Machine Learning Methods for RSS-Based User Positioning in Distributed Massive MIMO

K. N. R. Surya Vara Prasad, Ekram Hossain, *Fellow, IEEE*, and Vijay K. Bhargava, *Life Fellow, IEEE*

***Abstract*— We propose a supervised machine learning (ML) approach based on Gaussian process (GP) regression to position users in a distributed massive multiple-input multiple-output (DM-MIMO) system from their uplink received signal strength (RSS). The proposed approach serves as a *proof-of-concept* that we can localize users by training an ML model with noise-free RSS and using the trained model to estimate the test user locations from their noisy RSS. We consider two GP methods for localization, namely, the conventional GP (CGP) and the numerical approximation GP (NaGP). We find that the CGP provides unrealistically small 2σ error-bars on the location estimates. Therefore, we derive the true predictive distribution and employ NaGP to obtain realistic 2σ error-bars on the location estimates. Next, we derive a Bayesian Cramer–Rao lower bound (BCRLB) on the root-mean-squared-error (RMSE) performance of the two GP methods. Numerical studies reveal that: 1) the NaGP indeed provides realistic 2σ error-bars on the estimated locations; 2) both the CGP and NaGP achieve RMSEs that are close to the BCRLBs; 3) the presence of correlated shadowing improves the RMSE performance; and 4) extrapolation to the zero input noise scenario can significantly improve the RMSE achieved by the NaGP.**

***Index Terms*— Massive MIMO, localization, machine learning (ML), Gaussian process (GP) regression, Cramer–Rao bounds.**

I. Introduction

WIRELESS user positioning is an important research direction for the fifth generation (5G) networks because location information can be utilized to provide context-aware communication services. For example, approximate location information can facilitate area-specific advertisements, content caching, and personnel tracking under emergency calls. Satellite-based global positioning systems (GPSs) [2], which are currently being used in LTE to procure location information, do not provide reliable location estimates for indoor and users. They also require power-hungry sensors to be active on the user devices. Consequently, there is ongoing research on alternative local positioning systems, which use wireless signal information, such as angle-of-arrival (AOA), time of arrival (TOA), and received signal strength (RSS), to locate the users.

The massive multiple-input multiple-output (MIMO) technology [3], [4], which operates with a large number of antennas at the base station (BS), opens up new opportunities for machine learning in wireless positioning. Due to the large number of antennas, massive MIMO allows the BS to record large vectors of RSS on the uplink. Machine learning (ML) techniques can then be employed for user positioning, wherein we train an ML model with a database comprising RSS vectors for several known user locations. The RSS of a test user is then fed as input to the trained ML model, so as to obtain its location estimate.

In this work, we consider a distributed massive MIMO (DM-MIMO) system, where a large number of distributed single-antenna radio remote heads, connected via high-speed fronthaul links to a computing unit, serve multiple users simultaneously on the same time-frequency resource. As pointed out in few recent works [6], DM-MIMO systems provide high coverage probability and also uniform throughput service to all the users, with significant gains in the 95%-likely per-user throughput when compared to small cell schemes. When compared to co-located massive MIMO systems, DM-MIMO can also offer higher sum-throughput rate [5], average throughput rate [7], and bit-per-joule energy efficiency [8]. To achieve such performance gains, DM-MIMO systems make efficient use of spatial resources through the use of distributed antennas, which comes at the expense of increased deployment costs, fronthaul overhead, and complexity in channel modelling.

For a distributed massive MIMO setup with single-antenna remote radio heads, we propose a new machine learning approach based on Gaussian process regression (GP) to predict user locations from their uplink RSS. We choose RSS as the signal property for user positioning because RSS measurements are readily available at the BS, without the need for extra hardware to be installed. A major difficulty with using RSS for user positioning is that the RSS is corrupted by small-scale fading and shadowing effects of the wireless channel. Small scale fading can be mitigated by averaging over multiple timeslots [10], [11], but shadowing mitigation requires spatial averaging, which in turn relies on prior knowledge of the user location [10]. Since we are not aware of the test users' locations, we cannot therefore average out the shadowing noise in the test RSS. However, for training, we can synthetically generate noise-free RSS data.

Manuscript received January 5, 2018; revised October 7, 2018; accepted October 11, 2018. Date of publication October 25, 2018; date of current version December 10, 2018. This work was supported by the Natural Sciences and Engineering Research Council of Canada. This paper was presented in part at the 11th IEEE International Conference on Advanced Networks and Telecommunications Systems, Bhubaneswar, India, December 2017 [1]. The associate editor coordinating the review of this paper and approving it for publication was R. Dinis. *(Corresponding author: K. N. R. Surya Vara Prasad.)*

K. N. R. S. V. Prasad and V. K. Bhargava are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: surya@ece.ubc.ca; vijayb@ece.ubc.ca).

E. Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada (e-mail: ekram.hossain@umanitoba.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2876832

With the above constraint in mind, we investigate user positioning when the training RSS data is noise-free and the test RSS data is noisy due to shadowing. Firstly, we consider the conventional GP (CGP) method, which naively treats the test RSS data as noise-free for location prediction. Our simulation studies reveal that the CGP method provides unrealistically small 2σ error-bars on the predicted locations. To address this limitation, we consider using a moment-matching GP method, referred to in our work as the numerical approximation GP (NaGP) method. The NaGP method provides realistic 2σ error-bars on the estimated locations by learning from the statistical properties of the noise present in the test RSS.

The main contributions of our work are:

(i) For user positioning in massive MIMO, our work serves as a *proof-of-concept* that we can train a machine learning (ML) model with noise-free RSS and use the trained ML model to position users from their noisy RSS. Unlike most existing methods, we derive closed-form expressions for the estimated locations and their 2σ error-bars.

(ii) For the machine learning problem of user localization with noise-free training RSS and noisy test RSS in distributed massive MIMO systems, ours is the first work to identify that the CGP method provides unrealistically small 2σ error-bars on the estimated locations. To derive realistic 2σ error-bars, we are also the first to apply the NaGP method, which is a popular moment-matching GP method in time-series analysis [12].

(iii) When using GP methods for user positioning, the covariance between $x-$coordinates (or $y-$coordinates) needs to be modelled as a function of the RSS vectors. We analyze the mathematical relation between RSS and user locations and propose to use a weighted-sum covariance model comprising squared exponential and inner-product terms.

(iv) We derive a Bayesian Cramer-Rao lower bound (BCRLB) on the achievable RMSE performance of the two GP methods under study.

(v) We demonstrate the impact of correlated shadowing on RSS-based user positioning in distributed massive MIMO. When GP methods are used, we observe that the RMSE performance only improves in the presence of correlated shadowing and therefore, the case of uncorrelated shadowing represents the worst-case setup to study the localization.

(vi) We provide an example on how extrapolation may be employed to significantly improve the RMSE performance achieved by the NaGP method.

The paper is organized as follows. In Section II, we present a review of related works. In Section III, we present the system model for user positioning. In Section IV, details are presented on the training phase of the two GP methods under study. The CGP and NaGP methods are presented in Sections V-VI. Performance metrics are presented in Section VII and a Bayesian Cramer-Rao lower bound is derived on the achievable RMSE performance. Numerical studies are presented in Section VIII, followed by few concluding remarks in Section IX.

General Notation: We use regular font small letters for scalars, boldface small letters for vectors, and boldface capital letters for matrices, for example, a, $\mathbf{a}$, and $\mathbf{U}$, respectively. The notations $[\mathbf{a}]_i$, $[\mathbf{U}]_i$, and $[\mathbf{U}]_{ij}$ refer to the element i in vector $\mathbf{a}$, column i in matrix $\mathbf{U}$, and the element (i,j) in matrix $\mathbf{U}$, respectively. The overhead symbol $\widetilde{(.)}$ refers to training data and the overhead symbol $\widehat{(.)}$ refers to test data, respectively. An additional superscript $(.)^*$ is used to refer to the noise-free component in the data. The symbol $\approx$ denotes approximation. A random vector $\mathbf{a}$ that is Gaussian distributed with mean $\mathbf{u}$ and covariance $\mathbf{U}$ is referred to as $\mathbf{a} \sim \mathcal{N}(\mathbf{u}, \mathbf{U})$, and its probability density function is denoted as $\mathcal{N}(\mathbf{a}; \mathbf{u}, \mathbf{U})$.

II. Related Work

A. User Positioning in Massive MIMO

User positioning in massive MIMO is a recent research topic. The authors in [13] propose a compressed sensing approach to estimate user locations from TOA information recorded at multiple massive MIMO BSs. AOA information is used in [14]–[16], while the combined information of time delay, AOD, and AOA information is used in [16] and [17] for positioning users in massive MIMO. A millimeter wave massive MIMO system is considered in [16] to derive the necessary conditions under which a user location can be estimated from AOD, AOA, and TOA information under line-of-sight conditions. In [18], a highly directional 60 GHz millimeter wave network is considered and an environment-sensing method is proposed to estimate the reflector locations from the AOA, AOD, and path length information. For accurate AOD, AOA and/or TOA based localization, all of the above methods rely on the information recorded at adjacent antennas in a co-located MIMO setup, where each base station hosts an array of antennas. The localization becomes unreliable for the distributed massive MIMO system in our work because we consider single-antenna remote radio heads. The most related work to our study is [19], where RSS-based user positioning is investigated, but noisy RSS data is considered for both training and test purposes. Training with noisy RSS can be quite challenging because extensive and recurrent measurement campaigns would be needed to build the training database. Therefore, unlike [19], we consider noise-free RSS for training and noisy RSS for testing. We also make two improvements over [19]. Firstly, we identify that CGP method, which is also employed in [19], provides unrealistically small 2σ error-bars on the location estimates. Secondly, to address this concern, we apply a moment-matching based GP method, namely the NaGP, to position the users from their RSS. NaGP provides realistic 2σ error-bars on the estimated locations.

B. Machine Learning Techniques for User Positioning

Wireless source localization is a widely investigated topic. Traditionally, ranging-based methods such as [10] and [21] are employed, wherein, a single source is localized with the help of signal information, such as TOA, AOA and/or RSS recorded at a number of sensors. TOA-based methods, such as [20], require tight timing synchronization between the source and sensors and are therefore complicated and expensive. AOA-based methods require multiple antennas to

be deployed at each sensor. RSS-based methods, such as [21], typically assume that the path-loss exponent is known and use it to estimate the distance between source and sensors from the RSS. Trilateration is then employed to estimate the source location. Such an estimation procedure typically involves multiple matrix inversions, which can quite often result in numerical noise because the matrices to be inverted have low condition number. Moreover, we notice that there are no current works which extend the above methods to multi-source systems, such as massive MIMO, operating in multi-slope path-loss regions. The main added challenge here is that the localization needs to handle multiuser interference. Keeping these limitations in mind, we propose to take a supervised machine learning approach to position multiple sources simultaneously from their RSS. Unlike the above works, we estimate the source locations directly from their RSS by training and building a non-linear regression model. Such a data-driven approach allows us to localize users even though the signals experience multi-slope path-loss with non-line-of-sight conditions.

Several machine learning techniques, including, k-nearest neighbours [28], GP methods [19], and more recently, deep learning methods [36], [37], have been explored for wireless user positioning. From these techniques, we choose GP methods for our analysis for four reasons, as stated in Remark 2.

Localization methods proposed for WiFi systems, such as in [29]–[36] and references therein, do not extend in a straightforward manner to massive MIMO systems. This is because the localization in WiFi systems is fundamentally different from ours. Firstly, WiFi-based localization methods do not consider multi-user MIMO transmissions on the uplink and therefore, do not account for the associated inter-user interference. Secondly, most RSS-based WiFi localization methods (see [29]–[32] and references therein) focus the downlink, wherein, the users need to handle the computational cost of estimating their own locations. Whereas, in our system, localization is focused on the uplink and the base station estimates the user locations. Lastly, several AoA-based WiFi localization systems [33]–[35] operate on the uplink, but assume the presence of multiple co-located antennas at the base station. Such methods do not work well for single-antenna receiver systems such as the distributed MIMO system considered in our work.

Coming to GP methods for user positioning in wireless networks, we observe that most of the existing works [29], [40], [41] have opted for an indirect modelling approach, wherein the GP models take location estimates as inputs and provide RSS as outputs. Our work is different from the above in three ways. Firstly, above works take an indirect modelling approach, wherein, each user trains at least one GP model per base station. Such a training policy is not suitable for massive MIMO because we operate with a large number of Bs antennas and therefore, each user is required to train a large number of GP models. In contrast, our approach only requires a total of two GP models to be trained - one each for the x and y coordinates. Secondly, above works advocate the use of conventional GP methods, which do not provide realistic 2σ error-bars on the estimated locations. Lastly, in the above works, the users bear all the computational load for both training and prediction. In our work, the BS handles all the computational load for localization. Since the computations can be particularly high when training, our approach is more appealing for user devices operating with battery constraints.

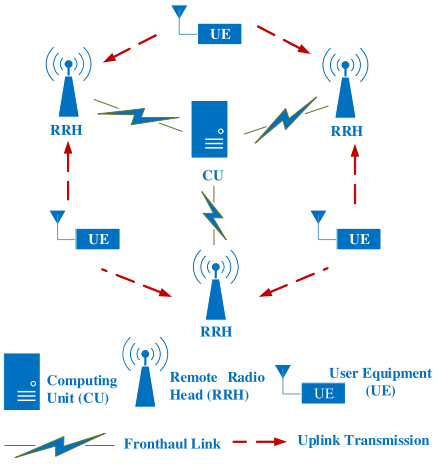


Fig. 1. Setup for user positioning in distributed massive MIMO: K single-antenna UEs transmit uplink signals simultaneously to M RRHs on the same time-frequency resource. Each RRH records forwards its RSS to the CU via high-speed fronthaul. The CU hosts a machine learning model which takes the RSS vectors as input and provides the users' location estimates as output.

Few previous works have studied noisy inputs in GP. The authors in [12], [42], and [43] propose moment-matching based GP methods for multi-step prediction in time series analysis, system identification and channel prediction respectively, but not for user positioning in massive MIMO. Also, unlike above works, we derive a Bayesian Cramer-Rao lower bound on the RMSE of the considered GP methods.

III. System Description

We consider a distributed multiuser massive MIMO setup, as shown in Fig. 1, where K user equipments (UEs) transmit uplink signals to M remote radio heads (RRHs) simultaneously on the same time-frequency resource. For simplicity, we assume that the RRHs are all single-antenna units and refer to them interchangeably as BS antennas.[1] We also assume that all the UEs are single-antenna units and refer to them as users. The RRHs are connected to a computing unit (CU) through high-speed fronthaul. When the K users transmit on the uplink, each RRH records its own received signal strength (RSS). The CU gathers the RSS from each RRH, processes them to extract the per-user RSS, and forms an $M \times 1$ RSS vector for each user. The RSS vectors thus formed are fed as input to a trained ML model for predicting the user locations. The CU hosts the ML model and handles the computations required for training and prediction.

[1]Extensions to the multiple-antenna scenario may consider augmenting RSS information with AoA information for improved localization accuracy because multiple-antenna receivers provide a very good setting for AoA-based localization [33]

A. Multi-User Transmissions

When user k transmits a symbol vector $\boldsymbol{\omega}_k$ with power ρ, the RRH m receives $\mathbf{r}_m$, given by

$$\mathbf{r}_m = \sqrt{\rho}\sum_{k=1}^{K} h_{mk}\boldsymbol{\omega}_k + \boldsymbol{\vartheta}_m, \tag{1}$$

where $h_{mk} = q_{mk}\sqrt{g_{mk}}$ is the flat-fading uplink channel gain with q_{mk} and g_{mk} being the small-scale and large-scale fading coefficients, and $\boldsymbol{\vartheta}_m \sim \mathcal{N}(\mathbf{0}, \sigma_\vartheta^2\mathbf{I})$ is the additive white Gaussian noise vector. The received vector $\mathbf{r}_m$ has the same dimension as that of the transmitted symbol vector $\boldsymbol{\omega}_k$. We assume that the small-scale fading coefficients q_{mk} are i.i.d complex Gaussian, i.e., $q_{mk} \sim \mathcal{CN}(0,1)$. If d_{mk} is the distance between user k and RRH m, b_0 is the path-loss at reference distance d_0, η is the path-loss exponent, and $z_{mk} \sim \mathcal{N}(0, \sigma^2_{z_{mk}})$ is the channel gain due to shadowing noise, we model the large-scale fading coefficient g_{mk} as

$$g_{mk} = b_0 d_{mk}^{-\eta} 10^{\frac{z_{mk}}{10}}. \tag{2}$$

B. Extracting Per-User RSS Values

From (1), we note that the RSS $||\mathbf{r}_m||^2$ at RRH m corresponds to the multiuser RSS because the received vector $\mathbf{r}_m$ is the sum of symbol vectors received from all K users. We cannot directly use the multiuser RSS $||\mathbf{r}_m||^2$ to position any given user k because we would then be unable to distinguish among the K users that are transmitting simultaneously. Instead, the RRH m should extract the per-user RSS p_{mk} of each user k from $\mathbf{r}_m$ and use it for positioning the user k. This can be done if the symbol vectors $\{\boldsymbol{\omega}_k\}$ in (1) are mutually orthogonal in time-frequency and are already known at the RRH,[2] for example, $\{\boldsymbol{\omega}_k\}$ can be pilot sequences transmitted for channel estimation [6]. The RSS p_{mk} of user k can then be obtained from $\mathbf{r}_m$ as

$$p_{mk} = \rho g_{mk}|q_{mk}|^2. \tag{3}$$

Observe from (3) that the per-user RSS can be noisy due to small-scale fading and shadowing of the wireless channel. We assume that small-scale fading is averaged out over multiple time-slots and focus on the case where only the shadowing noise exists. We do so because shadowing is space-dependent and requires access to the user location in order to be averaged out. The resulting RSS, after substituting the large-scale fading model in (2), is given in dB scale as

$$p_{mk}^{\text{dB}} = p_0^{\text{dB}} - 10\eta\log_{10}(d_{mk}) + z_{mk}, \tag{4}$$

where $p_0^{\text{dB}} = 10\log_{10}(\rho b_0)$ is the uplink RSS at the reference distance d_0. For each user k, the CU can then form an $M \times 1$ RSS vector $\mathbf{p}_k$ such that $[\mathbf{p}_k]_m = p_{mk}^{\text{dB}}$, i.e.,

$$\mathbf{p}_k = [p_{1k}^{\text{dB}} \;\; p_{2k}^{\text{dB}} \;\; \ldots \;\; p_{Mk}^{\text{dB}}]^T. \tag{5}$$

Remark 1 (On the Use of Machine Learning With Noise-Free Training Inputs): A major concern with the use of machine learning for user positioning is the availability of appropriate training data. Particularly, training with noisy RSS can be difficult because we would need a diverse training dataset that accommodates the signal variations across multiple environments and time periods. To circumvent this challenge, we propose to use noise-free RSS for training purposes. Instead of relying on extensive measurement campaigns, we can generate noise-free RSS datasets synthetically via software simulations. To do so, we would only require knowledge of the training locations (which we can randomly choose within the service area), the transmission power (which can be obtained from the transmission protocol), and the path-loss exponents (which are generally stable and can be obtained through calibration campaigns [44]). Our work serves as proof-of-concept that we can indeed obtain location estimates, as well as their 2σ error-bars, through machine learning with noise-free training inputs. A natural downside with the proposed approach is that the test RSS is generally noisy and this mismatch hampers the localization performance. To address this issue, in Section VIII, we present an illustrative example on how we may add artificial noise to the test RSS vectors and extrapolate the resulting location estimates to the zero-noise scenario for improvement in the localization performance.

C. Machine Learning Model

Let us define $f_x(.)$ and $f_y(.)$ as the functions which map the RSS vector $\mathbf{p}_k$ of any user k in the system to its 2D location coordinates (x_k, y_k), such that

$$x_k = f_x(\mathbf{p}_k) \;\; \text{and} \;\; y_k = f_y(\mathbf{p}_k) \;\; \forall x_k, y_k. \tag{6}$$

We employ supervised machine learning to learn the functions $f_x(.)$ and $f_y(.)$, wherein we first train a machine learning model with RSS vectors for several known user locations. The trained model is then fed with RSS vectors of test users as inputs, so as to obtain their location estimates.

For the machine learning task, we adopt two Gaussian process regression (GP) methods from time-series analysis, namely, the conventional GP (CGP) [9] and the numerical approximation GP (NaGP) [12] methods. Both the GP methods employ the same training procedure, but differ in terms of how the test user locations are predicted. Details on the training are presented first.[3]

Remark 2 (On the Choice of Gaussian Process Regression to Estimate User Locations): The problem of estimating the $x-$coordinate x_k (and $y-$coordinate y_k) of any user k from its uplink RSS vector $\mathbf{p}_k$ is a non-linear regression problem in machine learning. This is because x_k is a non-linear

[2]In order to employ mutually orthogonal pilot sequences, the length of each pilot sequence $\boldsymbol{\omega}_k$ needs to be greater than or equal to the number of users K being served simultaneously. This is very much feasible, as is illustrated in the following example. Let us consider an outdoor scenario with mobility up to $v = 37.5$ m/s. For a carrier frequency of $f_c = 2$ GHz and the speed of light $c = 3 * 10^8$m/s, the coherence time, given by $\frac{c}{4*f_c*v}$, is $T_c = 1$ms. Also, for a delay spread of $T_d = 2.5$ μs, the coherence bandwidth, given by $\frac{1}{2*T_d}$, is $B_c = 200$kHz. The coherence block therefore contains $T_c * B_c = 200$ symbols. If we allocate 30% of the coherence block for pilot transmissions and the remaining 70% for data transmissions, we can create $0.3 * 200 = 60$ orthogonal pilots and can therefore localize up to 60 users simultaneously.

[3]For simplicity, we focus on the $x-$coordinates only, but the same procedure can be applied for $y-$coordinates as well.

function of $\mathbf{p}_k$ with the additional presence of shadowing noise z_{mk} (c.f. (4)). From among several non-linear regression methods, we choose Gaussian process regression for our study because (i) GP methods are non-parametric and can be used to model any arbitrary function [9], (ii) GP methods are known to perform as good as most other machine learning methods in terms of multiple performance metrics, including the squared prediction error [38], [39], the correlation between prediction error and the 2σ error-bars provided by the ML model [38], the ability to learn from the training data [39], and the generalizability of the trained ML model [39], (iii) GP methods provide probabilistic estimates in the form of a full predictive distribution, wherein we obtain closed-form expressions for the estimated parameter (location coordinates in our case), as well as the 2σ error-bars on the estimates, and (iv) unlike most machine learning methods, including deep learning methods [36], [37], GP methods allow us to derive Cramer-Rao type bounds on the prediction performance.

IV. Training Phase of the GP Methods

At the core of all GP methods is the assumption that the function to be learned, i.e., $f_x(.)$, is drawn from a zero-mean Gaussian process prior specified by a user-defined covariance function $\phi(.,.)$ [9]. This means that any finite number of $f_x(.)$ realizations are assumed to follow a joint Gaussian distribution with mean zero and covariance $\mathbf{\Phi}$, whose elements are given by the function $\phi(.,.)$. We refer to this assumption as

$$f_x(.) \sim \mathcal{GP}(0, \phi(.,.)). \tag{7}$$

The function $\phi(.,.)$ models the covariance of $x-$coordinates of any two users in the system as a function of their RSS vectors. We choose $\phi(.,.)$ as the weighted-sum of squared-exponential (SE), inner product (IP) and delta functions, given for any two RSS vectors $\mathbf{p}_k$ and $\mathbf{p}_{k'}$, as

$$\begin{aligned}\phi(\mathbf{p}_k, \mathbf{p}_{k'}) = \alpha e^{-\frac{1}{2}(\mathbf{p}_k-\mathbf{p}_{k'})^T\mathbf{B}^{-1}(\mathbf{p}_k-\mathbf{p}_{k'})} + \gamma \mathbf{p}_k^T\mathbf{p}_{k'} + \sigma_{er}^2\delta_{kk'},\\ \text{where} \quad \mathbf{B} = \mathbf{diag}\{\beta_m\}, m = 1, \ldots, M, \quad \text{and}\\ \delta_{kk'} = \{1 \text{ if } k = k', 0 \text{ if otherwise}\}.\end{aligned} \tag{8}$$

In (8), the SE term $\alpha e^{-\frac{1}{2}(\mathbf{p}_k-\mathbf{p}_{k'})^T\mathbf{B}^{-1}(\mathbf{p}_k-\mathbf{p}_{k'})}$ captures the dependence of $\phi(\mathbf{p}_k, \mathbf{p}_{k'})$ on the distance between the RSS vectors $\mathbf{p}_k$ and $\mathbf{p}_{k'}$. The IP term $\gamma\mathbf{p}_k^T\mathbf{p}_{k'}$ captures the dependence of $\phi(\mathbf{p}_k, \mathbf{p}_{k'})$ on the actual RSS vectors $\mathbf{p}_k$ and $\mathbf{p}_{k'}$. The delta term $\sigma_{er}^2\delta_{kk'}$ captures the variance due to measurement errors. The rationale behind our choice of $\phi(.,.)$ is as follows.

On the Choice of Covariance Function $\phi(.,.)$

In Fig. 2, we have considered an example setup to motivate our choice of the covariance model. Any pair of users in our system can be categorized into three feasible types: (i) Type I: the users are closely-spaced in both $x-$coordinate space $\mathcal{X}$ and RSS space $\mathcal{P}$. The users U1 and U2 in Fig. 2 fall under this category. (ii) Type II: the users are well-separated in both $\mathcal{X}$ and $\mathcal{P}$. The users U1 and U5 fall under this category. (iii) Type III: the users are closely-spaced in $\mathcal{X}$ but are well-separated in $\mathcal{P}$. The users U2 and U3 fall under this category because their $y-$coordinates are well-separated.

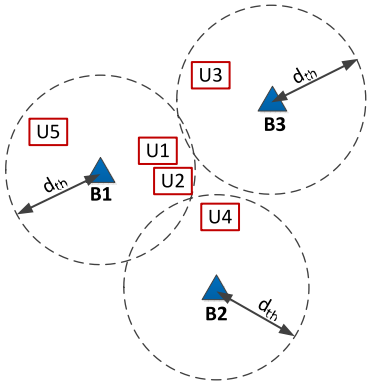


Fig. 2. Example setup with three RRHs, labelled B1, B2, and B3, and five users, labelled U1, U2, U3, U4, and U5. The radius d_{th} is the distance beyond which η changes in value (e.g., d_{th} = 45m in 3GPP UMi [22] because η changes from 2 to 6.7). For any pair of users, $\phi(.,.)$ should be high if the $x-$coordinates are closely-spaced and it should decrease with the separation.

Also, the users U2 and U4 fall under the same category because their path loss exponents are different.

Our choice of $\phi(.,.)$ needs to accommodate the three feasible types of user pairs mentioned above. Firstly, $\phi(.,.)$ needs to generate entries of a positive definite matrix because the resulting matrix is the covariance of a Gaussian distribution. Secondly, the $\phi(.,.)$ should exhibit local stationarity, i.e., for a given pair of users, the covariance between their $x-$coordinates should decrease with the distance between their RSS vectors. This will allow $\phi(.,.)$ to accommodate user pairs of Type I and II. The squared exponential function in an example stationary covariance function [9]. Lastly, the $\phi(.,.)$ should also exhibit some non-stationarity, i.e., the $\phi(.,.)$ model should allow users to have high covariance between their $x-$coordinates even if they are well-separated in the RSS space. This will allow $\phi(.,.)$ to accommodate user pairs of Type III. The inner-product function is an example non-stationary covariance function [9]. We therefore choose $\phi(.,.)$ as the weighted-sum of squared exponential and inner-product covariance functions. (8).

Our objective in the training phase is to train a GP model to learn the function $f_x(.)$. For this, it is sufficient to learn the free parameters introduced by (8) because $f_x(.)$ is fully specified by $\phi(.,.)$. Let us accumulate the free parameters in (8) into an $(M+2)\times 1$ vector $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = [\alpha \ \beta_1 \ \ldots \ \beta_M \ \gamma]^T. \tag{9}$$

Let us say we have $\widetilde{L}$ training locations. The $\widetilde{L}\times 1$ training $x-$coordinate vector $\widetilde{\mathbf{x}}$ and the corresponding $\widetilde{L}\times M$ noise-free training RSS matrix $\widetilde{\mathbf{P}}$ are defined as follows

$$\begin{aligned}\widetilde{\mathbf{x}} &= [\widetilde{x}_1 \ \widetilde{x}_2 \ \ldots \widetilde{x}_{\widetilde{L}}]^T,\\ \widetilde{\mathbf{P}} &= [\widetilde{\mathbf{p}}_1 \ \widetilde{\mathbf{p}}_2 \ \ldots \widetilde{\mathbf{p}}_{\widetilde{L}}]^T,\end{aligned} \tag{10}$$

where the row l in $\widetilde{\mathbf{P}}$, i.e., the training RSS vector $\widetilde{\mathbf{p}}_l$, corresponds to the training $x-$coordinate $\widetilde{x}_l$, $\forall\ l = 1, \ldots, \widetilde{L}$.

From the GP assumption, we know that $\widetilde{\mathbf{x}}$ is Gaussian distributed as

$$\begin{aligned}\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Phi}}), \quad \text{where}\\ [\widetilde{\mathbf{\Phi}}]_{ll'} = \phi(\widetilde{\mathbf{p}}_l, \widetilde{\mathbf{p}}_{l'}), \quad \forall\ l, l' = 1, \ldots, \widetilde{L}.\end{aligned} \tag{11}$$

From eq. (11), we can obtain the log-likelihood expression of $\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}},\boldsymbol{\theta}$. Subsequently, we can learn the vector $\boldsymbol{\theta}$ via maximum-likelihood as

$$\bar{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log(p(\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}},\boldsymbol{\theta})), \tag{12}$$

where $\bar{\boldsymbol{\theta}}$ is the learned parameter vector. The optimization problem in (12) is non-convex, but can be solved for a local optimum using gradient ascent methods, such as conjugate gradient [48]. Solving (12) for $\bar{\boldsymbol{\theta}}$ completes the training phase.

Next, in the prediction phase, let there be $\widehat{L}$ test users whose location coordinates need to be predicted. We now introduce the $\widehat{L}\times 1$ vector $\widehat{\mathbf{x}}$ of the test user $x-$coordinates, which needs to be predicted from an $\widehat{L}\times M$ matrix $\widehat{\mathbf{P}}$ of the noisy test RSS vectors, defined such that

$$\begin{aligned}\widehat{\mathbf{P}} &= [\widehat{\mathbf{p}}_1 \;\; \widehat{\mathbf{p}}_2 \;\; \ldots \widehat{\mathbf{p}}_{\widehat{L}}]^T \equiv [\widehat{\mathbf{p}}_1^* \;\; \widehat{\mathbf{p}}_2^* \;\; \ldots \widehat{\mathbf{p}}_{\widehat{L}}^*]^T + [\widehat{\mathbf{z}}_1 \;\; \widehat{\mathbf{z}}_2 \;\; \ldots \widehat{\mathbf{z}}_{\widehat{L}}]^T \\ \widehat{\mathbf{x}} &= [\widehat{x}_1 \;\; \widehat{x}_2 \;\; \ldots \;\; \widehat{x}_{\widehat{L}}]^T,\end{aligned} \tag{13}$$

where the RSS vector $\widehat{\mathbf{p}}_l$ corresponds to the test user whose $x-$coordinate is $[\widehat{\mathbf{x}}]_l = \widehat{x}_l$, $\forall = 1,\ldots,\widehat{L}$. In (13), $\widehat{\mathbf{p}}_l^*$ refers to the noise-free component and $\widehat{\mathbf{z}}_l$ to the shadowing noise component in $\widehat{\mathbf{p}}_l$. We now proceed to present details on the CGP and NaGP methods, with focus on their prediction phase only because the training procedure is the same as detailed above.

V. Location Prediction With Conventional GP Method (CGP)

We now consider CGP - a GP method which employs conventional GP principles [9] to predict user locations. CGP treats the noisy test RSS vectors as noise-free and uses the GP assumption (7) to obtain the joint distribution of the training and test $x-$coordinate vectors $\widetilde{\mathbf{x}}$ and $\widehat{\mathbf{x}}$ as

$$\begin{bmatrix}\widetilde{\mathbf{x}} \\ \widehat{\mathbf{x}}\end{bmatrix} \Bigg| \widetilde{\mathbf{P}},\widehat{\mathbf{P}} \sim \mathcal{N}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\widetilde{\boldsymbol{\Phi}} & (\boldsymbol{\Phi}^\dagger)^T \\ \boldsymbol{\Phi}^\dagger & \widehat{\boldsymbol{\Phi}}\end{pmatrix}\right], \tag{14}$$

where $\widetilde{\boldsymbol{\Phi}} \in \mathbb{R}^{\widetilde{L}\times\widetilde{L}}$, $\boldsymbol{\Phi}^\dagger \in \mathbb{R}^{\widetilde{L}\times\widehat{L}}$, and $\widehat{\boldsymbol{\Phi}} \in \mathbb{R}^{\widehat{L}\times\widehat{L}}$ are the covariance matrices between the noise-free training and noisy test RSS vectors, defined such that $[\widetilde{\boldsymbol{\Phi}}]_{ll'} = \phi(\widetilde{\mathbf{p}}_l,\widetilde{\mathbf{p}}_{l'}), \forall l,l' = 1,\ldots,\widetilde{L}$, $[\boldsymbol{\Phi}^\dagger]_{ll'} = \phi(\widehat{\mathbf{p}}_l,\widetilde{\mathbf{p}}_{l'}), \forall l = 1,\ldots,\widehat{L}, l' = 1,\ldots,\widetilde{L}$, and $[\widehat{\boldsymbol{\Phi}}]_{ll'} = \phi(\widehat{\mathbf{p}}_l,\widehat{\mathbf{p}}_{l'}), \forall l,l' = 1,\ldots,\widehat{L}$.

Conditioning the joint distribution in (14) over $\widetilde{\mathbf{x}}$ gives us [9]

$$\begin{aligned}&\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}} \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}, \widehat{\mathbf{C}}_x^{(\mathrm{CGP})}), \quad \text{where} \\ &\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})} = \boldsymbol{\Phi}^\dagger\widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{x}}, \quad \widehat{\mathbf{C}}_x^{(\mathrm{CGP})} = \widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}^\dagger\widetilde{\boldsymbol{\Phi}}^{-1}(\boldsymbol{\Phi}^\dagger)^T.\end{aligned} \tag{15}$$

Eq. (15) gives us the predicted mean $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}$ and the associated covariance $\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}$ of the test $x-$coordinate vector $\widehat{\mathbf{x}}$ when CGP is employed. The distribution of the $x-$coordinate $[\widehat{\mathbf{x}}]_l$ of any particular test user l can be obtained, through marginalization of $\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}}$, as

$$\begin{aligned}&[\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l \sim \mathcal{N}([\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l, [\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}), \text{ where} \\ &[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l = \sum_{i=1}^{\widetilde{L}} \phi(\widehat{\mathbf{p}}_l,\widetilde{\mathbf{p}}_i)[\boldsymbol{\psi}]_i, \text{ where } \boldsymbol{\psi} = \widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{x}}, \text{ and} \\ &[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll} = \phi(\widehat{\mathbf{p}}_l,\widehat{\mathbf{p}}_l) - \sum_{i=1}^{\widetilde{L}}\sum_{j=1}^{\widetilde{L}} \phi(\widehat{\mathbf{p}}_l,\widetilde{\mathbf{p}}_i)[(\widetilde{\boldsymbol{\Phi}})^{-1}]_{ij}\phi(\widetilde{\mathbf{p}}_j,\widehat{\mathbf{p}}_l).\end{aligned} \tag{16}$$

In (16), the terms $[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l$ and $[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}$ refer to the predicted mean and variance of the $x-$coordinate $[\widehat{\mathbf{x}}]_l$ of any test user l. Since the predictive distribution is Gaussian, the predicted mean $[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l$ gives us the maximum-a-posteriori (MAP) estimate of $[\widehat{\mathbf{x}}]_l$. Also, the predictive variance bears a strong correlation with the empirical mean-squared prediction error on the test points [9], [38]. Therefore, $[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}$ serves as a direct indicator of the prediction error associated with the estimate $[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l$. The 2σ error-bars, calculated as $\pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}}$, provide us with the 95% confidence interval $\left[[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}}\right]$ on choosing $[\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}]_l$ as the estimate of $[\widehat{\mathbf{x}}]_l$. We have observed through simulation studies in Section VIII that the 2σ error-bars obtained from CGP are urealistically small. This issue can be overcome by the NaGP method presented next because, unlike CGP, it accounts for the noisy nature of the test RSS vectors.

VI. Location Prediction With Numerical Approximation GP Method

We now consider the numerical approximation GP method (NaGP), which is a moment matching GP method [12], to estimate the test user locations. This method exploits the stochastic nature of the noisy test RSS vectors to provide more realistic 2σ error-bars on the estimated locations than the CGP method. Details are as follows. We observe from (4) that any noisy RSS value p_{mk}^{dB} recorded at the RRH m is the sum of a noise-free component, i.e., $p_0^{\mathrm{dB}} - 10\eta\log_{10}(d_{mk})$, and a shadowing noise component, i.e., z_{mk}. This allows us to express the test RSS vector $\widehat{\mathbf{p}}_l$ as

$$\widehat{\mathbf{p}}_l = \widehat{\mathbf{p}}_l^* + \widehat{\mathbf{z}}_l, \quad \text{such that } \widehat{\mathbf{z}}_l \sim \mathcal{N}(\mathbf{0},\widehat{\boldsymbol{\Sigma}}_l), \tag{17}$$

where $\widehat{\mathbf{p}}_l^*$ is the noise-free component in $\widehat{\mathbf{p}}_l$ and $\widehat{\mathbf{z}}_l$ is the shadowing noise with covariance $\widehat{\boldsymbol{\Sigma}}_l$. For simplicity, we assume that $\widehat{\boldsymbol{\Sigma}}_l$ is a diagonal matrix, in other words, we assume that the M uplink channels of the test user l experience mutually independent shadowing. We also assume that the diagonal elements of $\widehat{\boldsymbol{\Sigma}}_l$ are already known to the CU. We then have the conditional distribution of $\widehat{\mathbf{p}}_l^*$ from (17) as

$$\widehat{\mathbf{p}}_l^*|\widehat{\mathbf{p}}_l,\widehat{\boldsymbol{\Sigma}}_l \sim \mathcal{N}(\widehat{\mathbf{p}}_l,\widehat{\boldsymbol{\Sigma}}_l). \tag{18}$$

We can now treat $\widehat{\mathbf{p}}_l^*$ as a hidden variable and use (16) to obtain an estimate of $[\widehat{\mathbf{x}}]_l$ in terms of $\widehat{\mathbf{p}}_l^*$. After this, we can

use (18) to integrate out the hidden variable $\widehat{\mathbf{p}}_l^*$ and obtain the true predictive distribution of $[\widehat{\mathbf{x}}]_l$ in terms of $\widehat{\mathbf{p}}_l$ as follows[4]

$$p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l) = \int p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l^*)p(\widehat{\mathbf{p}}_l^*|\widehat{\mathbf{p}}_l,\widehat{\mathbf{\Sigma}}_l)d\widehat{\mathbf{p}}_l^*, \quad (19)$$

where $p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l^*)$ is obtained from (16) and $p(\widehat{\mathbf{p}}_n^*|\widehat{\mathbf{p}}_n,\widehat{\mathbf{\Sigma}}_n)$ from (18), respectively. The predictive distribution $p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l)$ in (19) is non-Gaussian and cannot be obtained in closed-form because the integral on the right hand side is intractable. As a consequence, we can only obtain an approximation to the true predictive distribution $p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l)$.

The NaGP method takes a numerical approach and approximates $p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l)$ using Markov-Chain Monte-Carlo sampling [45] as follows. We draw S i.i.d samples $\widehat{\mathbf{p}}_l^*(s)$, $1 \leq s \leq S$, from $\widehat{\mathbf{p}}_l^*|\widehat{\mathbf{p}}_l,\widehat{\mathbf{\Sigma}}_l \sim \mathcal{N}(\widehat{\mathbf{p}}_l,\widehat{\mathbf{\Sigma}}_l)$ and approximate the integral in (19) as

$$\begin{aligned} p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l) &\overset{(a)}{\approx} \sum_{s=1}^{S}\frac{1}{S}p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l^*(s)) \\ &\overset{(b)}{=} \sum_{s=1}^{S}\frac{1}{S}\mathcal{N}([\widehat{\mathbf{x}}]_l;[\widehat{\boldsymbol{\mu}}_x^{\text{CGP}}(s)]_l,[\widehat{\mathbf{C}}_x^{\text{CGP}}(s)]_{ll}). \end{aligned} \quad (20)$$

In (20), (a) follows from the Monte-Carlo approximation procedure [45], and (b) from (16), with the $[\widehat{\boldsymbol{\mu}}_x^{\text{CGP}}(s)]_l$ and $[\widehat{\mathbf{C}}_x^{\text{CGP}}(s)]_{ll}$ being the same as $[\widehat{\boldsymbol{\mu}}_x^{\text{CGP}}]_l$ and $[\widehat{\mathbf{C}}_x^{\text{CGP}}]_{ll}$ but with the test RSS vector $\widehat{\mathbf{p}}_l$ replaced by the Monte-Carlo sample $\widehat{\mathbf{p}}_l^*(s)$. Since the r.h.s of (20) is a mixture of S Gaussians with identical weights, we know from [46] (eq. (14.10)-(14.11)) that the l.h.s can be approximated as a Gaussian distribution with the same first and second order moments, i.e.,

$$\begin{aligned} p([\widehat{\mathbf{x}}]_l|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{p}}_l) &\approx \mathcal{N}([\widehat{\mathbf{x}}]_l;[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l,[\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll}), \quad \text{where,} \\ [\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l &= \frac{1}{S}\sum_{s=1}^{S}[\widehat{\boldsymbol{\mu}}_x^{\text{CGP}}(s)]_l, \\ [\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll} &= \frac{1}{S}\sum_{s=1}^{S}([\widehat{\boldsymbol{\mu}}_x^{\text{CGP}}(s)]_l - [\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l)^2 \\ &\quad + \frac{1}{S}\sum_{s=1}^{S}[\widehat{\mathbf{C}}_x^{\text{CGP}}(s)]_{ll}, \quad \forall l = 1,\ldots,\widehat{L}. \end{aligned} \quad (21)$$

In (21), $[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l$ refers to the estimate of the test $x-$coordinate $[\widehat{\mathbf{x}}]_l$ from NaGP and $[\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll}$ refers to the associated predictive variance. As is the case with with CGP, we know that the 2σ error-bars on the estimate $[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l$, calculated from the predictive variance as $\pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll}}$, provide us with the 95% confidence interval $\left[[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll}}\right]$ on choosing $[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l$ as the estimate of $[\widehat{\mathbf{x}}]_l$. By increasing S, we can increase the accuracy of the $[\widehat{\boldsymbol{\mu}}_x^{(\text{NaGP})}]_l$ and $[\widehat{\mathbf{C}}_x^{(\text{NaGP})}]_{ll}$ values because the numerical approximation in (20) becomes tighter with increasing S [45].

Remark 3: We may note from (19) that, unlike the CGP method which naively treats the noisy test RSS vectors as noise-free, the NaGP method treats the noise-free components in the test RSS vectors as hidden variables and integrates them out using statistical knowledge of the noise present. By doing so, the NaGP method learns from and incorporates the noise covariance matrices $\{\widehat{\mathbf{\Sigma}}_l\}$ into the predicted mean and variance expressions (c.f. (21)). This learning enables NaGP to provide more realistic 2σ error-bars on the predicted locations than the CGP.

Next, we present details on the performance metrics considered and also derive a Cramer-Rao lower bound on the root-mean-squared error performance of the two GP methods under study.

VII. Performance Metrics and Cramer-Rao Lower Bound

We measure prediction performance in terms of (i) the root-mean-squared prediction error (RMSE) and (ii) the log-predictive density (LPD), defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{l=1}^{\widehat{L}}([\widehat{\mathbf{x}}]_l - [\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l)^2 + ([\widehat{\mathbf{y}}]_l - [\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l)^2}{\widehat{L}}}, \quad \text{and}$$

$$\begin{aligned} \text{LPD} &= \frac{1}{\widehat{L}}(\log(p(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}})) + \log(p(\widehat{\mathbf{y}}|\widetilde{\mathbf{y}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}}))), \\ &= -\log(2\pi) - \frac{1}{2\widehat{L}}\sum_{l=1}^{\widehat{L}}\log([\widehat{\mathbf{C}}_x^{(\cdot)}]_{ll}) + \log([\widehat{\mathbf{C}}_y^{(\cdot)}]_{ll}) \end{aligned} \quad (22)$$

$$+ \frac{([\widehat{\mathbf{x}}]_l - [\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l)^2}{[\widehat{\mathbf{C}}_x^{(\cdot)}]_{ll}} + \frac{([\widehat{\mathbf{y}}]_l - [\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l)^2}{[\widehat{\mathbf{C}}_y^{(\cdot)}]_{ll}}, \quad (23)$$

where $[\widehat{\mathbf{x}}]_l$ and $[\widehat{\mathbf{y}}]_l$ are the actual x and y coordinates of the test user l, $[\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l$ and $[\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l$ are the estimates of $[\widehat{\mathbf{x}}]_l$ and $[\widehat{\mathbf{y}}]_l$ as given by the chosen GP method, and $[\widehat{\mathbf{C}}_x^{(\cdot)}]_{ll}$ and $[\widehat{\mathbf{C}}_y^{(\cdot)}]_{ll}$ are the variances associated with the estimates $[\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l$ and $[\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l$, respectively. For example, if we choose CGP, $[\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l = [\widehat{\boldsymbol{\mu}}_x^{(\text{CGP})}]_l$, $[\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l = [\widehat{\boldsymbol{\mu}}_y^{(\text{CGP})}]_l$, $[\widehat{\mathbf{C}}_x^{(\cdot)}]_{ll} = [\widehat{\mathbf{C}}_x^{(\text{CGP})}]_{ll}$ and $[\widehat{\mathbf{C}}_y^{(\cdot)}]_{ll} = [\widehat{\mathbf{C}}_y^{(\text{CGP})}]_{ll}$. The RMSE only takes the estimates $[\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l$ and $[\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l$ into account. In contrast, the LPD takes the entire predictive distribution into account. Observe from (22) that LPD penalizes overconfident location estimates by assigning larger weights to the prediction errors $([\widehat{\mathbf{x}}]_l - [\widehat{\boldsymbol{\mu}}_x^{(\cdot)}]_l)$ and $([\widehat{\mathbf{y}}]_l - [\widehat{\boldsymbol{\mu}}_y^{(\cdot)}]_l)$ when the associated uncertainties $[\widehat{\mathbf{C}}_x^{(\cdot)}]_{ll}$ and $[\widehat{\mathbf{C}}_y^{(\cdot)}]_{ll}$ are small. Lower RMSE and higher LPD indicate better prediction performance.

A. Cramer-Rao Lower Bound on the RMSE Performance

To evaluate the RMSE performance of the presented GP methods, we need a theoretical lower bound on the achievable RMSE. Towards this, we derive a Bayesian Cramer-Rao lower bound that reflects the best possible RMSE of any unbiased

[4]For notational ease, all integrals are written as indefinite integrals, but in reality, they are definite over appropriate sets.

estimator of the test user's location coordinates. We view the location prediction problem under study as an estimation problem in which we wish to estimate the test $x-$coordinate vector $\widehat{\mathbf{x}}$ from the training $x-$coordinate measurements $\widetilde{\mathbf{x}}$, given the training RSS $\widetilde{\mathbf{P}}$, the test RSS $\widehat{\mathbf{P}}$, and the vector $\boldsymbol{\theta}$ (available upon training the GP). For any of the two GP methods under study, the Bayesian Cramer Rao lower bound (BCRLB) on the expected squared-error matrix of the test users' $x-$coordinates can then be obtained as [49]

$$\begin{aligned}&\mathbb{E}((\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)})(\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)})^T)\\&\qquad\succeq \text{BCRLB}_x,\quad \text{where,}\\&\text{BCRLB}_x=-(\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}[\nabla_{\widehat{\mathbf{x}}}\log(p(\widetilde{\mathbf{x}},\widehat{\mathbf{x}}|\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta}))]^T))^{-1}.\end{aligned}\tag{24}$$

In (24), $\widehat{\boldsymbol{\mu}}_x^{(.)}$ is the estimate of $\widehat{\mathbf{x}}$ from the chosen GP method and BCRLB_x is the associated BCRLB. The expectation $\mathbb{E}(.)$ is with respect to $\widetilde{\mathbf{x}}$ and $\widehat{\mathbf{x}}$. The term $\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}(\nabla_{\widehat{\mathbf{x}}}(\log p(\widetilde{\mathbf{x}},\widehat{\mathbf{x}}|\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta}))))$ on the r.h.s of (24) is the Bayesian Information Matrix (BIM) on $\widehat{\mathbf{x}}$ [49], which we simplify as

$$\begin{aligned}&\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}[\nabla_{\widehat{\mathbf{x}}}(\log(p(\widetilde{\mathbf{x}},\widehat{\mathbf{x}}|\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta})))]^T)\\&\overset{(a)}{=}\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}[\nabla_{\widehat{\mathbf{x}}}\log(p(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta}))+\nabla_{\widehat{\mathbf{x}}}\log(p(\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta}))]^T)\\&\overset{(b)}{=}\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}[\nabla_{\widehat{\mathbf{x}}}(-\frac{\widehat{L}}{2}\log(2\pi)-\frac{1}{2}(|\widehat{\mathbf{C}}_x^{(.)}|)-\frac{1}{2}(\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)})^T\\&\quad(\widehat{\mathbf{C}}_x^{(.)})^{-1}(\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)}))+\nabla_{\widehat{\mathbf{x}}}(-\frac{\widetilde{L}}{2}\log(2\pi)-\frac{1}{2}\log|\widetilde{\boldsymbol{\Phi}}|\\&\quad-\frac{1}{2}\widetilde{\mathbf{x}}^T\widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{x}})]^T)\\&\overset{(c)}{=}-\mathbb{E}(\nabla_{\widehat{\mathbf{x}}}[\nabla_{\widehat{\mathbf{x}}}(\frac{1}{2}(\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)})^T(\widehat{\mathbf{C}}_x^{(.)})^{-1}(\widehat{\mathbf{x}}-\widehat{\boldsymbol{\mu}}_x^{(.)}))]^T)\\&\overset{(d)}{=}-(\widehat{\mathbf{C}}_x^{(.)})^{-1},\end{aligned}\tag{25}$$

where (a) follows from Bayes' rule, (b) from substituting $p(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta})$ from the chosen GP method (c.f. **Remark 4** below) and $p(\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta})=p(\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}},\boldsymbol{\theta})$ from (11), (c) from setting the gradient $\nabla_{\widehat{\mathbf{x}}}(.)$ of all the terms which are constant with respect to (w.r.t) $\widehat{\mathbf{x}}$ to zero, (d) from evaluating the gradient twice w.r.t $\widehat{\mathbf{x}}$, and then observing that the elements of $(\widehat{\mathbf{C}}_x^{(.)})^{-1}$ are independent of both $\widetilde{\mathbf{x}}$ and $\widehat{\mathbf{x}}$ (c.f. $\widehat{\mathbf{C}}_x^{(\text{CGP})}$ in (16), $\widehat{\mathbf{C}}_x^{(\text{NaGP})}$ in (21). Substituting (25) into (24), we have $\text{BCRLB}_x=\widehat{\mathbf{C}}_x^{(.)}$ In a similar fashion, we can obtain the BCRLB for the expected squared-error of the test users' $y-$coordinates as $\text{BCRLB}_y=\widehat{\mathbf{C}}_y^{(.)}$. Followed by this, we can obtain a BCRLB on the RMSE for predicting the test user locations as follows

$$\begin{aligned}\text{BCRLB}^{(\text{RMSE})}&=\sqrt{\frac{1}{\widehat{L}}\text{Tr}(\text{BCRLB}_x+\text{BCRLB}_y)}\\&\equiv\sqrt{\frac{1}{\widehat{L}}\text{Tr}(\widehat{\mathbf{C}}_x^{(.)}+\widehat{\mathbf{C}}_y^{(.)})},\end{aligned}\tag{26}$$

where $\widehat{L}$ is the number of test users. Eq. (26) shows that the $\text{BCRLB}^{(\text{RMSE})}$ for any chosen GP method can be obtained from its predictive covariances $\widehat{\mathbf{C}}_x^{(.)}$ and $\widehat{\mathbf{C}}_y^{(.)}$ through simple linear algebraic operations. To obtain a valid $\text{BCRLB}^{(\text{RMSE})}$, we must therefore ensure that the $\widehat{\mathbf{C}}_x^{(.)}$ and $\widehat{\mathbf{C}}_y^{(.)}$ values are accurate. **Remark 4** below summarizes our approach to obtain accurate $\widehat{\mathbf{C}}_x^{(.)}$ and $\widehat{\mathbf{C}}_y^{(.)}$ values for calculating the $\text{BCRLB}^{(\text{RMSE})}$.

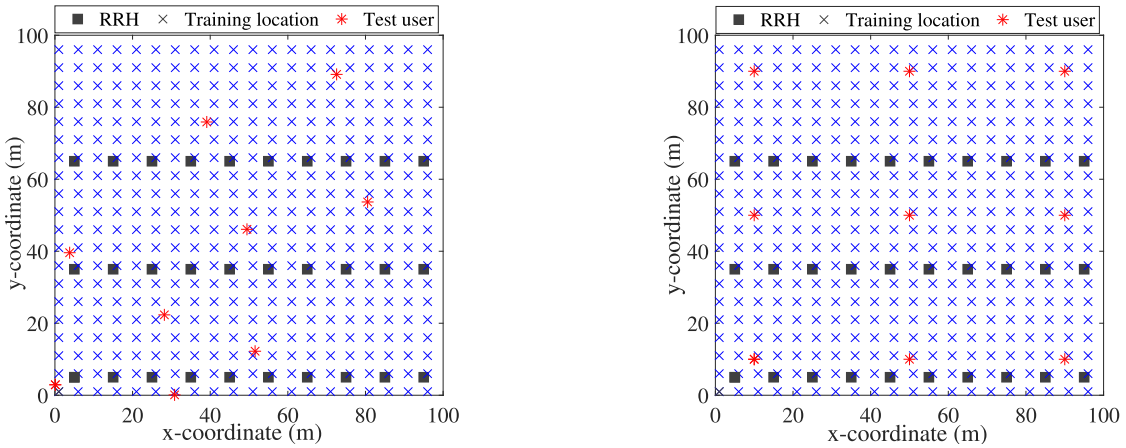
Remark 4: As noted in Section VI, the true predictive distribution $p(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta})$ cannot be obtained in exact form. Nevertheless, NaGP gives us a numerical approximation for $p(\widehat{\mathbf{x}}|\widetilde{\mathbf{x}},\widetilde{\mathbf{P}},\widehat{\mathbf{P}},\boldsymbol{\theta})$ (c.f. (20)). Therefore, to calculate the $\text{BCRLB}^{(\text{RMSE})}$ *for both the CGP and NaGP methods using (26), we advocate the use of $\widehat{\mathbf{C}}_x^{(\text{NaGP})}$ (and $\widehat{\mathbf{C}}_y^{(\text{NaGP})}$) obtained from (21) as the $\widehat{\mathbf{C}}_x^{(.)}$ (and $\widehat{\mathbf{C}}_y^{(.)}$).*

VIII. Numerical Studies and Discussions

We now present numerical examples to evaluate the RMSE and LPD performance of the two GP methods under study. We consider two example massive MIMO setups, as shown in Fig. 3, with $M=30$ RRH antennas and $\widetilde{L}=400$ training locations distributed in a grid-spaced manner over a service area of 100m $\times$ 100m. Our aim is to estimate the locations of $\widehat{L}=9$ test users distributed across the service area in a (i) uniformly random manner, as shown in Scenario A (c.f. Fig. 3a) and in a (ii) grid-spaced manner, as shown in Scenario B (c.f. Fig. 3b) respectively. The goals of our numerical studies are to (i) illustrate that we can obtain location estimates by training GP models with noise-free RSS, (ii) expose trends in the RMSE and LPD performance under varied shadowing noise levels and varied number of RRHs, (iii) verify whether the achieved RMSEs are close to the theoretical BCRLBs, (iv) examine the impact of correlated shadowing on the localization performance, and (v) demonstrate that extrapolation techniques may be employed to improve the RMSE performance of the CGP and NaGP methods. While we consider the general 3GPP Urban Micro propagation model [22] for most of our analysis, we also consider specific line-of-sight and non-line-of-sight conditions, as in the WINNER II A1 scenario [23], to evaluate the localization performance. In the future, field experimentations need to be conducted, for example via USRP implementations [54], to analyze the impact of multiple practical issues, such as the environmental changes due to human activities and the multiuser interference due to non-orthogonal pilots or pilot reuse. Suitable transfer learning [24]–[27] and pre/post-processing steps will need to be added to achieve similar localization performance across different times, areas, and device types.

A. Parameters and Setup

We assume that all the training user locations are available with a measurement error variance (σ_{er}^2) of -1dB. For the training phase, we generate noise-free training RSS $\widetilde{\mathbf{P}}$ using (4) with shadowing variance $\sigma_{z_{mk}}^2=0$ and other parameters as per Table I. Entries of Table I are chosen as follows. The path-loss parameters l_0, d_0, and η are chosen as per the 3GPP Urban Micro propagation model [22]. The user transmit power is chosen as per LTE standards to be 21dBm [51]. Total noise power in the system is set to -107.5dBm. The uplink receiver sensitivity, which represents the minimum detection threshold for the receiver to distinguish between the signal strength and the noise power, is set to -106.5dBm.



(a) Scenario A: Uniformly random test user locations

(b) Scenario B: Grid-spaced test user locations

Fig. 3. Simulation setup with $M = 30$ RRHs, $\widetilde{L} = 400$ training locations, and $\widehat{L} = 9$ test users.

TABLE I
SIMULATION PARAMETERS

System Parameters	Value
Path-loss parameters (3GPP UMi [22])	$d_0 = 10$m, $l_0 = -47.5$dB, $\eta = \begin{cases} 0 \text{ if } d_{mk} < 10\text{m}, \\ 2 \text{ if } 10\text{m} \le d_{mk} \le 45\text{m}, \\ 6.7 \text{ if otherwise.} \end{cases}$
UE transmit power	21dBm (125mW)
Noise power	-107.5 dBm
Receiver sensitivity	-106.5 dBm

Once the training RSS data is ready, a GP model is trained by solving the log-likelihood maximization problem in (12) using conjugate gradient (CG) method [48]. We run 20 trials with random initializations and choose the local optimum which gives the best training objective, i.e., $\log(p(\widetilde{\mathbf{x}}|\widetilde{\mathbf{P}}, \boldsymbol{\theta}))$ value. The same learned parameter vector $\bar{\boldsymbol{\theta}}$ is reused for evaluating the prediction performance of both CGP and NaGP methods because they use the same training procedure. We generate 200 Monte-Carlo test RSS matrices each for shadowing variance $\sigma_z^2 = 1, 2, \ldots, 5$dB, using (4) with parameters chosen as per Table I. During simulations, any instantaneous test RSS value that is lower than the receiver sensitivity is replaced with the noise power in the system. The RMSE and LPD values, averaged over the Monte-Carlo realizations, are reported. For the NaGP method, we set $S = 200$.

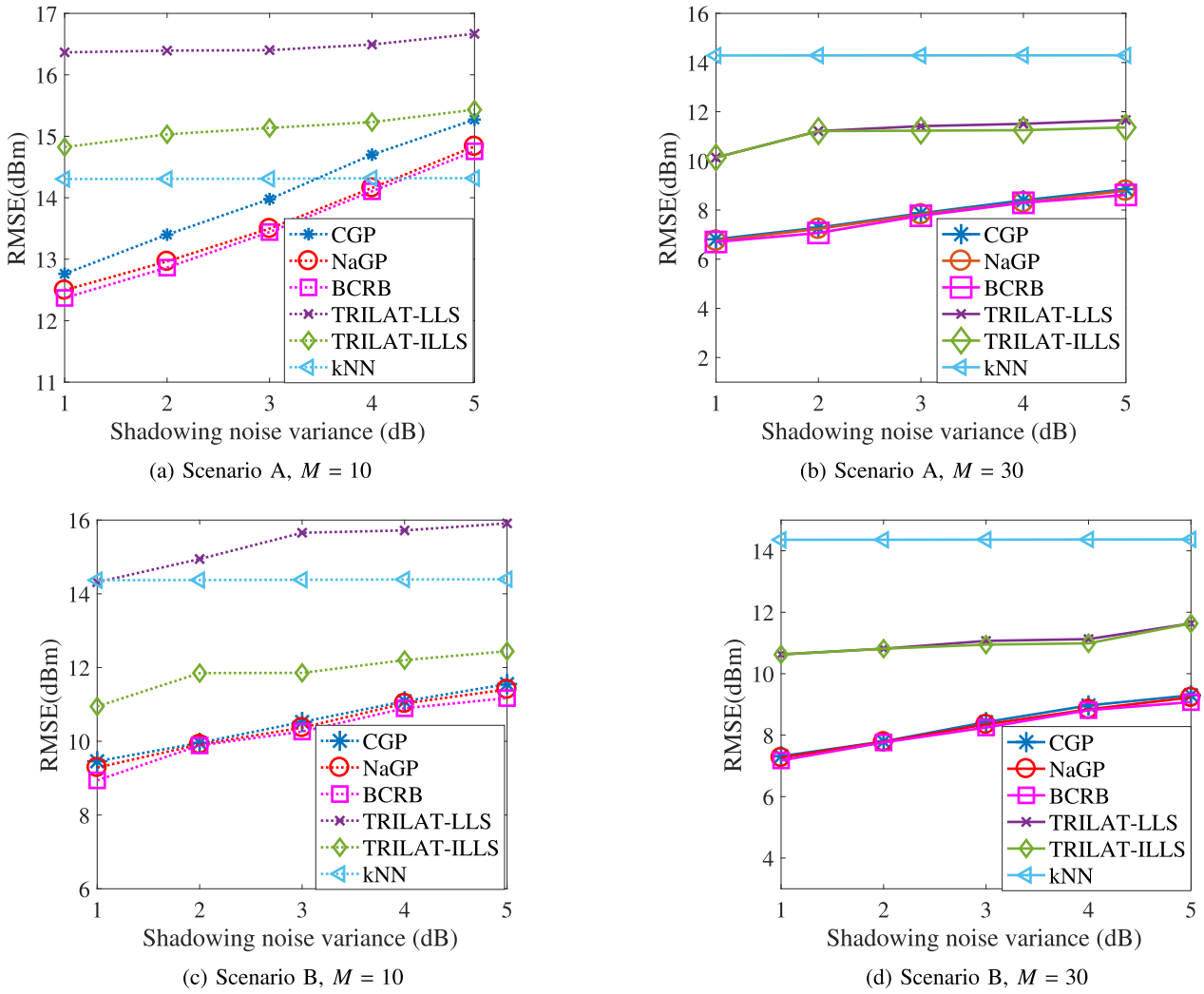
B. Baseline Schemes

As the first two baselines, we consider the linear least squares and improved linear least squares trilateration schemes proposed in [21] (eq. (19) and (30)). We refer to these methods as TRILAT-LLS and TRILAT-ILLS respectively. Both the TRILAT methods employ distance estimation, followed by trilateration, to estimate the user locations. Since these methods assume a single-slope path-loss model and three anchor nodes, for the implementation, we set $\eta = 2$ and choose the three RRHs with the largest RSS values as the anchor nodes for localization.

As the third baseline, we consider the k-nearest neighbours (KNN) [28], which is a popular unsupervised machine learning method for indoor positioning. In KNN, location estimates are obtained as the mean of the three training locations whose RSS vectors are closest in the RSS space to the test RSS vector $\mathbf{p}_k$. While the TRILAT methods serve as the non-ML baselines, the KNN method serves as the non-GP baseline.

C. RMSE Performance

In Fig. 4, we illustrate the average RMSE performance achieved by the CGP, NaGP, and the three baseline methods, when the shadowing variance (σ_z^2) is increased from 1dB to 5dB for $M = 10$ and 30. This plot also serves to demonstrate the impact of imperfect test RSS because the imperfectness due to residual multiuser interference can be modelled as additional noise in the test RSS vectors. Several interesting observations can be made from Fig. 4. Firstly, both the CGP and NaGP methods provide similar RMSE values for different shadowing noise levels. This is because the location estimates from the CGP and the NaGP methods are found to be similar in value. Secondly, the RMSE values increase with the shadowing noise level. This is expected because both the CGP and NaGP methods are trained with noise-free RSS and so, they tend to project the noise present in the input RSS onto the output location coordinate space. Thirdly, we notice that both the CGP and NaGP methods outperform both the TRILAT and KNN methods across different noise levels. The high RMSE values of the KNN method are a consequence of its unsupervised nature and also the presence of shadowing noise in the test RSS. The high RMSE values of TRILAT can be attributed to the single-slope path-loss model assumption. While on the other hand, we train the GP models with multi-slope path loss data and therefore, expect the GP methods to capture the non-linear relation between RSS and user locations in a more effective manner than the TRILAT method, leading to better RMSE performance. The performance gap reduces with increase in shadowing noise



(a) Scenario A, M = 10

(b) Scenario A, M = 30

(c) Scenario B, M = 10

(d) Scenario B, M = 30

Fig. 4. Average RMSE performance of the CGP, NaGP, TRILAT-LLS, TRILAT-ILLS, and KNN methods under Scenario A and B respectively for different shadowing noise levels, when $M = 10, 30$. The BCRLBs for the CGP and NaGP methods are also included in the plot.

levels because both the GP methods tend to project the input noise onto the output space. Lastly, we observe that the RMSE values are smaller when the number of RRHs M is increased from 10 to 30. This may be attributed to two reasons: (i) the increased average signal strength at the RRHs, which is in turn due to the reduced RRH-to-user distance, and (ii) the extra RSS information available to the GP models due to the increase in the number of RRHs.

D. Bayesian Cramer-Rao Lower Bounds

In Fig. 4, we have also plotted the BCRLBs on the RMSE performance of the CGP and NaGP methods. The BCRLBs are computed using (26) with $\widehat{\mathbf{C}}_x^{(.)} = \widehat{\mathbf{C}}_x^{(\text{NaGP})}$ and $\widehat{\mathbf{C}}_x^{(.)} = \widehat{\mathbf{C}}_y^{(\text{NaGP})}$. We notice that for both the GP methods, the achieved RMSE performance is very close to the BCRLBs. Note that the RMSE and BCRLB trends are similar for the Scenarios A and B. We notice similar trends for the 2σ error-bars and the log-predictive densities as well. Therefore, from here on, we present results for Scenario A only.

E. 2σ Error-Bar Performance

In Fig. 5, we plot the average 2σ error-bars on the test users' $x-$ coordinates and $y-$coordinates, respectively, as given by the CGP and the NaGP methods for Scenario A (c.f. Fig. 3a). The TRILAT and KNN methods do not provide 2σ error-bar estimates and are therefore not included in the plot. We also plot the number of true test user locations which are within the 2σ confidence range $([\widehat{\boldsymbol{\mu}}_x^{(.)}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(.)}]_{ll}}, [\widehat{\boldsymbol{\mu}}_y^{(.)}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_y^{(.)}]_{ll}})$ of the estimated locations. We observe that the CGP method tends to provide unrealistically small 2σ error-bars, which are very low even if the corresponding RMSE values in Fig. 4a and 4b are high. As a result, we note that less than 15% (for $M = 10$) and 5% (for $M = 30$) of the true test user locations fall inside the 2σ confidence range of the CGP method. On the other hand, we observe that the NaGP method consistently provides realistic 2σ error-bars on the estimated locations for varied shadowing noise levels and for both $M = 10, 30$. Consequently, for both $M = 10, 30$, we observe that more than 90% of the true user locations (c.f. Fig. 5) are within the 2σ confidence range provided by NaGP.

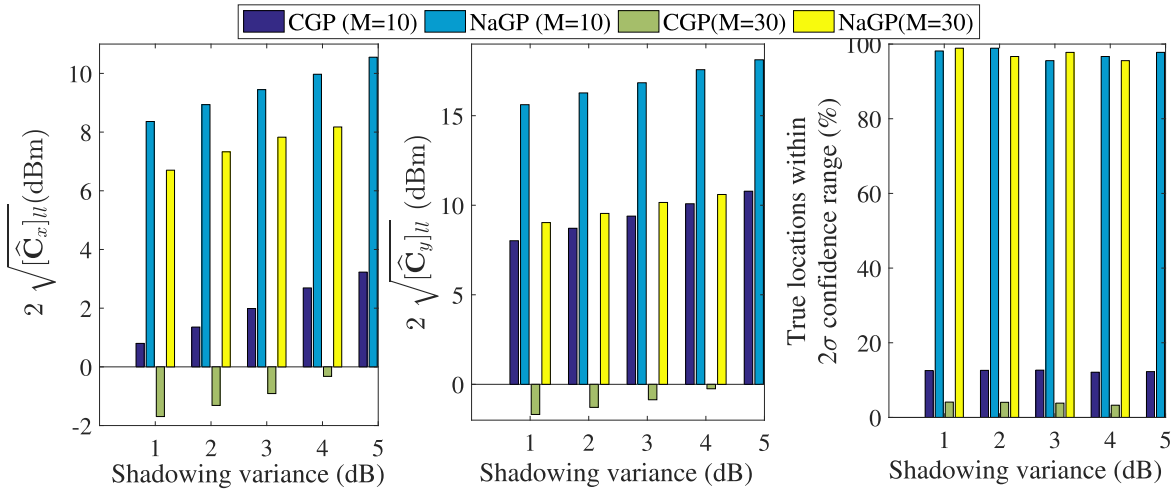


Fig. 5. Plots of the average 2σ error-bars on the test users' x and y coordinates and the number of true test user locations within the 2σ confidence range $([\widehat{\boldsymbol{\mu}}_x^{(.)}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_x^{(.)}]_{ll}}, [\widehat{\boldsymbol{\mu}}_y^{(.)}]_l \pm 2\sqrt{[\widehat{\mathbf{C}}_y^{(.)}]_{ll}})$ of the estimated locations, as provided by the CGP and NaGP methods for different shadowing noise levels, when $M = 10, 30$.

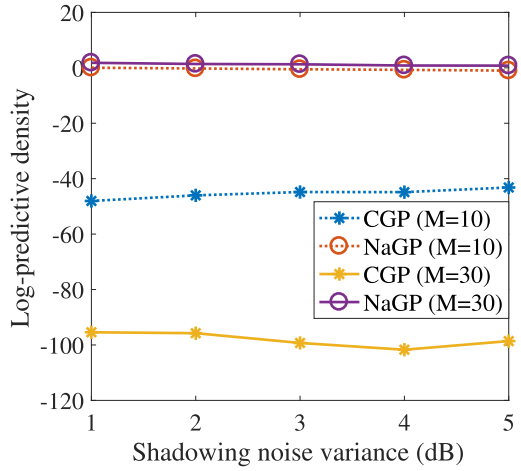


Fig. 6. Average LPD performance of the CGP and NaGP methods under Scenario A for different shadowing noise levels, when $M = 10, 30$.

F. LPD Performance

In Fig. 6, we plot the LPD performance of the CGP and NaGP methods when $M = 10$ and 30 for Scenario A (c.f. Fig. 3a). The TRILAT and KNN methods do not provide the log-predictive densities and are therefore not included in the plot. We observe that the CGP method achieves very low LPD values because it provides unrealistically small $[\widehat{\mathbf{C}}_x]_{ll}$ and $[\widehat{\mathbf{C}}_y]_{ll}$ values (c.f. Fig. 5), with less than 15% for $M = 10$ (and 5% for $M = 30$) of the true user locations falling inside the 2σ confidence range of the estimated locations. Note from (22) that the LPD metric penalizes such overconfident estimates by assigning large weights to the prediction error. NaGP achieves much higher LPD values than CGP because it provides realistic $[\widehat{\mathbf{C}}_x]_{ll}$ and $[\widehat{\mathbf{C}}_y]_{ll}$ values (c.f. Fig. 5), with more than 90% of the true user locations inside the 2σ confidence range of the location estimates).

Taking both the RMSE and LPD plots into perspective, we observe that NaGP achieves significantly better LPD performance than CGP, while achieving similar RMSE performance. We therefore conclude that NaGP provides better prediction performance among the two GP methods. Note that the superior LPD performance of NaGP is because it learns from the statistical properties of the noise in the test RSS to provide realistic 2σ error-bars on the estimated locations.

G. Impact of the number of RRHs

In Fig. 7, we plot the average RMSE performance of the CGP and NaGP methods, along with their BCRLBs, when the number of RRHs M is increased from 10 to 100 and when the shadowing noise variance $\sigma_z^2 = 3, 5$ dB. For comparison, we plot the RMSE performance of the TRILAT-ILLS and KNN baselines. Also, in the same figure, we plot the corresponding 2σ error-bars on the test users' x and y coordinates, as provided by the CGP and NaGP methods. We notice that for varied levels of M, the two GP methods consistently perform better than the TRILAT and KNN methods. As M increases, we also notice that the RMSEs and BCRLBs of both CGP and NaGP methods decrease initially, followed by saturation. Similar pattern is observed for the 2σ error-bars provided by NaGP. The decrease in RMSE with M may be attributed to two major reasons: (i) the increased average signal strength resulting from the reduced RRH-to-user distance and (ii) the extra RSS information that becomes available to the GP models because of the addition of the new RRHs.

On the RMSE Floor Phenomenon: Note from Fig. 7 that the RMSE performance of both CGP and NaGP methods saturates beyond a certain M. To identify the cause of this RMSE floor, we may apply singular value decomposition on the training RSS $\widetilde{\mathbf{P}}$ and observe the obtained singular values. Upon doing so, we realize that the first few singular values represent most of the energy contained in $\widetilde{\mathbf{P}}$, as is also reported in our recent work [47]. As a consequence, when we increase M beyond a certain point (30 RRHs for the Scenario A in our simulation setup), the amount of additional information provided by the

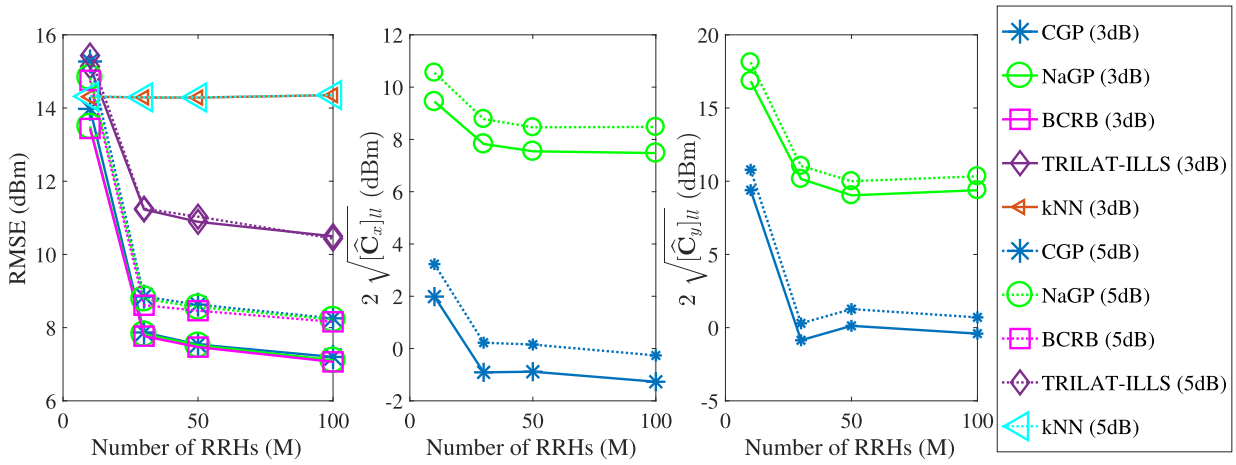


Fig. 7. Average RMSE, the associated BCRLBs, and the 2σ error-bars on the estimated x and y cordinates, as given by the NaGP and CGP methods for different M, when $\sigma_z^2 = 3, 5$dB. For comparison, we also plot the RMSE performance of the TRILAT-ILLS and KNN baselines.

new RRHs is very small. This leads to the RMSE floor phenomeon in the RMSE vs M plots. Also note that the BCRLB curves in Fig. 7 serve as a guideline to choose the number of RRHs for localization - for example, if we are operating with an excessively large number of RRHs in the given area (as typically done in massive MIMO systems for enhanced coverage and throughput services), we can simply choose a small subset from the total number of RRHs for user positioning (for example, a subset comprising 30 RRHs) and still be able to achieve similar RMSE as achieved with all the RRHs in use.

H. Impact of Channel Correlation

So far, we have assumed that the shadowing noise coefficients $\{z_{mk}\}$ of different uplink channels in the system are mutually uncorrelated. This assumption may not always hold in practice because the RRH antennas (or users) that are close to each other may share a common set of obstacles. Therefore, it is important to study how the presence of correlated shadowing among the test RSS vectors impacts the localization performance of the GP methods under study.

To generate correlated shadowing noise coefficients, we consider the two-component model which has been validated through field studies in [52]. The shadowing noise coefficient z_{mk} for the uplink channel between user k and RRH m is modelled as

$$z_{mk} = \sigma_{z_{mk}}(\sqrt{\varrho}\epsilon_m + \sqrt{1-\varrho}\varepsilon_k), \quad \text{where,}$$
$$\epsilon_m \sim \mathcal{N}(0,1),\ \varepsilon_k \sim \mathcal{N}(0,1), \quad \text{and } 0 \leq \varrho \leq 1. \quad (27)$$

In (27), $\sigma_{z_{mk}}$ is the standard deviation of the shadowing noise, ϵ_m models the shadowing due to obstacles around the RRH m, ε_k models the shadowing due to obstacles around the user k, and ϱ is a weight parameter. The random variables ϵ_m and ε_k are independent of each other. For any two RRHs m and m' and any two users k and k', the correlation between ϵ_m and $\epsilon_{m'}$ and the correlation between ε_k and $\varepsilon_{k'}$ are modelled as, $\forall m, m' = 1, \ldots M$, and $k, k' = 1, \ldots, K$,

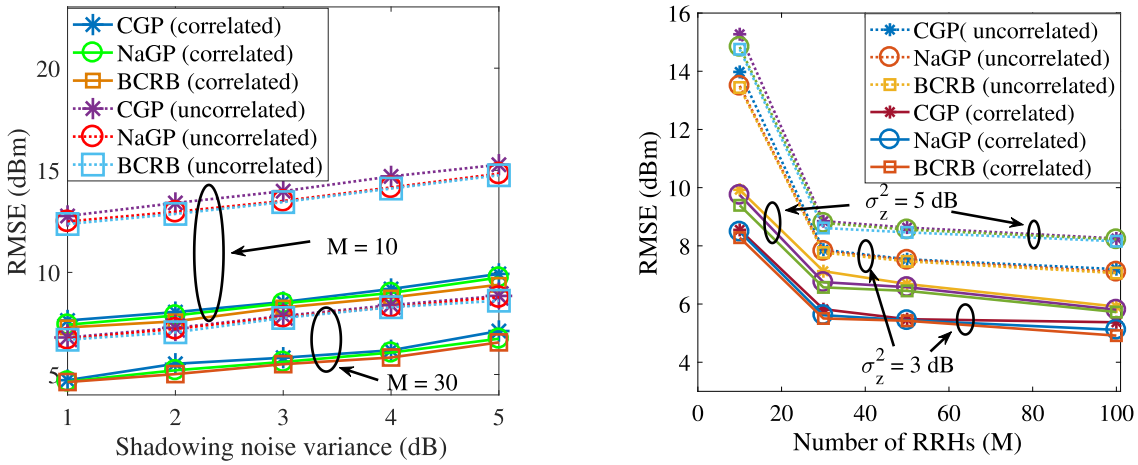
$$\mathbb{E}(\epsilon_m \epsilon_{m'}) = 2^{-\frac{d^{(bs)}_{mm'}}{d^{(\text{dec})}}} \quad \text{and } \mathbb{E}(\varepsilon_k \varepsilon_{k'}) = 2^{-\frac{d^{(ue)}_{kk'}}{d^{(\text{dec})}}}, \quad (28)$$

where we have denoted $d^{(bs)}_{mm'}$ as the distance between RRHs m and m' and $d^{(ue)}_{kk'}$ as the distance between users k and k'. The term $d^{(\text{dec})}$ denotes the decorrelation distance, which depends on the type of environment. For our simulations, we have set $\varrho = 0.5$ and $d^{(\text{dec})} = 20$m.

To study the impact of correlated shadowing, we plot in Fig. 8 the RMSE performance of the CGP and NaGP methods, along with the associated BCRLBs, for different shadowing noise levels (c.f. Fig. 8a) and different number of RRHs (c.f. Fig. 8b). From Fig. 8a and Fig. 8b, we notice that the RMSE performance improves in the presence of correlated shadowing. We provide the following intuition for the improvement in RMSE. GP methods perform non-linear regression by mapping the correlation among inputs to the correlation among outputs, through the function $\phi(.,.)$. When the correlation among the inputs is more prominent, the GP methods would need to make lesser number of independent predictions on the test dataset. That is, when an accurate prediction is made on a given test RSS vector, the GP model tends to make accurate predictions on other test vectors that have high correlation with the considered test RSS vector. This leads to the improvement in the RMSE performance. Lastly, similar to the case with uncorrelated shadowing (c.f. Fig. 4 and 7), we notice from Fig. 8a and 8b that (i) for a given M, the RMSE values increase with σ_z^2, and (ii) for a given noise variance, the RMSE values decrease with M initially, followed by saturation.

I. Localization on Indoor LoS and NLoS Conditions

We now simulate the A1 indoor line of sight (LoS) and non-line-of-sight (NLoS) scenarios proposed in the WINNER-II project [23] and evaluate the localization performance of the two GP methods under study. For the LoS and NLoS scenarios, the shadowing noise standard deviation σ_z is 3 dB and 4dB respectively, and the correlation distance $d^{(\text{dec})}$ is 6m and 4m respectively [23]. For the training and test user locations, we have considered the setup in Fig. 3a. In Fig. 9, we plot the average RMSE, LPD, and the number of true user locations within the estimated 2σ confidence range of the



(a) Average RMSE vs σ_z^2 for $M = 10, 30$

(b) Average RMSE vs M for $\sigma^2 = 3, 5$dB

Fig. 8. Average RMSE performance, and the corresponding BCRLBs, of the CGP and NaGP methods under the presence of correlated shadowing among the test RSS vectors.

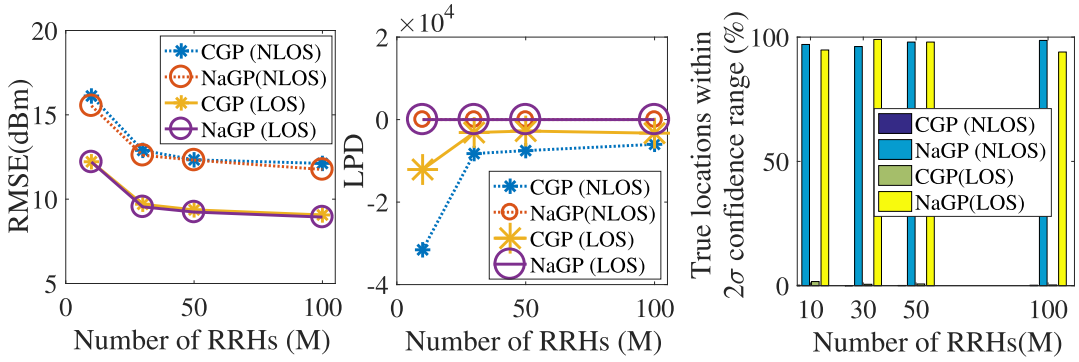


Fig. 9. Localization performance of the CGP and NaGP methods under the WINNER II A1-LOS and A1-NLOS scenarios. We plot the average RMSE, LPD, and the number of true test user locations within the 2σ confidence range, as given by the GP methods for different M.

CGP and NaGP methods, when M is varied from 10 to 100. We observe that the trends in the localization performance are similar to the case with the general 3GPP UMi model (c.f. Fig. 5, 6, and 7).

J. Extrapolation to Improve the RMSE Performance

In all our numerical studies so far, we notice that the RMSE levels achieved by both the CGP and NaGP methods are too high for use in practical location-aware wireless applications. To address this limitation, we now provide an illustration example on how the achieved RMSE values can be lowered through addition of artificial noise, followed by extrapolation of the resulting location estimates to the zero input noise scenario.

We add artificial noise $\widehat{\boldsymbol{\epsilon}}_l$ to the test RSS $\widehat{\mathbf{p}}_l$, such that $\widehat{\boldsymbol{\epsilon}}_l$ is zero-mean and i.i.d. as $\widehat{\mathbf{z}}_l$, and define the extrapolation function $f_x^{(\text{ext})}$ as

$$f_x^{(\text{ext})}(\lambda, \widehat{\mathbf{p}}_l) = \int f_x(\widehat{\mathbf{p}}_l + \sqrt{\lambda}\widehat{\boldsymbol{\epsilon}}_l) dp(\widehat{\boldsymbol{\epsilon}}_l), \overset{(a)}{=} \int f_x(\widehat{\mathbf{p}}_l^* + (\sqrt{1+\lambda})\widehat{\boldsymbol{\epsilon}}_l) dp(\widehat{\boldsymbol{\epsilon}}_l). \quad (29)$$

In (29), λ denotes the variance of the added noise and (a) follows from the fact that $\widehat{\mathbf{p}}_l = \widehat{\mathbf{p}}_l^* + \widehat{\mathbf{z}}_l$ (c.f. (17)) and that the noise vectors $\widehat{\boldsymbol{\epsilon}}_l$ and $\widehat{\mathbf{z}}_l$ are i.i.d. Since we have modelled $f_x(.)$ as a GP, we then know from (29) that the realizations of the extrapolation function $f_x^{(\text{ext})}$ are the same as the perturbed x-coordinate estimates from the NaGP method (c.f. (17)-(21)), obtained by replacing the original test RSS $\widehat{\mathbf{p}}_l$ with the noisier test RSS $\widehat{\mathbf{p}}_l + \sqrt{\lambda}\widehat{\boldsymbol{\epsilon}}_l$.

We can now model the function $f_x^{(\text{ext})}$ as a new GP (ExGP) with the inputs as λ, for example, $\lambda = [0, 0.3, 0.6, \ldots, 3]$, and the outputs as the realizations of $f_x^{(\text{ext})}(\lambda, \widehat{\mathbf{p}}_l)$, i.e., the perturbed $x-$coordinate estimates from the NaGP method. The predictive variances associated with the perturbed $x-$coordinate estimates serve as output noise. If we then feed a test input of $\lambda = -1$ to the newly trained GP model, we know from (29) that we can obtain location estimates corresponding to the zero input noise scenario. Such extrapolation to the zero-noise inputs would significantly lower the RMSE values obtained from NaGP because we have trained the NaGP with noise-free RSS. For illustration, we plot in Fig. 10 the RMSE performance of the ExGP for the WINNER-II A1 LoS and NLoS scenarios. We have considered $M = 30, 50$ and the

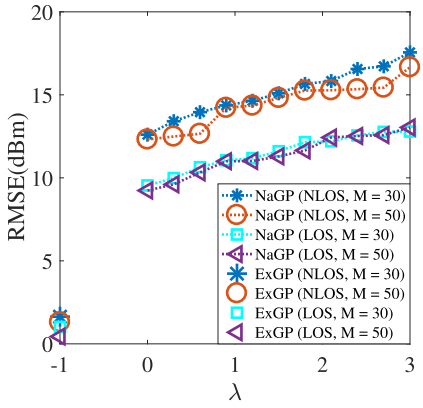


Fig. 10. RMSE performance of the extrapolation GP (ExGP) method for the WINNER-II A1-LOS and A1-NLOS scenarios, when the training inputs are $\lambda = [0, 0.3, 0.6, \ldots, 3]$, the training outputs are the corresponding location estimates from NaGP, and the test input is $\lambda = -1$. For reference, we also plotted the RMSE achieved by NaGP for each training λ.

same training and test location setup as in Fig. 3a. For both the LoS and NLoS scenarios, we notice that $m-$level accuracy is achievable.

IX. Conclusion

We have proposed a machine learning approach to estimate user locations from their uplink received signal strength (RSS) data in a distributed massive multiple-input multiple-output (MIMO) system. The proposed approach considers noise-free RSS for training and noisy RSS for the test purposes. For the machine learning task, we have applied two Gaussian process regression (GP) methods, namely, the conventional GP (CGP) method and the moment-matching based numerical approximation GP (NaGP) method. Firstly, we have identified that the CGP method provides unrealistically small 2σ error-bars on the estimated locations because it naively treats the noisy test RSS as noise-free. We have overcome this limitation using the NaGP method because it learns from the statistical properties of the noise present in the test RSS. We have also derived a Bayesian Cramer-Rao lower bound (BCRLB) on the achievable root-mean-squared error (RMSE) performance of the two GP methods under study.

Through numerical studies, we have provided several new insights on RSS-based user positioning in distributed massive MIMO: (i) it is possible to derive realistic 2σ error-bars on the location estimates if the noisy nature of the test RSS is taken into account, as done in the NaGP method (ii) the RMSE values of the CGP and NaGP methods increase with the shadowing noise level, and therefore, noise reduction may be attempted to improve the localization performance, (iii) when we increase the number of RRHs, the improvement in RMSE becomes marginal beyond a certain point, (iv) the presence of correlation among shadowing improves the RMSE of the two GP methods under study, and (v) extrapolation to the zero input noise scenario can significantly improve the RMSE performance.

Several exciting research directions arise from the presented work. Firstly, we have modelled the x and y coordinates of the users as independent random variables. Better prediction performance may be achieved if the training procedure takes the correlation between the x and y coordinates into account. Also, theoretical investigation may be done to prove that the RMSE of the presented GP methods attain the BCRLBs. Secondly, future work may analyze the case of non-orthogonal pilots and the resulting impact of pilot interference on the localization performance. Thirdly, compressive sensing techniques [53] may be employed to perform dimensionality reduction on the RSS vectors. This can help in reducing the computational complexity and also in reducing the noise level in the test RSS. Lastly, field studies, for example using USRP implementations [54], need to be conducted to verify the localization performance of the studied GP methods under practical settings involving environmental changes due to human activities.

Appendix

A. Complexity Analysis of the CGP and NaGP Methods

1) Training: The CGP and NaGP methods share the same training procedure. When the training objective is to maximize the log-likelihood as in (12) and a gradient-based method, such as conjugate gradient, is used to obtain the optimum points, we would incur $\mathcal{O}(\widetilde{L}^3)$ computations [9], where $\widetilde{L}$ is the number of training points.

2) CGP Method: From (15), we note that $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}$ involves one calculation of $\widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{x}}$, which incurs $\mathcal{O}(\widetilde{L}^3)$ complexity (due to Cholesky decomposition of $\widetilde{\boldsymbol{\Phi}}$), and $\widehat{L}\widetilde{L}$ calculations of $\phi(\widehat{\mathbf{p}}_l, \widetilde{\mathbf{p}}_i)$, which incurs $\mathcal{O}(\widehat{L}\widetilde{L}M)$ complexity. Typically, we have $M \ll \widetilde{L}$ and also $\widehat{L} \ll \widetilde{L}$. Therefore, $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}$ requires $\mathcal{O}(\widetilde{L}^3)$ computations. Next, from (16), we note that the computation of $[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}$ for the $\widehat{L}$ test users involves $\widehat{L}\widetilde{L}$ computations of $\phi(\widehat{\mathbf{p}}_l, \widetilde{\mathbf{p}}_i)$, which incurs $\mathcal{O}(\widehat{L}\widetilde{L}M)$ operations, $\widehat{L}\widetilde{L}$ computations of the product $[(\widetilde{\boldsymbol{\Phi}})^{-1}]_{ij}\phi(\widetilde{\mathbf{p}}_j, \widehat{\mathbf{p}}_l)$, which incurs $\mathcal{O}(\widetilde{L}\widehat{L}M + \widetilde{L}^2\widehat{L})$ complexity, and $\widehat{L}$ computations of the sum $\sum_{i=1}^{\widetilde{L}}\sum_{j=1}^{\widetilde{L}}\phi(\widehat{\mathbf{p}}_l, \widetilde{\mathbf{p}}_i)[(\widetilde{\boldsymbol{\Phi}})^{-1}]_{ij}\phi(\widetilde{\mathbf{p}}_j, \widehat{\mathbf{p}}_l)$, which incurs another $\mathcal{O}(\widetilde{L}\widehat{L})$ complexity. Typically, we know $M \ll \widetilde{L}$ and $\widehat{L} \ll \widetilde{L}$. Therefore, we require a total of $\mathcal{O}(\widetilde{L}^2\widehat{L})$ computations to calculate the $[\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}]_{ll}$ for the $\widehat{L}$ test users.

3) NaGP Method: To calculate $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{NaGP})}$, observe from (21) that we need a total of S evaluations of $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}(s)$. This amounts to $\mathcal{O}(\widetilde{L}^3)$ computations for $\boldsymbol{\psi} = \widetilde{\boldsymbol{\Phi}}^{-1}\widetilde{\mathbf{x}}$ and $\mathcal{O}(S\widetilde{L}\widehat{L}M)$ computations for $\phi(\widehat{\mathbf{p}}_l^*(s), \widetilde{\mathbf{p}}_i)$. Typically, $S\widetilde{L}\widehat{L}M$ is of the same order as $\widetilde{L}^2$, and therefore, $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{NaGP})}$ incurs $\mathcal{O}(\widetilde{L}^3)$ complexity. After calculating $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{NaGP})}$, the calculation of $\widehat{\mathbf{C}}_x^{(\mathrm{NaGP})}$ requires S evaluations of $\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}(s)$, leading to $\mathcal{O}(S\widehat{L}^2\widetilde{L})$ complexity.

From the above discussion, we note that the calculation of $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{CGP})}$ and $\widehat{\boldsymbol{\mu}}_x^{(\mathrm{NaGP})}$ incurs similar complexity, i.e., $\mathcal{O}(\widetilde{L}^3)$. Also, the complexity of $\widehat{\mathbf{C}}_x^{(\mathrm{NaGP})}$ is S times higher than that of $\widehat{\mathbf{C}}_x^{(\mathrm{CGP})}(s)$. This is the cost we pay to obtain realistic 2σ error-bars on the location estimates. Lastly, we note that the complexities of CGP and NaGP are not affected when we increase M to a few 100s because the number of training points $\widetilde{L}$ is the dominating factor.

REFERENCES

[1] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "A numerical approximation method for RSS-based user positioning in distributed massive MIMO," in *Proc. IEEE 11th Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2017, pp. 1–6.

[2] M. S. Grewal, L. R. Weill, and A. P. Andrews, *Global Positioning Systems, Inertial Navigation, and Integration*. New York, NY, USA: Wiley, 2001.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.

[5] K. T. Truong and R. W. Heath, Jr., "The viability of distributed antennas for massive MIMO systems," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 1318–1323.

[6] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[7] J. Wang and L. Dai, "Asymptotic rate analysis of downlink multi-user systems with co-located and distributed antennas," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3046–3058, Jun. 2015.

[8] J. Joung, Y. K. Chia, and S. Sun, "Energy-efficient, large-scale distributed-antenna system (L-DAS) for multiple users," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 954–965, Oct. 2014.

[9] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[10] A. Zanella, "Best practice in RSS measurements and ranging," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2662–2686, 4th Quart., 2016.

[11] N. Patwari, A. O. Hero, M. Perkins, N. S. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2137–2148, Aug. 2003.

[12] A. Girard, C. E. Rasmussen, J. Q. Candela, and R. Murray-Smith, "Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2003, pp. 545–552.

[13] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich, and M. Coulon, "Direct localization for massive MIMO," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2475–2487, May 2017.

[14] S. A. Shaikh and A. M. Tonello, "Localization based on angle of arrival in EM lens-focusing massive MIMO," in *Proc. IEEE Int. Conf. Consum. Electron.-Berlin (ICCE-Berlin)*, Sep. 2016, pp. 124–128.

[15] A. Hu, T. Lv, H. Gao, Z. Zhang, and S. Yang, "An ESPRIT-based approach for 2-D localization of incoherently distributed sources in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 996–1011, Oct. 2014.

[16] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "5G position and orientation estimation through millimeter wave MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[17] A. Guerra, F. Guidi, and D. Dardari, "Position and orientation error bound for wideband massive antenna arrays," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 853–858.

[18] T. Wei, A. Zhou, and X. Zhang, "Facilitating robust 60 GHz network deployment by sensing ambient reflectors," in *Proc. 14th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Mar. 2017, pp. 213–226.

[19] V. Savic and E. G. Larsson, "Fingerprinting-based positioning in distributed massive MIMO systems," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC Fall)*, Sep. 2015, pp. 1–5.

[20] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.

[21] H. C. So and L. Lin, "Linear least squares approach for accurate received signal strength based source localization," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 4035–4040, Aug. 2011.

[22] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, document TS 36.814, 3GPP, Mar. 2010.

[23] Y. J. Bultitude and T. Rautiainen, *IST-4-027756 WINNER II D1.1.2 V1.2 WINNER II Channel Models*, document D1.1.2 V1.2, EBITG, TUI, UOULU, CU/CRC, NOKIA, Espoo, Finland, Sep. 2007.

[24] V. W. Zheng, E. W. Xiang, Q. Yang, and D. Shen, "Transferring localization models over time," in *Proc. AAAI*, 2008, pp. 1421–1426.

[25] S. J. Pan, D. Shen, Q. Yang, and J. T. Kwok, "Transferring localization models across space," in *Proc. AAAI*, 2008, pp. 1383–1388.

[26] V. W. Zheng, S. J. Pan, Q. Yang, and J. J. Pan, "Transferring multi-device localization models using latent multi-task learning," in *Proc. AAAI*, 2008, pp. 1427–1432.

[27] M. Kandemir, "Asymmetric transfer learning with deep Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 730–738.

[28] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proc. IEEE Infocom*, Tel Aviv, Israel, Mar. 2000, pp. 775–784.

[29] S. Kumar, R. M. Hegde, and N. Trigoni, "Gaussian process regression for fingerprinting based localization," *Ad Hoc Netw.*, vol. 51, pp. 1–10, Nov. 2016.

[30] S. Yiu and K. Yang, "Gaussian process assisted fingerprinting localization," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 683–690, Oct. 2016.

[31] B. Roberts and K. Pahlavan, "Site-specific RSS signature modeling for WiFi localization," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2009, pp. 1–6.

[32] Y. Wen, X. Tian, X. Wang, and S. Lu, "Fundamental limits of RSS fingerprinting based indoor localization," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 2479–2487.

[33] J. Xiong and K. Jamieson, "ArrayTrack: A fine-grained indoor location system," in *Proc. 10th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, 2013, pp. 71–84.

[34] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter level localization using WiFi," in *Proc. ACM SIGCOMM*, 2015, pp. 269–282.

[35] S. Sen, J. Lee, K.-H. Kim, and P. Congdon, "Avoiding multipath to revive inbuilding WiFi localization," in *Proc. MobiSys*, 2013, pp. 249–262.

[36] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 763–776, Jan. 2017.

[37] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson. (Aug. 2017). "Deep convolutional neural networks for massive MIMO fingerprint-based positioning." [Online]. Available: https://arxiv.org/abs/1708.06235

[38] A. G. de G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani. (Apr. 2018). "Gaussian process behaviour in wide deep neural networks." [Online]. Available: https://arxiv.org/abs/1804.11271

[39] C. E. Rasmussen, "Evaluation of Gaussian processes and other methods for non-linear regression," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1996.

[40] A. Schwaighofer, M. Grigoraş, V. Tresp, and C. Hoffmann, "GPPS: A Gaussian process positioning system for cellular networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2003, pp. 579–586.

[41] B. Ferris, D. Hähnel, and D. Fox, "Gaussian processes for signal strength-based location estimation," in *Robotics: Science and Systems*. Cambridge, MA, USA: MIT Press, Aug. 2006.

[42] H. Bijl, T. B. Schön, J.-W. van Wingerden, and M. Verhaegen. (2016). *System Identification Through Online Sparse Gaussian Process Regression With Input Noise*. [Online]. Available: http://hildobijl.com/Downloads/SONIG.pdf

[43] L. S. Muppirisetty, T. Svensson, and H. Wymeersch, "Spatial wireless channel prediction under location uncertainty," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1031–1044, Feb. 2016.

[44] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[45] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-93-1, 1993.

[46] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[47] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Low-dimensionality of noise-free RSS and its application in distributed massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 486–489, Aug. 2018.

[48] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.

[49] H. L. Van Trees, K. L. Bell, and Z. Tian, *Detection, Estimation, and Modulation Theory: Part I*. New York, NY, USA: Wiley, 2013.

[50] 4G Americas. (Feb. 2014). *4G Mobile Broadband Evolution: 3GPP Release 11 and Release 12 and Beyond*. [Online]. Available: http://www.5gamericas.org/files/2614/0758/7473/4GMobileBroadbandEvolutionRel11Rel12andBeyondFeb2014FINALv2.pdf

[51] J. Salo, M. Nur-Alam, and K. Chang, "Practical introduction to LTE radio planning," Eur. Commun. Eng., Espoo, Finland, White Paper, Nov. 2010.

[52] Z. Wang, E. K. Tameh, and A. R. Nix, "Joint shadowing process in urban peer-to-peer radio channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, Jan. 2008.
[53] Z. Gao, L. Dai, S. Han, I. Chih-Lin, Z. Wang, and L. Hanzo. (Sep. 2017). "Compressive sensing techniques for next-generation wireless communications." [Online]. Available: https://arxiv.org/abs/1709.01757
[54] E. Hamed, H. Rahul, M. A. Abdelghany, and D. Katabi, "Real-time distributed MIMO systems," in *Proc. ACM SIGCOMM Conf.*, Aug. 2016, pp. 412–425.

K. N. R. Surya Vara Prasad received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology (IIT), Bhubaneswar, India, in 2012, and the M.A.Sc. degree in electrical and computer engineering from The University of British Columbia (UBC), Vancouver, Canada, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. He is an NSERC Alexander Graham Bell Scholar. His current research focus is on machine learning methods and their applications in wireless systems. He was a recipient of the Best Demo and Exhibits Award at the IEEE COMSNETS 2014.

Ekram Hossain (F'15) received the Ph.D. degree in electrical engineering from the University of Victoria, Canada, in 2001. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada. His current research interests include design, analysis, and optimization of wireless, mobile, cognitive, and green communication networks with an emphasis on 5G and beyond, 5G cellular, applied game theory, and network economics. He has authored/edited several books in these areas. He was elevated to an IEEE fellow for his contributions to spectrum management and resource allocation in cognitive and cellular radio networks. He is a member (Class of 2016) of the College of the Royal Society of Canada. He received the 2017 IEEE ComSoc Technical Committee on Green Communications and Computing Distinguished Technical Achievement Recognition Award for outstanding technical leadership and achievement in green wireless communications and networking. He has received several research awards, including the 2017 IEEE Communications Society Best Survey Paper Award, the University of Manitoba Merit Award for Research and Scholarly Activities in 2010, 2013, 2014, and 2015, and the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award. He is the current Editor-in-Chief of the IEEE Press. He served as the Editor-in-Chief for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2012 to 2016, as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in the area of Resource Management and Multiple Access from 2009 to 2011, as an Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING from 2007 to 2012, and as an Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS—Cognitive Radio Series from 2011 to 2014. He served as a Distinguished Lecturer of the IEEE Communications Society. He is currently a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is an Elected Member of the Board of Governors of the IEEE Communications Society for the term 2018–2020. He was listed as a Clarivate Analytics Highly Cited Researcher in computer science in 2017 and 2018.

Vijay K. Bhargava (S'70–M'74–SM'82–F'92–LF'13) was born in Beawar, India, in 1948. He came to Canada in 1966. He received the B.A.Sc., M.A.Sc., and Ph.D. degrees from Queen's University at Kingston in 1970, 1972, and 1974, respectively. He was with the Indian Institute of Science from 1974 to 1975, the University of Waterloo in 1976, the Concordia University from 1976 to 1984, and the University of Victoria from 1984 to 2003. He has held visiting appointments at the Ecole Polytechnique de Montreal, the NTT Research Lab, the Tokyo Institute of Technology, the University of Indonesia, The Hong Kong University of Science and Technology, Tohoku University, and the University of Erlangen-Nuremberg, Germany. He is currently a Professor with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, where he served as the Department Head from 2003 to 2008. He is also an Honorary Professor at UESTC, Chengdu, and a Gandhi Distinguished Professor at IIT Bombay. He is in the Institute for Scientific Information (ISI) Highly Cited list. He served as the Founder and the President of Binary Communications Inc., from 1983 to 2000. He is a co-author/co-editor of seven books, the latest of which is *Wireless-Powered Communication Networks* (Cambridge University Press, 2016). He is a fellow of The Royal Society of Canada, The Canadian Academy of Engineering, and the Engineering Institute of Canada. He is a Foreign Fellow of the National Academy of Engineering, India, and has served as a Distinguished Visiting Fellow at the Royal Academy of Engineering, U.K. He has received awards for his teaching, research, and service to the IEEE. The latest awards are the Killam Prize in Engineering received from the Canada Council for the Arts and the Humboldt Research Prize received from the Alexander von Humboldt Foundation of Germany. A long-time volunteer of the IEEE, he has served as the Director for Region 7 from 1992 to 1993, as the Vice President for the Regional Activities Board-RAB (now MGA) from 1994 to 1995, as the President for the Information Theory Society in 2000, and as the President for the IEEE Communications Society from 2012 to 2013. He is currently the Director of Division III for 2018–2019. He has served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and as the Editor-in-Chief for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.