Proposition 2 that (except for one particular case) excluding prior information regarding monotonicity worsens at least one of the bounds in Fact 5(a) during generalization on unseen data and this therefore increases uncertainty and error. We also evaluate this fact empirically in Section VII-B1 in a realistic wireless network.

The lower and upper bounds without monotonicity constraints in Fact 5 are given by $(i \in \overline{1, M}) \quad \eta_l^i(\mathbf{r}) = \max_k \{\rho_i^k - L_i \|\mathbf{r}^k - \mathbf{r}\|\}$ and $\eta_u^i(\mathbf{r}) = \min_k \{\rho_i^k + L_i \|\mathbf{r} - \mathbf{r}^k\|\}$. Let $U_{\text{mon}}(\mathbf{r}) := \frac{|\sigma_u^i(\mathbf{r}) - \sigma_l^i(\mathbf{r})|}{2}$ denote the magnitude of uncertainty calculated from the bounds in Fact 5, and let $U(\mathbf{r}) := \frac{|\eta_u^i(\mathbf{r}) - \eta_l^i(\mathbf{r})|}{2}$ denote the magnitude of uncertainty without monotonicity in the framework.

Proposition 2. Let $\mathbf{r} \notin \mathcal{D} = \{(\mathbf{r}^k, \rho^k) \in \mathcal{R} \times \mathcal{L}\}_{k=1}^K$, where \mathcal{D} is the data set in Fact 5. Then $U_{\text{mon}}(\mathbf{r}) \leq U(\mathbf{r})$ if

- a). $(k^* \in \arg\max_k \{\rho_i^k - L_i \|(\mathbf{r}^k - \mathbf{r})_+\|\}) \quad \mathbf{r}^{k^*} \geq \mathbf{r}$, and
b). $(j^* \in \arg\min_j \{\rho_i^j + L_i \|(\mathbf{r} - \mathbf{r}^j)_+\|\}) \quad \mathbf{r}^{j^*} \leq \mathbf{r}$;

otherwise $U_{\text{mon}}(\mathbf{r}) < U(\mathbf{r})$.

Proof. Consider two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_{\geq 0}^N$ such that $\mathbf{x} \neq \mathbf{y}$. If $\mathbf{x} \geq \mathbf{y}$, then $\|(\mathbf{x} - \mathbf{y})_+\| = \|(\mathbf{x} - \mathbf{y})\|$ and $\|(\mathbf{y} - \mathbf{x})_+\| < \|(\mathbf{x} - \mathbf{y})\|$. Similarly, if $\mathbf{x} \leq \mathbf{y}$, then $\|(\mathbf{x} - \mathbf{y})_+\| < \|(\mathbf{x} - \mathbf{y})\|$ and $\|(\mathbf{y} - \mathbf{x})_+\| = \|(\mathbf{x} - \mathbf{y})\|$. If \mathbf{x} and \mathbf{y} are incomparable then $\|(\mathbf{y} - \mathbf{x})_+\| < \|(\mathbf{x} - \mathbf{y})\|$ and also $\|(\mathbf{x} - \mathbf{y})_+\| < \|(\mathbf{x} - \mathbf{y})\|$.

Now, if conditions a) and b) are satisfied simultaneously, then (by condition a)) for the lower bound we have

$$\begin{aligned} \eta_l^i(\mathbf{r}) &= \{\rho_i^{k^*} - L_i \|(\mathbf{r}^{k^*} - \mathbf{r})_+\|\} \\ &= \{\rho_i^{k^*} - L_i \|(\mathbf{r}^{k^*} - \mathbf{r})_+\|\} \\ &\leq \max_k \{\rho_i^k - L_i \|(\mathbf{r}^k - \mathbf{r})_+\|\} = \sigma_l^i(\mathbf{r}). \end{aligned}$$

Similarly, (by condition b)) $\sigma_u^i(\mathbf{r}) \leq \eta_u^i(\mathbf{r})$. This proves the first claim of the proposition. Now suppose condition a) is violated, i.e., either $\mathbf{r}^{k^*} \leq \mathbf{r}$ or \mathbf{r}^{k^*} and \mathbf{r} are incomparable, then from the above discussion

$$\begin{aligned} \eta_l^i(\mathbf{r}) &= \{\rho_i^{k^*} - L_i \|(\mathbf{r}^{k^*} - \mathbf{r})\|\} \\ &< \{\rho_i^{k^*} - L_i \|(\mathbf{r}^{k^*} - \mathbf{r})_+\|\} \\ &\leq \max_k \{\rho_i^k - L_i \|(\mathbf{r}^k - \mathbf{r})_+\|\} = \sigma_l^i(\mathbf{r}). \end{aligned}$$

Similarly, if condition b) is violated, $\sigma_u^i(\mathbf{r}) < \eta_u^i(\mathbf{r})$ and the second claim follows. \square

The consequence of Proposition 2 is that $U_{\text{mon}}(\mathbf{r}) < U(\mathbf{r})$ whenever \mathbf{r} violates either of the two conditions in Proposition 2. Therefore, including prior knowledge in our framework regarding monotonicity provably improves generalization/prediction on unseen data.

VI. IMPLEMENTATION IN A WIRELESS NETWORK

We have shown in Theorem 2 that there exists an implicit function $(\forall i \in \overline{1, M}) \quad f_i : \mathcal{R} \rightarrow]0, 1]$ mapping every $\mathbf{r} \in \mathcal{R}$ to a cell-load value ρ_i at base station i . Furthermore, Fact 5 shows that given a sample set $\mathcal{D}(i) = \{(\mathbf{r}^k, f_i(\mathbf{r}^k))\}_{k=1}^K$ at base station i and the knowledge of the Lipschitz constant L_i , we can easily approximate the cell-load value $f_i(\mathbf{r})$ for $\mathbf{r} \notin \mathcal{D}(i)$. In this section we show how to implement our framework in an OFDMA-based wireless cellular network. To this end, we first look at how to calculate the cell-load, and then we show how to obtain an appropriate sample set at a base station.

A. Cell-load Calculation

In OFDMA-based networks, such as LTE networks, time is divided into fixed length slots known as *subframes*. During a subframe, if a base station is active, it transmits to one or more users on a block of frequencies in its cell. Therefore, users are allocated subframes in time and bandwidth in frequency to match their rate requirements. A subframe together with its bandwidth is commonly referred to as a *physical resource block*. To calculate the cell-load, we record the fraction of the total available physical resource blocks allocated by a base station on average during a total time period of $T_{\text{avg}} > 0$, where T_{avg} is a design parameter.

Algorithm 1 Cell-load Learning for each Base Station

→ **Initialization**

- Fix $K > 0$ and $T_{\text{avg}} > 0$.

→ **Sample Acquisition** (while $t < T_{\text{obv}}$)

- Exchange user rate with other base stations.
- Observe the sample set $\mathcal{D}^{\text{noise}} = \{(\mathbf{r}^k, y^k = f(\mathbf{r}^k) + \epsilon(\mathbf{r}^k))\}_{k=1}^K$ (Section VI-B).

→ **Training** (at $t = T_{\text{obv}}$)

- Perform the estimation of L (Section VI-C).
- Perform data smoothing to obtain a compatible $\mathcal{D}^{\text{com}} = \{(\mathbf{r}^k, \tilde{\rho}^k)\}_{k=1}^K$ (Section VI-C).

→ **On-Demand Prediction** (at $t > T_{\text{obv}}$)

- Given a new rate vector $\mathbf{r} \in \mathcal{R}$, perform the computation (11) in Fact 5

$$g(\mathbf{r}) = \frac{1}{2}(\max_k \{\tilde{\rho}^k - L\|(\mathbf{r}^k - \mathbf{r})_+\|\}) + \frac{1}{2}(\min_k \{\tilde{\rho}^k + L\|(\mathbf{r} - \mathbf{r}^k)_+\|\}).$$

B. Obtaining a Sample Set

We denote by $T_{\text{net}} > 0$ the network coherence time during which the environment (network topology, channels, rate distribution, etc.) is assumed to be constant (see Section I-A). Let $T_{\text{obv}} < T_{\text{net}}$ denote the sample observation time. We divide T_{obv} in $K \in \mathbb{N}$ time windows of duration T_{avg} each as shown in Figure 1. To obtain a sample set $\mathcal{D}(i) = \{(\mathbf{r}^k, \rho_i^k = f_i(\mathbf{r}^k))\}_{k=1}^K$ at each base station $i \in \overline{1, M}$, the cell-load values $\rho_i^k = f_i(\mathbf{r}^k)$ can be calculated as in Section VI-A for each time window $k \in \overline{1, K}$. The base stations can exchange the rate values of users associated with them with other base stations to obtain the rate vectors \mathbf{r}^k .

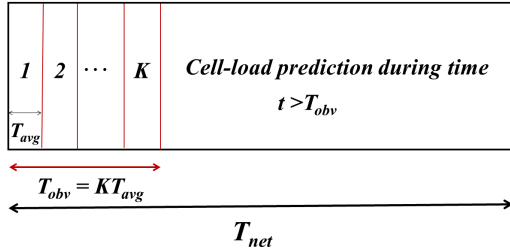


Fig. 1. Learning Timeline: During each slot $k \in \overline{1, K}$ of length T_{avg} we obtain a sample (\mathbf{r}^k, ρ_i^k) by observing the proportion of resource blocks consumed to support rate \mathbf{r}^k on average during T_{avg} .

In the following, we assume that a sample set $\mathcal{D}(i) = \{(\mathbf{r}^k, \rho_i^k = f_i(\mathbf{r}^k))\}_{k=1}^K$ is available at time $t = T_{\text{obv}}$ at base station $i \in \overline{1, M}$. We also omit the index i since the same procedure is carried out at each base station.

C. Obtaining a Compatible Sample Set

Note that the cell-load values calculated in a real network do not follow the cell-load model exactly. In more detail, instead of the sample set $\mathcal{D} = \{(\mathbf{r}^k, \rho^k = f(\mathbf{r}^k))\}_{k=1}^K$, we assume that an inaccurate sample set $\mathcal{D}^{\text{error}} = \{(\mathbf{r}^k, y^k = f(\mathbf{r}^k) +$

$\epsilon(\mathbf{r}^k))\}_{k=1}^K$ is available; $\epsilon(\mathbf{r}^k) \geq 0$ is the inaccuracy/error which is assumed to be bounded.⁴ As a consequence, for a given value of the Lipschitz constant $L \in \mathbb{R}_{\geq 0}$, $\mathcal{D}^{\text{error}}$ may not be compatible with the monotonicity of f . Therefore, and if required, it must be smoothed to obtain a compatible set. Furthermore, in practice the prior information about the Lipschitz constant L is often unavailable, so its value must be estimated from the set $\mathcal{D}^{\text{error}}$. In more detail, we first estimate the Lipschitz constant by $\tilde{L} := \max_{k \neq j} \frac{|y^k - y^j|}{\|\mathbf{r}^k - \mathbf{r}^j\|}$ [39].⁵ Given an estimate \tilde{L} of the Lipschitz constant, we perform monotone-smoothing of $\mathcal{D}^{\text{error}}$. The details are provided in Appendix F.

D. Algorithm

The robust cell-load learning algorithm is presented in Algorithm 1. The *Sample Acquisition* step corresponds to the acquisition of the training sample set as explained in Section VI-B, whereas *Training* refers to Lipschitz constant estimation and the data smoothing process as presented in Appendix F. The *On-Demand Prediction* refers to the approximation of the cell-load value for a new rate vector during time period $T_{\text{net}} - T_{\text{obv}}$ (also see Figure 1).

VII. NUMERICAL EVALUATION

In this section we evaluate the robust learning framework presented in Section V-A by simulation. To evaluate the learning techniques in a realistic cellular network, simulations are performed in the *network simulator* (NS3) [40]. We focus on the following aspects in this numerical evaluation:

- 1) We only use the load-coupling model (see Section III-A) in this study to establish some prior knowledge about the cell-load in a real cellular network. We show in the simulations that our learning framework is able to predict the cell-load sufficiently accurately in a realistic cellular network in NS3. This is significant because models are only idealizations, and they may not capture the true behavior of cellular networks.
- 2) We have shown in Proposition 2 that including prior knowledge decreases the uncertainty. We demonstrate this by comparing our learning framework with full prior knowledge with the case in which the prior information regarding the monotonicity of the cell-load with respect to rate is not included in the framework.
- 3) Finally, we compare our method to standard multivariate regression techniques. We show the effect of sample size K and the size of the network (i.e., the number of users N and base stations M) on the quality of approximation.

In the next section we present the LTE simulation framework in NS3.

⁴Our approximation framework is a special case of bounded error estimation/robust set-membership estimation [37], [38] which was developed for scenarios where the inaccuracy is unknown but bounded.

⁵There exist more sophisticated methods of estimating the Lipschitz constant such as the method proposed in [28]. But these methods are not the focus of this study and they add substantial complexity to the algorithm.

TABLE II
NS3 SIMULATION PARAMETERS

Description	Value
Number of base stations M	3
Number of users N	30
Base station height	30 m
User height	1.5 m
Noise figure base station	5 dB
Noise figure user	9 dB
Min/Max user rate	$0.1 \times 10^6 / 1 \times 10^6$
Simulation area	200×200 m
Simulation time	1 s
Total bandwidth	10 MHz
Total number of resource blocks	50
Path-loss model	Log-Distance Propagation Loss
SRS periodicity	80×10^{-3} s
Internet application	On-Off with Ipv4

A. Network Simulator (NS3) and Scenario

We perform simulation in NS3 using the LTE model, the details of which can be found in [40]. The load coupling model is evaluated in the LTE downlink in certain scenarios in [9]. Briefly, NS3 is a well-known discrete-event network simulator widely used in educational research and industry due to its accuracy in simulating computer networks such as LTE. The granularity of the LTE model in NS3 is up to the resource block level which allows for accurate packet scheduling and calculation of inter-cell interference. We chose the *Round Robin* scheduler at the MAC layer. The reason is that the fairness inherent in the simple cyclic scheduling is more likely to ensure that the minimum data rate requirement of all users are met, which may not be the case with other more complex scheduling algorithms [41]. The *modulation and coding scheme* and the resource block allocation are chosen based on the wide-band *channel quality indicator* (CQI). The CQI is calculated based on the average received SINR. Users and base stations are distributed uniformly in the service area of 200×200 meters. We perform simulations for $M = \{3, 5, 7, 9, 10\}$ base stations with $N = \{30, 50, 70, 80, 90, 100\}$ users. Users are associated with the base station to which they have the lowest path-loss. To generate training and test data, the data rates are distributed uniformly between 0.1×10^6 bits/s and 1×10^6 bits/s. The important simulation parameters are shown in Table II. Other parameters were chosen as default in NS3. The simulation time was chosen to be 1 second which is equal to the length T_{avg} of each averaging time slot/window in Figure 1 and Algorithm 1. The cell-load values are calculated according to Section VI-A.

B. Results

We now present our numerical results. We use Algorithm 1 to perform the robust learning of cell-load proposed in this study. We present the results for cell-load learning at a single base station. To obtain reliable statistics we consider 50 topologies (with different user locations, base station locations, and user associations) for each value of N and we let $M = N/10$. Note that scaling the number of base stations with an increase in the number of users is necessary to ensure that rate requirements of users are met. The objective of the simulation

is to observe the effect of sample size and the network size on the approximation. For each fixed topology, we perform 100 experiments for each value of $K \in \{10, 20, \dots, 100\}$. During each experiment, a sample set $\mathcal{D}^{\text{error}} = \{(\mathbf{r}^k, y^k)\}_{k=1}^K$ is generated independently at random and the *Training Step* is performed in Algorithm 1 to obtain a compatible training sample set \mathcal{D}^{com} . Validation/prediction is performed for an independent test sample set of size 1000 with rate vectors $\mathbf{r} \notin \mathcal{D}^{\text{com}}$. All results are averaged over 100 experiments and then over 50 topologies to obtain reliable statistics.

1) *Effect of Prior Information*: In this section we compare our framework's performance with and without the prior information regarding the monotonicity of the cell-road with respect to rate (see Remark 3). For this simulation we consider $M = 3$ and $N = 30$. Note that the objective of this rather theoretical comparison is to confirm the result of Proposition 2 in a realistic simulation. This comparison is performed with an ideal Lipschitz constant L^{ideal} that can be obtained by using the method in Section VI-C but by using both the training sample set and the test sample set. This way L^{ideal} is a good approximation of the true Lipschitz constant. We chose an ideal Lipschitz constant because in this section we want to focus only on the effect of including prior knowledge regarding monotonicity of the cell-load in rate in a realistic cellular network, and this requires an accurate calculation of function bounds in Section V. However, the comparison with *state-of-art* techniques in Section VII-B2, which is of a more practical significance, is performed with the Lipschitz constant that is estimated from only the training data set.

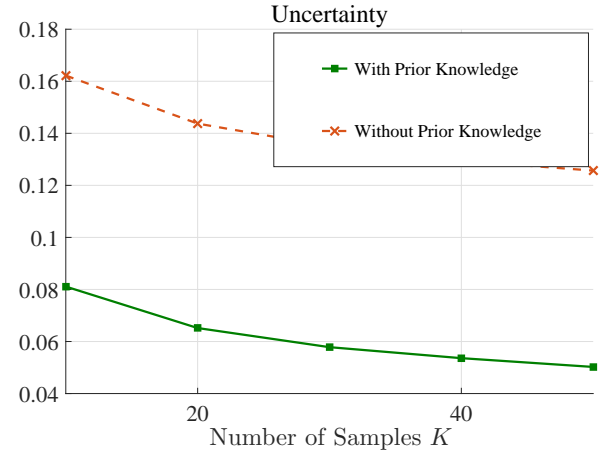


Fig. 2. We compare the performance of our framework with the case where prior knowledge about the monotonicity of the cell-load has not been considered.

We perform the comparison in terms of two metrics, namely the *magnitude of uncertainty* given as $\frac{|\sigma_u(\mathbf{r}) - \sigma_l(\mathbf{r})|}{2}$ (see Section V), where the rate \mathbf{r} is a test sample point and $\sigma_u(\mathbf{r})$ and $\sigma_l(\mathbf{r})$ are upper and lower bounds, and the *correlation* with test sample set that we measure in terms of the popular *Pearson's correlation coefficient*.

The results are shown in Figure 2 and Figure 3. Figure 2 shows that uncertainty about the cell-load values decreases with the increasing training sample set size K in both cases.

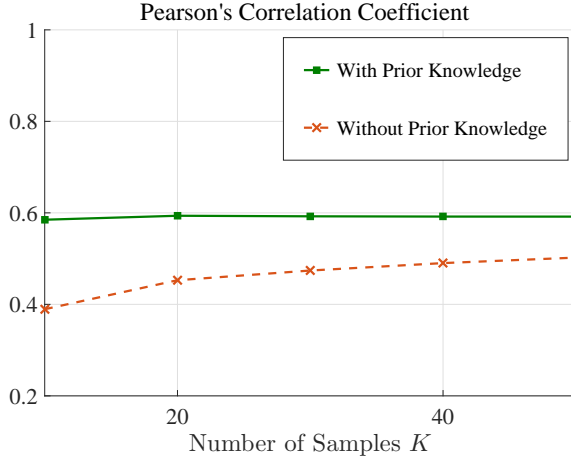


Fig. 3. We compare the performance of LIMF learning framework with the case where prior knowledge about the monotonicity of the cell-load has not been considered.

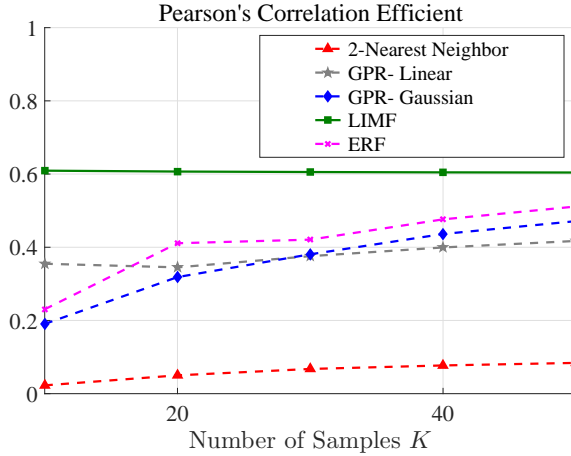


Fig. 4. We compare the the 5 techniques in terms of the linear correlation between predictions and true values for increasing K .

However, we observe that the prior information regarding the monotonicity always results in less uncertainty than the case where monotonicity of the cell-load is ignored. The results are therefore of a theoretical significance and they justify the inclusion of monotonicity as part of the prior knowledge in the framework (see Remark 3). The same effect is seen in Figure 3 where we can clearly see that the case with all prior information included in the framework results in more correlation with the test sample set.

2) *Comparison with State-of-Art Techniques:* In this section we compare our learning framework with some low-complexity state-of-art techniques for various training sample and network sizes. Throughout this section, we estimate L from the available training sample set. We compare our method with four multivariate techniques, namely the state-of-art methods *Gaussian process regression* (GPR) and *ensemble learning with random forests* (ERF), and the simple *2-nearest neighbor interpolation*. The GPR technique is well-known for its universal approximation of continuous functions defined over compact sets. Note that, in addition to the state-of-art

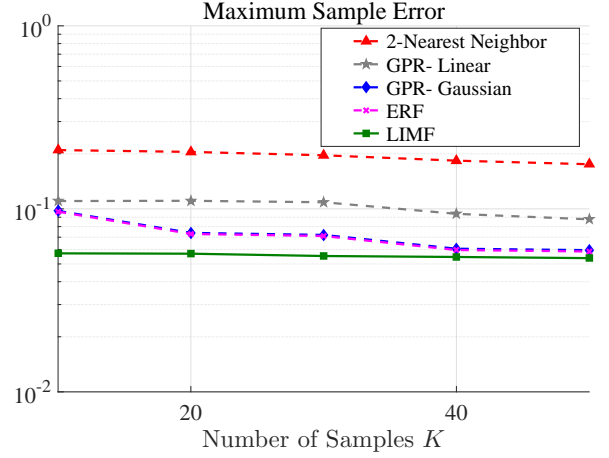


Fig. 5. We compare the the 5 techniques in terms of the maximum error between predictions and true values for increasing K .

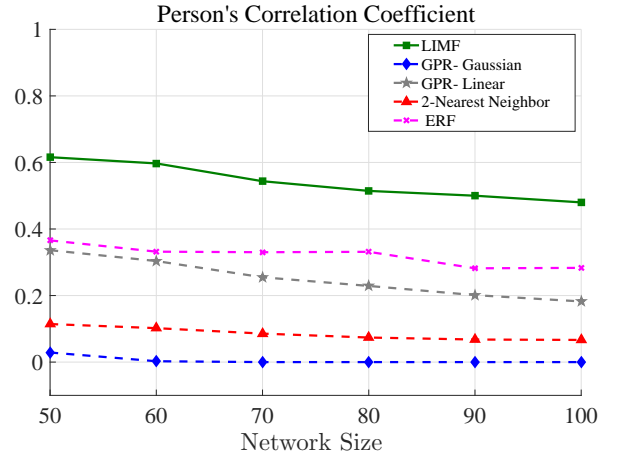


Fig. 6. We compare the the 5 techniques in terms of the linear correlation between predictions and true values for increasing network size.

methods, it is important to compare the performance with a simple method such as the *2-nearest neighbor interpolation* to highlight the difficulty of learning with small sample sets. We stress again that we consider very small sizes.

Figure 4 shows a comparison of (linear) *Pearson's* correlation coefficient, which is a popular measure of the strength and direction of the linear relationship between the predicted and the real test values, for an increasing sample size and fixed number of users $N = 30$. In particular, we use this coefficient as a measure of the “quality” of approximation. A high positive value of *Pearson's* correlation coefficient means that the predictions made by the learning method have a strong linear relationship with the test sample set. Figure 5 shows the maximum or worst-case error encountered while predicting on the test sample set for an increasing sample size K and fixed number of users $N = 30$. The maximum error is more suitable for comparing the robustness of the approximation techniques than some other popular error metrics because it shows that all error residuals remain below this level. Therefore, the maximum error is a reasonable substitute for the maximum error of approximation in (9) which we cannot

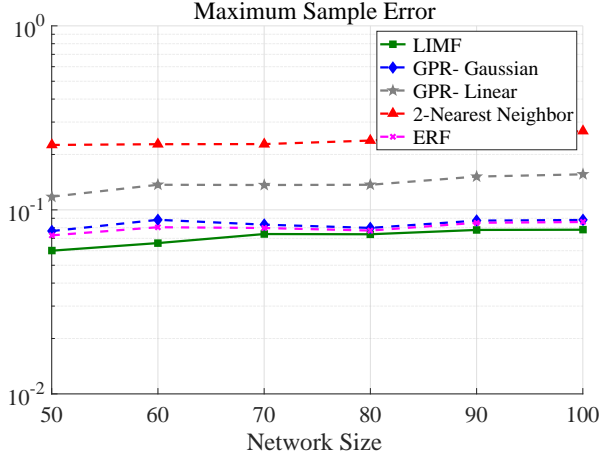


Fig. 7. We compare the 5 techniques in terms of the maximum error between predictions and true values for increasing network size.

TABLE III
TRAINING TIME COMPARISON ON STANDARD PC

Technique	Average Training Time
LIMF	10×10^{-3} seconds
Nearest Neighbor	not applicable
GPR	80×10^{-3} seconds
ERF	60×10^{-3} seconds

compute directly.

It is important to analyze maximum error and correlation together to better understand the comparison between our learning framework and other techniques. We observe that even for an inexact value of Lipschitz constant L , our method outperforms other techniques. An interesting observation is the fact that the GPR method (with the Gaussian function) and ERF show a relatively good error performance in Figure 5 but a considerably smaller correlation in Figure 4 than our method for small sample sizes $K < 30$. This is because of the fact that our method incorporates prior knowledge about the cell-load and other methods do not. The poorest performance is seen in the case of the *2-nearest neighbor interpolation* whose performance improves slowly with increasing sample size. Clearly, this shows that we do not have enough samples to perform such a simple interpolation.

Finally, Figure 6 and Figure 7 show the effect of network size (in terms of number of users N) on the performance of all techniques for a small sample size of $K = 20$. We see that, as expected, there is a gradual degradation of performance for all techniques. In particular, we observe in Figure 6 that the GPR with Gaussian function performs poorly due to insufficient training.

VIII. CONCLUSION

We have studied the problem of robust learning of cell-load in dynamic wireless cellular networks with small sample sets. In this challenging setting, we have proposed a learning framework that is robust against uncertainties that result from learning based on a small training sample set. We have shown that robustness can be achieved with the help of some

prior knowledge about the cell-load and its relationship with downlink rates. For example, an inherent property of the cell-load is that it is monotonic in rates so this property can be used as prior knowledge. To obtain additional prior knowledge, we have shown that the feasible rate region is compact, and that there exists a Lipschitz continuous function mapping feasible rates to the cell-load. These properties enable us to use the classical framework of minimax approximation. In this framework the objective is to minimize the worst-case error given a training sample set and the prior knowledge. We have shown by simulations in NS3 that, in a realistic scenario, our method outperforms other popular learning techniques. An extension of this study is to develop sophisticated methods for estimation of the Lipschitz constant from small sample sets.

APPENDIX

A. Proof of Equicontinuity of \mathbf{L} -Lipschitz functions

Let $\mathcal{F} \subset C(\mathcal{X}, \mathcal{Y})$ denote the set of \mathbf{L} -Lipschitz functions with $\mathbf{L} := [L_1, L_2, \dots, L_M]^T \in \mathbb{R}_{\geq 0}^M$. Since each component of $\mathbf{f} \in \mathcal{F}$ is Lipschitz on $\mathcal{X} \subset \mathbb{R}_{>0}^N$, we have that

$$(\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}) (\forall i \in \overline{1, M}) |f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq L_i \|\mathbf{x} - \mathbf{y}\|. \quad (12)$$

Define $L_{\max} := \max_{i \in \overline{1, M}} L_i$ and note that

$$(\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}) \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_{\infty} \leq L_{\max} \|\mathbf{x} - \mathbf{y}\|. \quad (13)$$

From the equivalence of norms in finite dimensional normed spaces it follows that $(\exists C > 0)$ such that

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq C \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_{\infty} \leq C L_{\max} \|\mathbf{x} - \mathbf{y}\|. \quad (14)$$

Given $\epsilon > 0$ and for every $\mathbf{x}_o \in \mathcal{X}$, choose $\delta := \frac{\epsilon}{L_{\max} C}$ as the radius of $B_{\mathcal{X}}(\mathbf{x}_o, \delta)$. We have from (14) that

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_o)\| \leq C L_{\max} \|\mathbf{x} - \mathbf{x}_o\| < \epsilon, \quad (15)$$

whenever $\|\mathbf{x} - \mathbf{x}_o\| < \delta$. We have shown that δ can be chosen independently of \mathbf{x}_o . Now since (15) holds for every $\mathbf{f} \in \mathcal{F}$, the proof is complete.

B. Jacobian of \mathbf{g} with respect to \mathbf{r}

The entry $[\nabla_{\mathbf{r}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})]_{i,j}$ of the $M \times N$ Jacobian $\nabla_{\mathbf{r}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})$ is given by

$$[\nabla_{\mathbf{r}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})]_{i,j} = \begin{cases} -\frac{1}{RB \log(1+\gamma_{ij})}, & \text{if } j \in N(i) \\ 0, & \text{otherwise} \end{cases}$$

where $\gamma_{ij} := \frac{p_i G_{i,j}}{\sum_{k \in M \setminus \{i\}} p_k G_{k,j} \rho_k + \sigma^2}$.

C. Jacobian of \mathbf{g} with respect to $\boldsymbol{\rho}$

The entry $[\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})]_{i,k}$ of the $M \times M$ Jacobian $\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})$ is given by

$$[\nabla_{\boldsymbol{\rho}}^{\mathbf{g}}(\mathbf{r}, \boldsymbol{\rho})]_{i,k} = \begin{cases} -\sum_{j \in N(i)} \ln(2) \frac{r_j}{RB} \frac{\frac{p_i G_{i,j}}{p_k G_{k,j}}}{\ln^2(1+\gamma_{i,j})(\gamma_{i,j}^{-2} + \gamma_{i,j}^{-1})}, & \text{if } i \neq k \\ 1, & \text{if } i = k \end{cases}$$

where $\gamma_{ij} := \frac{p_i G_{i,j}}{\sum_{k \in M \setminus \{i\}} p_k G_{k,j} \rho_k + \sigma^2}$.

D. Invertibility of the Jacobian $\nabla_{\rho}^g(\mathbf{r}, \rho)$

We follow the analysis in [21] which exploits the sufficient conditions for invertibility of a generalized diagonal dominant matrix [42] on the whole domain. In more detail, we show that the matrix $\nabla_{\rho}^g(\mathbf{r}, \rho)$ is invertible because it is an invertible generalized diagonal dominant matrix. For any $\rho \in \mathbb{R}_{>0}^M$

$$[\nabla_{\rho}^g(\mathbf{r}, \rho)]_i \rho = \rho_i - \sum_{j \in N(i)} \frac{r_j}{RB \log(1 + \gamma_{i,j})} \times \frac{\frac{\sum_{k \in M \setminus \{i\}} \rho_k p_k G_{k,j}}{p_i G_{i,j}}}{\ln(1 + \gamma_{i,j})(\gamma_{i,j}^{-2} + \gamma_{i,j}^{-1})},$$

where $[\nabla_{\rho}^g(\mathbf{r}, \rho)]_i$ is the i th row of $\nabla_{\rho}^g(\mathbf{r}, \rho)$. Since $\frac{\sum_{k \in M \setminus \{i\}} \rho_k p_k G_{k,j}}{p_i G_{i,j}} < \frac{\sum_{k \in M \setminus \{i\}} \rho_k p_k G_{k,j} + \sigma^2}{p_i G_{i,j}} = \gamma_{i,j}^{-1}$ and $\ln(1 + \gamma_{i,j})(\gamma_{i,j}^{-2} + \gamma_{i,j}^{-1}) > \gamma_{i,j}^{-1}$ [21], we have

$$\sum_{j \in N(i)} \frac{r_j}{RB \log(1 + \gamma_{i,j})} \times \frac{\frac{\sum_{k \in M \setminus \{i\}} \rho_k p_k G_{k,j}}{p_i G_{i,j}}}{\ln(1 + \gamma_{i,j})(\gamma_{i,j}^{-2} + \gamma_{i,j}^{-1})} < \sum_{j \in N(i)} \frac{r_j}{RB \log(1 + \gamma_{i,j})} = \rho_i$$

which implies that $[\nabla_{\rho}^g(\mathbf{r}, \rho)]_i \rho > 0$. Since the off-diagonal entries are all non-positive and diagonal entries are all non-negative, $\nabla_{\rho}^g(\mathbf{r}, \rho)$ satisfies the sufficient conditions for it to be an invertible generalized diagonal dominant matrix [21], [42].

E. Proof of Proposition 1

Proof. The class $\mathcal{F} \subset C(\mathcal{R}, \mathcal{L})$ satisfies the following properties:

- Boundedness:** \mathcal{F} is bounded because $(\forall \mathbf{f} \in \mathcal{F}) \|\mathbf{f}\|_{C(\mathcal{R})} \leq 1$.
- Equicontinuity:** Since \mathcal{F} is a set of \mathbf{L} -Lipschitz functions, \mathcal{F} is an equicontinuous subset of $C(\mathcal{R}, \mathcal{L})$ (see Remark 1).
- Closedness:** The class \mathcal{F} can be written as $\mathcal{F} = \mathcal{F}^{\text{Lip}} \cap \mathcal{F}^{\text{mon}}$, where \mathcal{F}^{Lip} and \mathcal{F}^{mon} are the sets of \mathbf{L} -Lipschitz functions and continuous monotone functions, respectively, in $C(\mathcal{R}, \mathcal{L})$. Recall that the intersection of two closed sets is closed. Therefore, it is sufficient to show that \mathcal{F}^{Lip} and \mathcal{F}^{mon} are closed sets. For completeness, we show in Lemma 2 that \mathcal{F}^{mon} and \mathcal{F}^{Lip} are closed sets.

The proposition now follows from Fact 1. \square

Lemma 2. Consider the space $C(\mathcal{X}, \mathcal{Y})$.

- The set of monotonic functions \mathcal{F}^{mon} in $C(\mathcal{X}, \mathcal{Y})$ is closed.
- The set of \mathbf{L} -Lipschitz functions \mathcal{F}^{Lip} in $C(\mathcal{X}, \mathcal{Y})$ is closed.

Proof. a). Let $(\mathbf{f}_n)_{n \in \mathbb{N}} \subset \mathcal{F}^{\text{mon}} \subset C(\mathcal{R}, \mathcal{L})$ be an arbitrary convergent sequence of continuous monotone functions converging to some $\mathbf{g} \in C(\mathcal{R}, \mathcal{L})$. Then from Definition

2, and the fact that inequalities are preserved in the limit, it follows that:

$$\begin{aligned} (\forall \mathbf{x}, \mathbf{y} \in \mathcal{R}) \mathbf{x} \leq \mathbf{y} &\implies (\forall n \in \mathbb{N}) \mathbf{f}_n(\mathbf{x}) \leq \mathbf{f}_n(\mathbf{y}) \\ (\forall \mathbf{x}, \mathbf{y} \in \mathcal{R}) \mathbf{x} \leq \mathbf{y} &\implies \lim_{n \rightarrow \infty} \mathbf{f}_n(\mathbf{x}) \leq \lim_{n \rightarrow \infty} \mathbf{f}_n(\mathbf{y}) \\ (\forall \mathbf{x}, \mathbf{y} \in \mathcal{R}) \mathbf{x} \leq \mathbf{y} &\implies \mathbf{g}(\mathbf{x}) \leq \mathbf{g}(\mathbf{y}), \end{aligned}$$

which means that $\mathbf{g} \in \mathcal{F}^{\text{mon}}$. Since $(\mathbf{f}_n)_{n \in \mathbb{N}}$ was chosen arbitrarily, the above holds for every sequence in \mathcal{F}^{mon} showing that \mathcal{F}^{mon} is closed.

- Following the same idea as above, we show that the limit function $\mathbf{g} \in C(\mathcal{R}, \mathcal{L})$ of an arbitrary sequence $(\mathbf{f}_n^{\text{Lip}})_{n \in \mathbb{N}} \subset \mathcal{F}^{\text{Lip}} \subset C(\mathcal{R}, \mathcal{L})$ is Lipschitz with the same \mathbf{L} , i.e., $\mathbf{g} \in \mathcal{F}^{\text{Lip}}$ also. Note that $\|\mathbf{f}_n^{\text{Lip}} - \mathbf{g}\|_{C(\mathcal{R})} \rightarrow 0$ if and only if $(\forall i \in \overline{1, M}) \|\mathbf{f}_n^{\text{Lip}} - \mathbf{g}_i\|_{C(\mathcal{R})} \rightarrow 0$. Therefore, it suffices to show that $(i \in \overline{1, M}) g_i$, the limit of the sequence $(f_{i,n}^{\text{Lip}})_{n \in \mathbb{N}}$, is Lipschitz with L_i , the i th component of \mathbf{L} . Now, since $f_{i,n}^{\text{Lip}} \rightarrow f_i$ uniformly, for some $\epsilon > 0$ there exists $N_1^\epsilon \in \mathbb{N}$ such that $(\forall \mathbf{x} \in \mathcal{R}) |f_i(\mathbf{x}) - f_{i,N_1^\epsilon}^{\text{Lip}}(\mathbf{x})| < \epsilon$ which implies that there exists $N^\epsilon > N_1^\epsilon$ such that $(\forall \mathbf{x} \in \mathcal{R}) |f_i(\mathbf{x}) - f_{i,N^\epsilon}^{\text{Lip}}(\mathbf{x})| < \epsilon/2$. Then,

$$\begin{aligned} (\forall \mathbf{x} \in \mathcal{R}) (\forall \mathbf{y} \in \mathcal{R}) |f_i(\mathbf{x}) - f_i(\mathbf{y})| &= |f_i(\mathbf{x}) + f_{i,N^\epsilon}^{\text{Lip}}(\mathbf{x}) \\ &\quad - f_{i,N^\epsilon}^{\text{Lip}}(\mathbf{x}) + f_{i,N^\epsilon}^{\text{Lip}}(\mathbf{y}) \\ &\quad - f_{i,N^\epsilon}^{\text{Lip}}(\mathbf{y}) + f_i(\mathbf{y})| \\ &< \epsilon/2 + \epsilon/2 + L_i \|\mathbf{x} - \mathbf{y}\| \\ &= \epsilon + L_i \|\mathbf{x} - \mathbf{y}\|. \end{aligned}$$

Since the above holds for all $\epsilon > 0$, it follows that

$$(\forall \mathbf{x} \in \mathcal{R}) (\forall \mathbf{y} \in \mathcal{R}) |f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq L_i \|\mathbf{x} - \mathbf{y}\|.$$

\square

F. Monotone Smoothing of the Sample set

We consider the monotone-smoothing problem which is formulated as a standard convex optimization problem. The author in [28] has shown that a sample set $\mathcal{D}^{\text{com}} := \{(\mathbf{r}^k, \tilde{\rho}^k)\}_{k=1}^K$ is compatible with the monotonicity if and only if it satisfies the following set of linear constraints [28, Proposition 4.1]

$$(\forall k \in \overline{1, K}) (\forall j \in \overline{1, K}) \tilde{\rho}^k - \tilde{\rho}^j \leq \tilde{L} \|(\mathbf{r}^k - \mathbf{r}^j)_+\|. \quad (16)$$

Given the measured sample set $\mathcal{D}^{\text{error}} = \{(\mathbf{r}^k, y^k)\}_{k=1}^K$, we look for a compatible set $\mathcal{D}^{\text{com}} = \{(\mathbf{r}^k, \tilde{\rho}^k)\}_{k=1}^K$ (that satisfies (16)) that is closest to $\mathcal{D}^{\text{error}}$ in the $\|\cdot\|_1$ sense. In more detail, let $\mathbf{y} = [y^1, \dots, y^K]^\top$ and $\tilde{\rho} = [\tilde{\rho}^1, \dots, \tilde{\rho}^K]^\top$, then we minimize

$$\|\mathbf{y} - \tilde{\rho}\|_1 = \sum_{k=1}^K |\tilde{\rho}^k - y^k|. \quad (17)$$

We now formalize this problem as a standard linear program (LP) which can be solved easily by any standard convex solver. Denote the k th residual in (17) by $q^k := \tilde{\rho}^k - y^k$ and split q^k into two parts q_+^k and q_-^k such that $q^k = q_+^k - q_-^k$. Substituting

($\forall l \in \overline{1, K}$) $q^l + y^l$ for $\tilde{\rho}^l$ into (16) and (17), the monotone-smoothing problem can be written as an LP [28]

$$\begin{aligned} & \underset{q_+^k, q_-^k \geq 0}{\text{minimize}} && \sum_{k=1}^K |q^k| \\ & \text{subject to} && (\forall k \in \overline{1, K}) \quad (\forall j \in \overline{1, K}) \\ & && q^k - q^j \leq y^j - y^k + \tilde{L} \|(\mathbf{r}^k - \mathbf{r}^j)_+\|, \end{aligned} \quad (18)$$

where $|q^k| = q_+^k + q_-^k$, and where $q_+^k, q_-^k \geq 0$ are the optimization variables. The smoothed compatible values follow from $\tilde{\rho}^k = y^k + q^k$.

Note that since we consider very small sample sizes K and the constraint matrix, with rows given by (18), is sparse, the above LP can be solved efficiently with standard convex solvers that exploit sparsity [43]. Therefore, the complexity of the smoothing step, which is performed only once after sample acquisition, is not of a practical concern.

ACKNOWLEDGMENT

The work was supported by the German Federal Ministry of Education and Research under grant 16KIS0605. This work is also supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF) in the framework of the project 5G NetMobil with funding number 16KIS0691. The authors alone are responsible for the content of the paper.

REFERENCES

- [1] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, "A low complexity algorithm for proportional resource allocation in OFDMA systems," in *IEEE Workshop on Signal Processing Systems*, Oct 2004, pp. 1–6.
- [2] I. Siomina, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.
- [3] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for interference-coupled cellular data networks," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 5102–5107.
- [4] K. Majewski and M. Koonert, "Conservative cell load approximation for radio networks with Shannon channels and its application to LTE network planning," in *2010 Sixth Advanced International Conference on Telecommunications*, May 2010, pp. 219–225.
- [5] C. K. Ho, D. Yuan, and S. Sun, "Data offloading in load coupled networks: A utility maximization framework," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1921–1931, 2014.
- [6] P. Skillermarck and P. Frenger, "Enhancing energy efficiency in lte with antenna muting," in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, May 2012, pp. 1–5.
- [7] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Energy-efficient virtual base station formation in optical-access-enabled cloud-ran," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1130–1139, May 2016.
- [8] P. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugi, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, no. 1, 2007, pp. 1234–1238.
- [9] Y. Shen, "Load coupling model evaluation and feasibility study of power allocation in OFDMA networks," 2015.
- [10] R. L. G. Cavalcante, E. Pollakis, S. Stańczak, F. Penna, and J. Bühler, "GreenNets deliverables," GreenNets Project, FP7-SME.2011.1, Tech. Rep., 2013.
- [11] A. G. Sukharev, *Minimax Models in the Theory of Numerical Methods*. Norwell, MA, USA: Kluwer Academic Publishers, 1992.
- [12] J. Traub and H. Woźniakowski, *A general theory of optimal algorithms*, ser. ACM monograph series. Academic Press, 1980.
- [13] M. Golomb and H. Weinberger, *On Numerical Approximation*, R.E. Langer ed. The University of Wisconsin Press, Madison, 1959.
- [14] J.-P. Calliess, "Conservative decision-making and inference in uncertain dynamical systems," Ph.D. dissertation, Department of Engineering Science, University of Oxford, 2014.
- [15] G. Marcus, "Deep learning: A critical appraisal," *CoRR*, vol. abs/1801.00631, 2018. [Online]. Available: <http://arxiv.org/abs/1801.00631>
- [16] M. Diligenti, S. Roychowdhury, and M. Gori, "Integrating prior knowledge into deep learning," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 920–923.
- [17] L. You, D. Yuan, L. Lei, S. Sun, S. Chatzinotas, and B. Ottersten, "Resource Optimization With Load Coupling in Multi-Cell NOMA," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4735–4749, July 2018.
- [18] M. A. Gutierrez-Estevéz, R. L. G. Cavalcante, S. Stanczak, J. Zhang, and H. Zhuang, "A distributed solution for proportional fairness optimization in load coupled OFDMA networks," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP' 16)*, 2016.
- [19] D. Awan, R. L. G. Cavalcante, and S. Stanczak, "Distributed RAN and backhaul optimization for energy efficient wireless networks," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP' 16)*.
- [20] E. Pollakis, R. L. G. Cavalcante, and S. Stanczak, "Traffic demand-aware topology control for enhanced energy-efficiency of cellular networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, pp. 1–17, 2016.
- [21] Z. Ren, S. Stanczak, and P. Fertl, "Activation of nomadic relay nodes in dynamic interference environment for energy saving," *2014 IEEE Global Communications Conference*, pp. 4466–4471, 2014.
- [22] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 1357–1361.
- [23] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem*. Boston(MA): Birkhauser, 2003.
- [24] I. Siomina and D. Yuan, "Optimizing small-cell range in heterogeneous and load-coupled LTE networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 5, pp. 2169–2174, May 2015.
- [25] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, Sep 1995.
- [26] K. Sun, S. Mou, J. Qiu, T. Wang, and H. Gao, "Adaptive fuzzy control for nontriangular structural stochastic switched nonlinear systems with full state constraints," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 8, pp. 1587–1601, Aug 2019.
- [27] J. Qiu, K. Sun, T. Wang, and H. Gao, "Observer-based fuzzy adaptive event-triggered control for pure-feedback nonlinear systems with prescribed performance," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 11, pp. 2152–2162, Nov 2019.
- [28] G. Beliaikov, "Monotonicity preserving approximation of multivariate scattered data," *BIT Numerical Mathematics*, vol. 45, no. 4, pp. 653–677, 2005.
- [29] W. Kotłowski, "Online isotonic regression," in *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23–26, 2016*, 2016, pp. 1165–1189.
- [30] D. A. Awan, R. Cavalcante, and S. Stanczak, "A robust machine learning method for cell-load approximation in wireless networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.
- [31] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.
- [32] J. Munkres, *Topology (Second Edition)*. Prentice Hall, Inc., 2000.
- [33] R. F. Brown, *A Topological Introduction To Nonlinear Analysis*. Birkhauser Basel, 2014.
- [34] R. L. G. Cavalcante, Y. Shen, and S. Stanczak, "Elementary properties of positive concave mappings with applications to network planning and optimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1774–1783, April 2016.
- [35] G. G. Belford, "Uniform approximation of vector-valued functions with a constraint," *Mathematics of Computation*, vol. 26, no. 118, pp. 487–492, 1972.
- [36] G. Beliaikov, "Interpolation of Lipschitz functions," *Journal of Computational and Applied Mathematics*, vol. 196, no. 1, pp. 20 – 44, 2006.
- [37] M. Milanese and R. Tempo, "Optimal algorithms theory for robust estimation and prediction," *IEEE Transactions on Automatic Control*, vol. 30, no. 8, pp. 730–738, August 1985.

- [38] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: An overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, Nov. 1991. [Online]. Available: [http://dx.doi.org/10.1016/0005-1098\(91\)90134-N](http://dx.doi.org/10.1016/0005-1098(91)90134-N)
- [39] R. G. Strongin, "On the convergence of an algorithm for finding a global extremum," *Engineering in Cybernetics*, 1973.
- [40] "The network simulator NS3," <https://www.nsnam.org/>, accessed: 2018-07-13.
- [41] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 2nd ed. Orlando, FL, USA: Academic Press, Inc., 2014.
- [42] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, 1994.
- [43] L. Liberti, P. Poirion, and K. Vu, "Fast approximate solution of large dense linear programs," http://www.optimization-online.org/DB_FILE/2016/11/5737.pdf, accessed: 2018-07-13.