A Theoretical Analysis on Sampling Size in WiFi Fingerprint-Based Localization

Baoqi Huang, *Member, IEEE*, Runze Yang, Bing Jia, *Member, IEEE*, Wuyungerile Li, and Guoqiang Mao, *Fellow, IEEE*

***Abstract*—This paper deals with a key problem in WiFi fingerprint-based localization, namely how to sample a sufficient number of received signal strength (RSS) measurements during an offline site survey. To this end, a probabilistic framework is firstly presented to characterize the ability of distinguishing two fingerprints, and is then applied in both the ideally infinite sampling case and the realistically finite sampling case with correlated samples. On these grounds, it is shown that the Euclidean distance between the vectors of mean RSS measurements at any two positions is the key factor of determining localization performance, and how several other factors affect localization performance. More importantly, based on the central limit theory, a quantitative analysis is conducted to describe the degradation in localization performance introduced by finite and correlated samples. In addition, provided that correlated samples satisfy the first order autoregressive model, an explicit formula is derived to describe the relationship between the correlation coefficient and the sampling sizes, which can be employed to guide the offline site survey. Extensive simulations are conducted to confirm the effectiveness of the probabilistic framework as well as the correctness of different analytical results, and an experiment is also carried out for validation. This paper not only helps to understand the basic mechanism of WiFi fingerprint-based localization, but also provides insightful guidelines for efficiently building a fingerprint database.**

***Index Terms*—Central limit theory, performance analysis, sampling size, WiFi fingerprint-based localization.**

I. Introduction

WITH the widespread of WiFi (or IEEE 802.11) enabled infrastructures and mobile devices in our daily life, it is promising to develop and deploy WiFi based indoor localization systems. Great efforts have been invested in the past decades to enable reliable and precise WiFi indoor positioning and navigation [1]. In particular, due to its simplicity and tolerance to pervasive multipath effects in indoor environments, WiFi fingerprint-based localization [2]–[6] has gained most attention in both academic and industrial fields.

Basically, WiFi fingerprint-based localization involves two phases, namely an offline site survey phase and an online localization phase. In the offline site survey phase, a fingerprint database consisting of a number of WiFi fingerprints (i.e., consisting of a vector of mean RSS measurements associated with multiple APs) labelled with reference locations within a service area is constructed. In the online localization phase, when a mobile device sends an location query containing its current received signal strength (RSS) measurements from multiple WiFi access points (APs), its location can be inferred by searching the existing fingerprint database through, e.g., the k nearest neighbor (KNN) method.

Though various WiFi fingerprint-based localization techniques [7]–[10] and relevant performance analyses [11]–[15] have been reported, it is still challenging to efficiently develop and deploy WiFi fingerprint-based localization systems [16]–[18]. Specifically, the fingerprint at any reference location is usually represented by the statistical attribute (e.g., mean, variance, and histogram) of associated received signal strength (RSS) measurements from multiple APs, so that building a fingerprint database normally demands labor-intensive and time-consuming measurement campaigns. Sampling size that specifies the number of RSS measurements requested for producing one fingerprint plays a vital role, due to the fact that it not only determines the overheads of building a fingerprint database, but also affects the localization performance [19]. Intuitively, the larger is the sampling size, the better is localization performance, but it is infeasible to conduct infinite sampling. Thus, it is of great value to derive *an efficient sampling size* for the offline site survey, such that having more samples does not significantly improve localization performance, with the result that both manpower and time are substantially saved without sacrificing localization accuracy. To be best of our knowledge, this problem is still open in the literature.

Essentially, WiFi fingerprint-based localization can be cast as a classification problem [20], [21], in the sense that dedicated classification algorithms are employed to select one or multiple most probable fingerprints in the fingerprint database which matches a vector of RSS measurements made in real time from multiple APs, such that the final location can be determined based on the location labels of such selected fingerprints. Therefore, the quality of the fingerprint database, i.e., the ability of discriminating any two fingerprints, determines the resulting localization performance, motivating us to focus on studying how to discriminate one fingerprint from another.

Manuscript received March 18, 2020; revised August 30, 2020 and December 10, 2020; accepted March 11, 2021. Date of publication March 17, 2021; date of current version May 5, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 41871363, 41761086, and 61761035, and in part by the Natural Science Foundation of Inner Mongolia Autonomous Region under Grant 2019MS06030. The review of this article was coordinated by Prof. Takahiro Hara. *(Corresponding author: Bing Jia.)*

Baoqi Huang, Runze Yang, Bing Jia, and Wuyungerile Li are with the College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China (e-mail: cshbq@imu.edu.cn; yangrunze@mail.imu.edu.cn; jiabing@imu.edu.cn; gerile@imu.edu.cn).

Guoqiang Mao is with the Research Institute of Smart Transportation, Xidian University, Xi'an 710071, China (e-mail: g.mao@ieee.org).

Digital Object Identifier 10.1109/TVT.2021.3066380

In general, a fingerprint is produced by using the sample average of sequentially sampled RSS measurements at the corresponding reference location, which is obviously affected by the sampling size and correlations, and is always distinguishable from most other fingerprints in a fingerprint database except a few distant but similar fingerprints (where the similarity is often measured by, e.g., Euclidean distance), which are responsible for introducing significant localization errors. As such, understanding how to discriminate two fingerprints in various circumstances is a fundamental approach for investigating the influence of sampling size, and also paves the way for analyzing localization performance, calibrating fingerprint databases, designing advanced localization algorithms, and etc.

Therefore, in this paper, a probabilistic framework is initially established to characterize the probability of discriminating one fingerprint from another, in both the ideally infinite sampling case and realistically finite sampling with correlation. Then, as a mathematical tool, the framework is leveraged to investigate the influences of different factors on localization performance. As a result, it is concluded that the localization performance only relies on the Euclidean distance between the vectors of the mean RSS measurements associated with such two fingerprints as well as the corresponding standard deviation. Moreover, by using the central limit theorem, the influences of the finite sampling size and correlation between RSS measurements are further investigated. In addition, supposing correlated samples satisfying the first order autoregressive model, a formula is obtained and shows the mathematical relationship among the localization performance, sampling sizes in both uncorrelated and correlated cases, and correlation coefficient, suggesting the efficient sampling size for developing WiFi fingerprint-based localization systems. Extensive simulations and an experiment are carried out and confirm the effectiveness of the probabilistic framework as well as various results obtained.

The rest of the paper is organized as follows. Section II reviews the performance analysis studies. Section III and Section IV respectively establish the theory of localization performance analysis with infinite and finite sampling. In Section V, both simulations and experimental results are reported. We conclude this paper and shed light on future works in Section VI.

II. Related Works

Due to the complexities of indoor wireless signal propagations, it is challenging to characterize the performance of WiFi fingerprint-based localization. Most existing results are obtained through either experiments or simulations [12], [22], and theoretical studies are quite limited.

In [23], the average and probability of the localization error were formulated, but closed-form formulas were unavailable, so that only simulation results were presented. In [11], a preliminary analytical model based on the Euclidean distance between vectors of RSS measurements was developed for a localization system with simplified assumptions on signal propagation and system design, and the influences of the AP number and signal propagation parameters were investigated. In [24], an analytical model that employs proximity graphs for predicting localization performance was developed and also verified through simulations. This model can be used to analyze the internal structure of fingerprints, so as to identify and eliminate unnecessary location fingerprints stored in a fingerprint database, thereby saving on computation while performing location estimation. In [25], an analytic expression for the cumulative distribution function (CDF) of the localization error was presented to investigate the effects of the number of fingerprints and the distance between adjacent fingerprints on the localization error, which were further verified through experiments. In [26], the probability density function (PDF) of the localization error was formulated and further approximated by using nonparametric kernel density estimation techniques. As a result, accurate online evaluation of the localization error is possible, but it is difficult to derive common knowledge about WiFi fingerprint-based localization. In [14], the Cramer-Rao Lower Bound (CRLB) was leveraged to investigate the fundamentals of WiFi based localization, revealing that how the number of APs and RSS gradients affect localization performance. In [13], a general but complicated probabilistic model was presented to investigate the accuracy and reliability of WiFi fingerprint-based localization with different numbers of RSS measurements from one or more AP during the online localization phase. In [15], a new localization error bound was derived by analogizing WiFi based localization into one of information propagation in a parallel Gaussian noisy channel and turned out to outperform the CRLB. However, all these studies cannot directly and comprehensively answer how WiFi fingerprint-based localization is affected by different factors.

Besides, there exist extensive studies relying on, e.g., pedestrian dead reckoning (PDR) [27], crowdsourcing [17], and etc., to alleviate the overheads of the offline site survey. In what follows, we shall review some studies on the optimization of the key parameters in the offline site survey.

In [28], the relationship between localization accuracy and the distance between adjacent reference locations was investigated, and an optimization method was developed based on the Gaussian process model to balance the survey workload and localization accuracy. Experiments were conducted to validate the efficiency of the mechanism, and show that the method can largely reduce the workload of the site survey. In [16], an algorithm was developed to determine the minimum number of reference locations which can be randomly distributed within the area of interest, in order to achieve the predefined localization accuracy with a desired confidence level. Both experiments and simulations were conducted and validated the effectiveness of the proposed algorithm.

In summary, though extensive efforts have been invested on WiFi fingerprint-based localization as well as its performance analyses, there still exist obvious gaps between the realistic localization systems and various analytical models. Specifically, existing studies focused on either designing advanced fingerprint matching algorithms used in the online localization phase, or controlling the density of reference locations in the offline site survey phase, but ignored the problem of how many RSS measurement samples are sufficient at each reference location in the offline site survey phase in order to produce a nearly ideal fingerprint. In this paper, we shall first propose a probabilistic

TABLE I
NOTATIONS DEFINITION

Symbol	Description
p	the AP index, $p \in \mathbb{N}^+$
$\mathbf{x}$	the RSS measurement vector from p APs
σ	the standard deviation of $\mathbf{x}$
x_p	the RSS measurement from the p-th AP
i	the position index, $i \in \mathbb{N}^+$
$\mathbf{m}_i$, $\hat{\mathbf{m}}_i$	the mean and sample mean of RSS measurements from p APs at the i-th reference location
$\mathbf{I}_p$	the identity matrix of order p
$\mathbf{l}_i$	the coordinates of the i-th reference location
$\Pr(\cdot)$	the probability of an event
$f_{m,\sigma}(\cdot)$	the PDF of a normal variable with the mean m and standard deviation σ
$\Phi(\cdot)$	the CDF of a standard normal variable
$\Phi_{0,\sigma}(\cdot)$	the CDF of a normal variable with zero mean and the standard deviation σ
$\Vert \cdot \Vert$	the Euclidean distance operator
$\mathrm{E}_{\hat{\mathbf{m}}}$	the expectation taken with respect to $\hat{\mathbf{m}}$
n_u, n_c	the sampling sizes in the finite uncorrelated and finite correlated cases
σ_u, σ_c	the standard deviations of RSS measurement samples in the finite uncorrelated and finite correlated cases
$\mathbf{R}$, $\mathbf{H}$	an orthogonal matrix and a Hessian matrix
α	the degree of autocorrelation of the original RSS measurements, $\alpha \in (0, 1)$
$Var(\cdot)$	the variance operator

framework to analyze the performance of WiFi fingerprint-based localization, and then solve the problem in relation to the sampling size from a theoretical approach.

III. Performance Analysis of Infinite Sampling

In view of WiFi fingerprint-based localization techniques, if two fingerprints are similar to each other, it is hard for the online localization phase adopting say the KNN method to discriminate one from the other given an arbitrary location query at nearby locations, such that the localization result tends to be erroneous. In other words, given a WiFi fingerprint based localization system, the key factor for determining its localization performance is the ability of discriminating any two fingerprints regardless of how many fingerprints involved.

Therefore, we introduce a probabilistic framework by formulating the probability of discriminating two fingerprints from each other in an ideal case, namely that the mean RSS measurement associated with each fingerprint is derived through infinite sampling. For ease of presentation, we define the following symbols used in this paper in Table I.

A. WiFi RSS Measurement Model

As was commonly assumed [11], [13], [24], the RSS measurements of the signals propagated from p APs to a receiver at the position $\mathbf{l}$, denoted $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$, are independently and normally distributed, namely

$$\mathbf{x} \sim N(\mathbf{m}(\mathbf{l}), \sigma^2 \mathbf{I}_p), \tag{1}$$

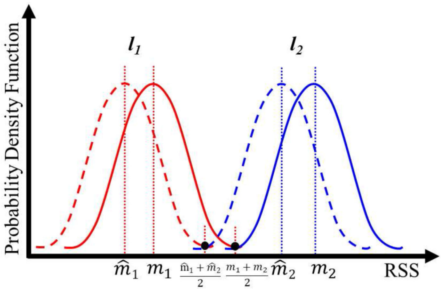

Fig. 1. The illustration of infinite sampling and finite sampling.

where $\mathbf{m}(\mathbf{l}) \in \mathbb{R}^p$ is a vector function calculating the mean RSS measurements (in dBm) at any given position $\mathbf{l} \in \mathbb{R}^2$ from p APs, and $\mathbf{I}_p$ is the identity matrix of order p.

Regarding the i-th reference location, the corresponding vector of the mean RSS measurements is abbreviated as $\mathbf{m}_i$, and if only one AP (i.e., $p = 1$) is considered, the mean RSS measurement is represented by m_i.

B. Discriminating Two Fingerprints

Instead of directly measuring localization errors, the probability of discriminating different fingerprints given a RSS measurement is defined as the metric for localization performance.

The following Lemma takes into account the simplest case that two fingerprints with only one AP are involved.

Lemma 1: Suppose that RSS measurements made at $\mathbf{l}_1$ and $\mathbf{l}_2$ are independently and normally distributed with means m_1 and m_2, respectively, and the standard deviation σ. Then, the probability that an arbitrary RSS measurement made at $\mathbf{l}_1$ is classified as belonging to $\mathbf{l}_1$ other than $\mathbf{l}_2$ is

$$\Pr(f_{m_1,\sigma}(x) > f_{m_2,\sigma}(x)) = \Phi\left(\frac{|m_2 - m_1|}{2\sigma}\right), \tag{2}$$

where $f_{m_1,\sigma}(\cdot)$ denotes the PDF of a normal variable with the mean m_1 and standard deviation σ, and $\Phi(\cdot)$ denotes the CDF of a standard normal variable.

Proof: Without loss of generality, suppose $m_1 < m_2$. An arbitrary RSS measurement made at $\mathbf{l}_1$, say $x \sim N(m_1, \sigma^2)$, is classified as belonging to $\mathbf{l}_1$ other than $\mathbf{l}_2$ if and only if it is more probable to obtain x at $\mathbf{l}_1$ than $\mathbf{l}_2$, i.e.,

$$f_{m_1,\sigma}(x) > f_{m_2,\sigma}(x). \tag{3}$$

As is depicted in Fig. 1, the two solid curves represent the true PDFs of RSS measurements at $\mathbf{l}_1$ and $\mathbf{l}_2$ from one AP, and it is evident that (3) holds if and only if the integral range is $x \in (-\infty, (m_1 + m_2)/2)$. As such, the probability that x is correctly classified as belonging to $\mathbf{l}_1$ other than $\mathbf{l}_2$ is

$$\Pr(f_{m_1,\sigma}(x) > f_{m_2,\sigma}(x)) = \int_{-\infty}^{\frac{m_1+m_2}{2}} f_{m_1,\sigma}(x)dx$$
$$= \int_{-\infty}^{\frac{m_1+m_2}{2}} f_{0,\sigma}(x - m_1)dx = \int_{-\infty}^{-\frac{m_1-m_2}{2}} f_{0,\sigma}(x)dx$$
$$= \Phi_{0,\sigma}\left(\frac{m_2 - m_1}{2}\right) = \Phi\left(\frac{m_2 - m_1}{2\sigma}\right),$$

where $\Phi_{0,\sigma}(\cdot)$ is the CDF of the normal variable $N(0,\sigma^2)$. Similarly, the formulation of the probability in the case of $m_1 > m_2$ can be derived. Finally, the lemma is proved. ■

Then, considering the case of two fingerprints with more APs, we can obtain the following theorem.

Theorem 1: Suppose that the vectors of RSS measurements from p APs and made at $\mathbf{l}_1$ and $\mathbf{l}_2$ are independently and normally distributed with mean vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ and the standard deviation σ. Then, the probability that an arbitrary RSS measurement vector made at $\mathbf{l}_1$ is classified as belonging to $\mathbf{l}_1$ other than $\mathbf{l}_2$ is

$$\Pr(f_{\mathbf{m}_1,\sigma}(\mathbf{x}) > f_{\mathbf{m}_2,\sigma}(\mathbf{x})) = \Phi\left(\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|}{2\sigma}\right). \tag{4}$$

Proof: Let $\mathbf{x}$ be a vector of RSS measurements made at $\mathbf{l}_1$. Similar to the treatments with one AP, the necessary and sufficient condition of correctly classifying $\mathbf{x}$ is that the Euclidean distance between $\mathbf{x}$ and $\mathbf{m}_1$ is less than that between $\mathbf{x}$ and $\mathbf{m}_2$. Then, it is simple to obtain (4). ■

Remark 1: Theorem 1 essentially simplifies the continuous area for localization into a set of discrete reference locations. Since we are not concerned with the precise values of localization errors, but intend to evaluate how localization performance is changed with different factors, it is sufficient to adopt the probabilistic framework in the performance analysis. Moreover, (4) holds if and only if the Euclidean distance between $\mathbf{x}$ and the mean RSS measurement vector at $\mathbf{l}_1$ is less than that associated with $\mathbf{x}$ and $\mathbf{l}_2$, which is factually the same as the comparing approach adopted in the online localization phase. Therefore, the proposed framework is congruent with the practical WiFi fingerprint-based localization techniques.

Remark 2: Irrespective of the number of APs (i.e., p), the similarity between any two fingerprints, namely $\|\mathbf{m}_2 - \mathbf{m}_1\|$, plays a vital role in discriminating the corresponding two reference locations, in the sense that reducing the similarity by increasing $\|\mathbf{m}_2 - \mathbf{m}_1\|$ contributes to localization performance. Moreover, for an existing localization system, adding one or more extra AP (namely increasing p), will definitely increase $\|\mathbf{m}_2 - \mathbf{m}_1\|$, so that the probability in (4) will rise as well and localization performance can be improved. In addition, the standard deviation of RSS measurements at any reference location, i.e., σ, is another important factor of determining localization performance, and it is intuitive that reducing σ tends to improve localization performance.

Remark 3: The result in Theorem 1 is consistent with existing relevant studies. For instance, in [14], a CRLB based approach is applied to analyze the performance of WiFi fingerprint-based localization, revealing that the larger is the gradient of the mean RSS measurement at any location, the better is the localization performance, which evidently admits Theorem 1; in [29], the difference between the mean RSS measurements of two fingerprints is empirically adopted as a metric for optimally deploying WiFi APs.

Remark 4: As was mentioned before, discriminating two fingerprints is key to the localization performance of a WiFi fingerprint-based localization system. For instance, provided that a pair of fingerprints associated with two distant reference locations in a fingerprint database have high similarity, RSS measurements sampled at one of them are prone to be classified as belonging to the other, with the result that large localization errors are eventually incurred. According to Theorem 1, one can reduce such localization errors by introducing a new AP(s) or moving an existing AP(s) to an appropriate position, such that the similarity is lowered. Therefore, evaluating and improving the quality of a fingerprint database can be implemented based on discriminating two fingerprints. However, since it is beyond the main purpose of this paper, we shall work towards it in our future works.

In summary, the proposed probabilistic framework based on formulating the probability of discriminating two fingerprints is correct and feasible, and also as a basis, paves the way for further analysis in practical scenarios.

IV. Performance Analysis of Finite Sampling

Since knowing an exact mean RSS measurement demands infinite sampling, which is infeasible in practice, sample average based on finite sampling is alternatively used as approximations. Thus, in what follows, the performance analysis shall be conducted given a limited sampling size. In addition, we take into account two fingerprints with multiple APs.

A. Discriminating Two Fingerprints With Finite Samples

Given any AP and any reference location, due to the limited size of RSS measurements made in the offline site survey, the resulting sample average usually deviates from the true mean value, as depicted in Fig. 1 where the dashed curves represent the approximate PDFs with the sample averages $\hat{m}_1$ and $\hat{m}_2$ as the mean values. As a result, fingerprint-based localization algorithm has to employ inaccurate fingerprints during localization. Then, we can obtain the following theorem.

Theorem 2: Suppose that RSS measurements made at $\mathbf{l}_1$ and $\mathbf{l}_2$ from p APs are independently and normally distributed with means $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively, and the standard deviation σ. Then, given the sampling size n_u, the probability that an arbitrary vector of RSS measurements made at $\mathbf{l}_1$, denoted $\mathbf{x}$, is classified as belonging to $\mathbf{l}_1$ other than $\mathbf{l}_2$ is

$$\Pr(f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x}) > f_{\hat{\mathbf{m}}_2,\sigma}(\mathbf{x})) = \mathrm{E}_{\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2}\left(\Phi\left(\frac{b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2)}{\sigma}\right)\right), \tag{5}$$

where $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$ are the sample averages, $\mathrm{E}_{\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2}$ is the expectation taken with respect to $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$, and

$$b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2) = \frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)^T}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}\left(\frac{\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2}{2} - \mathbf{m}_1\right) \tag{6}$$

Proof: $\Pr(f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x}) > f_{\hat{\mathbf{m}}_2,\sigma}(\mathbf{x}))$ can be evaluated by calculating the integral of $f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x})$ when $f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x}) > f_{\hat{\mathbf{m}}_2,\sigma}(\mathbf{x})$, which will hold as long as $\mathbf{x}$ is closer to $\hat{\mathbf{m}}_1$ than $\hat{\mathbf{m}}_2$, i.e., the integral range (denoted D) of $\mathbf{x}$ is defined by

$$\begin{aligned}
&(\mathbf{x} - \hat{\mathbf{m}}_1)^T(\mathbf{x} - \hat{\mathbf{m}}_1) \le (\mathbf{x} - \hat{\mathbf{m}}_2)^T(\mathbf{x} - \hat{\mathbf{m}}_2) \\
&\Leftrightarrow \frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)^T}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}(\mathbf{x} - \mathbf{m}_1) \le b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2) \\
&\Leftrightarrow \frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)^T}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}\mathbf{R}\mathbf{x}' \le b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2)
\end{aligned}$$

where

$$\mathbf{x}' = \mathbf{R}^T(\mathbf{x} - \mathbf{m}_1) \tag{7}$$

and $\mathbf{R}$ is the orthogonal matrix satisfying

$$\frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)^T}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}\mathbf{R} = [1\ 0 \cdots 0]. \tag{8}$$

It is clear that the first column of $\mathbf{R}$ equals to $\frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}$. As such, for $\mathbf{x}'$, its first element, denoted x_1', satisfies

$$x_1' = \frac{(\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1)^T}{\|\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1\|}(\mathbf{x} - \mathbf{m}_1) \le b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2), \tag{9}$$

and all the other elements are free. From the perspective of vector view, $b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2)$ is actually the projection distance of the vector from $\mathbf{m}_1$ to $\frac{\hat{\mathbf{m}}_1 + \hat{\mathbf{m}}_2}{2}$ on the vector $\hat{\mathbf{m}}_2 - \hat{\mathbf{m}}_1$.

Given a specific pair of $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$, we have

$$\begin{aligned}
&\oint_D f_{m_1^1,\sigma}(x_1)\cdots f_{m_1^p,\sigma}(x_p)dx_1\cdots dx_p \\
&= \oint_D \frac{1}{(\sqrt{2\pi})^{p-1}} f_{0,\sigma}(\|\mathbf{R}^T(\mathbf{x} - \mathbf{m}_1)\|)dx_1\cdots dx_p \\
&= \oint_D \frac{1}{(\sqrt{2\pi})^{p-1}} f_{0,\sigma}(\|\mathbf{x}'\|)dx_1'\cdots dx_p' \\
&= \oint_D f_{0,\sigma}(x_1')\cdots f_{0,\sigma}(x_p')dx_1'\cdots dx_p' \\
&= \Phi\left(\frac{b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2)}{\sigma}\right).
\end{aligned} \tag{10}$$

Thus, the theorem is proved. ■

Remark 5: Following the proof of Theorem 2, we can put it further by transforming $\mathbf{m}_2$ and $\mathbf{m}_1$ to any other two vectors with their Euclidean distance equal to $\|\mathbf{m}_2 - \mathbf{m}_1\|$, and obtain the same value of (5), namely that the probability (5) relies on $\|\mathbf{m}_2 - \mathbf{m}_1\|$ other than the specific vectors $\mathbf{m}_1$ or $\mathbf{m}_2$. This is the same as in the infinite sampling case.

Furthermore, it is of great value to understand how localization performance is affected by the sampling size, i.e., n_u. The following theorem describes the degradation in localization performance caused by the finite sampling size n_u.

Theorem 3: Provided that the same scenario as in Theorem 2 is considered and the sample averages $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$ are sufficiently close to their true means $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively, the probabilities given in (4) and Theorem 2 satisfy

$$\begin{aligned}
&\Pr(f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x}) > f_{\hat{\mathbf{m}}_2,\sigma}(\mathbf{x})) \approx \Pr(f_{\mathbf{m}_1,\sigma}(\mathbf{x}) > f_{\mathbf{m}_2,\sigma}(\mathbf{x})) \\
&- \frac{\sigma_u^2 \exp\left(-\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|^2}{8\sigma^2}\right)}{\sqrt{2\pi}}\left(\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|}{4\sigma^3} + \frac{(p-1)}{\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|}\right)
\end{aligned} \tag{11}$$

where σ_u denotes the standard deviation of the sample average $\hat{\mathbf{m}}_1$, and it follows from the central limit theory that

$$\sigma_u = \frac{\sigma}{\sqrt{n_u}}. \tag{12}$$

Proof: First of all, since $b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2)$ is infinitely differentiable except $\hat{\mathbf{m}}_1 = \hat{\mathbf{m}}_2$, we can apply the Taylor series expansions on $\Phi(\frac{b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2)}{\sigma})$ around $(\mathbf{m}_1, \mathbf{m}_2)$ with $\mathbf{m}_1 \neq \mathbf{m}_2$ and obtain

$$\begin{aligned}
&\mathrm{E}_{\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2}\left(\Phi\left(\frac{b_{\mathbf{m}_1}(\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2)}{\sigma}\right)\right) \approx \Phi\left(\frac{b_{\mathbf{m}_1}(\mathbf{m}_1, \mathbf{m}_2)}{\sigma}\right) \\
&+ \mathrm{E}_{\hat{\mathbf{m}}_1,\hat{\mathbf{m}}_2}\left(\begin{bmatrix}\hat{\mathbf{m}}_1 - \mathbf{m}_1 \\ \hat{\mathbf{m}}_2 - \mathbf{m}_2\end{bmatrix}^T \mathbf{H} \begin{bmatrix}\hat{\mathbf{m}}_1 - \mathbf{m}_1 \\ \hat{\mathbf{m}}_2 - \mathbf{m}_2\end{bmatrix}\right)
\end{aligned} \tag{13}$$

where $\mathbf{H}$ is the Hessian matrix of $\Phi(\frac{b_{\mathbf{m}_1}(\mathbf{m}_1, \mathbf{m}_2)}{\sigma})$.

Secondly, by applying fundamental and tedious matrix calculus operations, we can obtain the blocks of $\mathbf{H}$ on the main diagonal as follows.

$$\begin{aligned}
\mathbf{H}_{11} = &\Phi'' \frac{(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T}{4\sigma^2\|\mathbf{m}_2 - \mathbf{m}_1\|} \\
&- \Phi'\left(\frac{3\mathbf{I}_p}{2\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|} - \frac{3(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T}{2\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|^3}\right)
\end{aligned} \tag{14}$$

$$\begin{aligned}
\mathbf{H}_{22} = &\Phi'' \frac{(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T}{4\sigma^2\|\mathbf{m}_2 - \mathbf{m}_1\|} \\
&+ \Phi'\left(\frac{\mathbf{I}_p}{2\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|} + \frac{(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T}{2\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|^3}\right).
\end{aligned} \tag{15}$$

where

$$\Phi' = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|^2}{8\sigma^2}\right) \tag{16}$$

$$\Phi'' = -\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|}{2\sqrt{2\pi}\sigma}\exp\left(-\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|^2}{8\sigma^2}\right) \tag{17}$$

Thirdly, with σ_u we have

$$\mathrm{E}_{\hat{\mathbf{m}}_1}\left((\hat{\mathbf{m}}_1 - \mathbf{m}_1)(\hat{\mathbf{m}}_1 - \mathbf{m}_1)^T\right) = \sigma_u^2\mathbf{I}_p. \tag{18}$$

Finally, by substituting (14), (15), (16), (17) and (18) into (13), we can obtain (11). Thus, the theorem is proved.

Remark 6: The negative difference given in (11) indicates that localization performance is degraded due to finite sampling; specifically, the degradation is approximately proportional to σ_u^2, and also depends on $\|\mathbf{m}_2 - \mathbf{m}_1\|$. Supposing that $\|\mathbf{m}_2 - \mathbf{m}_1\|$ rises up from 0, $\exp(-\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|^2}{8\sigma^2})$ drops from 1 to 0, whereas $(\frac{\|\mathbf{m}_2 - \mathbf{m}_1\|}{4\sigma^3} + \frac{(p-1)}{\sigma\|\mathbf{m}_2 - \mathbf{m}_1\|})$ initially drops and then gradually rises up after $\|\mathbf{m}_2 - \mathbf{m}_1\| > 2\sigma\sqrt{p-1}$. Therefore, it is conjectured that the degradation will diminish with $\|\mathbf{m}_2 - \mathbf{m}_1\|$ increasing.

Remark 7: According to (11), the localization performance is dependent on the following parameters, σ, $\|\mathbf{m}_2 - \mathbf{m}_1\|$, p and n_u, where σ and $\|\mathbf{m}_2 - \mathbf{m}_1\|$ are actually dependent on wireless channels, p is often predetermined, and only n_u is an optional parameter relying on the offline site survey. Therefore, gain in localization performance can be attained by making more samples, namely increasing n_u, but will die off with n_u above a threshold due to the inversely proportional relationship. As such, (11) enables us to find out efficient sampling sizes in the finite and uncorrelated sampling case, so as to avoid wasting manpower and time in collecting excessive RSS measurements in the offline site survey.

B. Discriminating Two Fingerprints With Finite and Correlated Samples

When RSS measurements associated with finite sampling are also correlated, (18) will not hold any more. As a result, we can readily derive the following corollary about the influence of finite and correlated samples on localization performance.

Corollary 1: The same scenario as in Theorem 2 is considered, except that the RSS measurements between the same pair of AP and reference location are correlated. Then, the probabilities given in (4) and Theorem 2 satisfy

$$\Pr(f_{\hat{\mathbf{m}}_1,\sigma}(\mathbf{x}) > f_{\hat{\mathbf{m}}_2,\sigma}(\mathbf{x})) \approx \Pr(f_{\mathbf{m}_1,\sigma}(\mathbf{x}) > f_{\mathbf{m}_2,\sigma}(\mathbf{x})) - \frac{\sigma_c^2 \exp\left(-\frac{\|\mathbf{m}_2-\mathbf{m}_1\|^2}{8\sigma^2}\right)}{\sqrt{2\pi}} \left(\frac{\|\mathbf{m}_2-\mathbf{m}_1\|}{4\sigma^3} + \frac{(p-1)}{\sigma\|\mathbf{m}_2-\mathbf{m}_1\|}\right). \tag{19}$$

where $\sigma_c^2\mathbf{I}_p$ denotes the covariance matrix of $\hat{\mathbf{m}}_1$ obtained by using n correlated samples, namely

$$\mathrm{E}_{\hat{\mathbf{m}}_1}\left((\hat{\mathbf{m}}_1 - \mathbf{m}_1)(\hat{\mathbf{m}}_1 - \mathbf{m}_1)^T\right) = \sigma_c^2\mathbf{I}_p. \tag{20}$$

Remark 8: Similarly to the cases of infinite sampling in Theorem 2 and finite sampling without correlation in Theorem 3, the difference in the mean RSS measurements still plays a vital role, but the standard deviation σ_c, which relies on sample correlation, essentially determines the magnitude of the diminishing effect caused by sample correlation. In other words, the degradation in localization performance is approximately proportional to σ_c^2, indicating that one should pay attention to the value of σ_c in practice.

C. Analysis Based on the First Order Autoregressive Model

Treat the RSS measurements from one AP as a discrete stationary time series, which can be modelled by a first order autoregressive model [30]. Particularly, let s_t be the stationary time series representing the RSS measurement from one AP at time t. Then, s_t can be represented as follows

$$s_t = (1-\alpha)m + \alpha s_{t-1} + v_t, \tag{21}$$

where the correlation coefficient α lies between 0 and 1, and v_t denotes a white noise process independent from s_t. In essence, α is a parameter that determines the degree of autocorrelation of the original RSS measurements. Moreover, different samples from v_t are identically and independently distributed, i.e., $v_t \sim N(0, \sigma_v^2)$.

On these grounds, we can obtain the following theorem about the sampling distribution given correlated samples.

Theorem 4: Suppose that RSS measurements made at any location from any AP follows the first order autoregressive model in (21). Then, given the sampling size n_c, the sample average is normally distributed with mean m and variance

$$\sigma_c^2 = \frac{\sigma^2}{n_c}\left(\frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^{n_c})}{n_c(1-\alpha)^2}\right). \tag{22}$$

Proof: First of all, since the RSS measurements are normal, regardless of their correlation, it is evident that the sample average is also normally distributed with the same mean, i.e., m. Then, it follows from (21) that

$$s_t = (1-\alpha^t)m + \alpha^t s_0 + \sum_{i=1}^{t} \alpha^{t-i} v_t. \tag{23}$$

Since s_t follows the normal distribution with mean m and standard deviation σ, we can obtain

$$\sigma_v^2 = (1-\alpha^2)\sigma^2. \tag{24}$$

Given n sequential correlated RSS measurements, the variance of the sample average has the below formulation.

$$\begin{aligned} \sigma_c^2 &= Var\left(\frac{1}{n_c}\sum_{j=0}^{n_c-1} s_j\right) \\ &= \frac{1}{n_c^2} Var\left(\frac{1-\alpha^{n_c}}{1-\alpha} s_0 + \sum_{j=1}^{n_c-1}\sum_{i=1}^{j} \alpha^{j-i} v_i\right) \\ &= \frac{\sigma^2}{n_c}\left(\frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^{n_c})}{n_c(1-\alpha)^2}\right), \end{aligned}$$

where $Var(\cdot)$ denotes the variance operation. ■

In what follows, we intend to make comparisons between the cases with finite correlated samples and finite uncorrelated samples based on Theorem 4.

Remark 9: Firstly, suppose that the same number of RSS samples are obtained in both cases, namely $n_u = n_c$, and then, it follows from Theorem 4 that

$$\frac{\sigma_c^2}{\sigma_u^2} = \frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^{n_c})}{n_c(1-\alpha)^2} \tag{25}$$

which is only dependent on α and n_c, and has nothing to do with the noise level σ. According to Theorem 3 and Corollary 1, σ_u and σ_c respectively determine the degradations in localization performance achieved in the two cases in comparison with the infinite sampling case, so that the ratio actually characterizes the degree of the degradation induced by sample correlation.

Remark 10: Secondly, suppose that different numbers of RSS samples are obtained in both cases but produce the same standard deviation of the sample averages, namely $\sigma_c = \sigma_u$. It follows from Theorem 4 that

$$\frac{n_c}{n_u} = \frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^{n_c})}{n_c(1-\alpha)^2}. \tag{26}$$

The significance of (26) is two-fold.

- The efficient sampling size can be obtained through (26): firstly, the efficient sampling size in the finite uncorrelated case (i.e., n_u) can be determined, in the sense that using as few as n_u uncorrelated samples to produce one fingerprint is able to approximately achieve the localization performance with infinite sampling; secondly, the correlation coefficient (i.e., α) in the target space for site survey can be empirically or experimentally evaluated; thirdly, given specific α and n_u, the efficient sampling size in the finite correlated case (i.e., n_c) can be calculated based on (26).

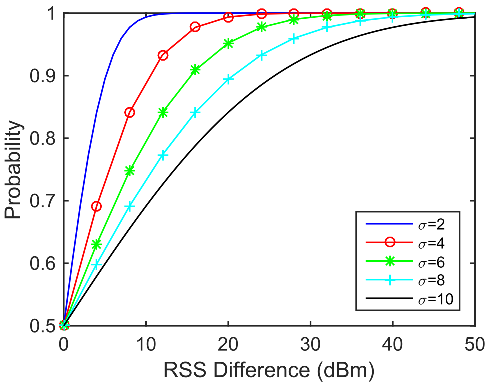

Fig. 2. The probability of successful localization with infinite sampling.

- Given α and n_c, (26) determines a unique n_u, which essentially functions as a scalar indicator of localization performance in light of (7), so that different values of α and n_c can be intuitively compared through n_u for the purpose of evaluating different sampling strategies.

To sum up, we establish the relationship between localization performance and several other factors with uncorrelated and correlated samples in the finite sampling cases, which can be used to guide the efficient construction of fingerprint databases in practice.

V. Simulations and Experiments

In this section, both simulations and experiments are carried out to validate the proposed framework and results.

A. Simulations for Infinite Sampling

Based on (4), the probabilities of successful localization with respect to different values of σ and differences in the mean RSS measurements are plotted in Fig. 2. It is shown that, the probability is always greater than 0.5, indicating a at least fifty-fifty chance to discriminate two fingerprints, and normally rises up with the difference increasing; if the difference is sufficiently large, the probability approaches to 1, meaning that discriminating such two fingerprints is almost surely successful. For instance, when the difference is above 50 dB, the probabilities approach to 1 regardless of the value σ. Additionally, it can be observed that increasing σ significantly reduces the probabilities. For instance, given a difference around 10 dB, the probability is nearly 1 when $\sigma = 2$ dB, but reduces to around 0.7 when $\sigma = 10$ dB.

B. Simulations for Finite Sampling

Firstly, in order to validate the approximate formula in Theorem 3, i.e., (11), the probabilities of successful localization with respect to different values of σ and n are evaluated through both (11) and the numerical integration approach. As depicted in Fig. 3, given $\|\mathbf{m}_2 - \mathbf{m}_1\| = 10$ dB and $p = 1$, the gap between (11) and the numerical result reduces with n increasing, and particularly, when $n = 30$, the curve of (11) almost overlaps with its counterpart derived by the numerical approach regardless of the value of σ. As such, the approximation in Theorem 3 is feasible and thus acceptable.

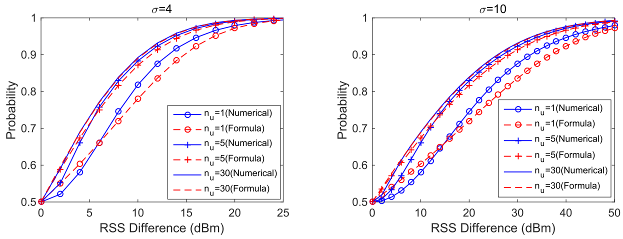

Fig. 3. The probability of successful localization with finite sampling evaluated by (11) and a numerical approach.

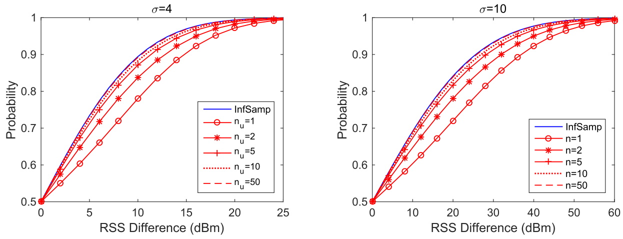

Fig. 4. The probability of successful localization with finite and uncorrelated samples evaluated by (11).

Secondly, given finite uncorrelated samples, the probabilities of successful localization with respect to different values of σ and n_u are plotted in Fig. 4 based on (11), where the probabilities associated with infinite sampling are plotted with the legend "InfSamp" for comparison. Note that the gap between the probabilities associated with finite sampling and the corresponding infinite sampling represents the degradation induced by finite sampling. As such, it can be found that the degradation can be mitigated by increasing either the sampling size n_u, or the difference in the mean RSS measurements (i.e., $\|\mathbf{m}_2 - \mathbf{m}_1\|$), both of which tend to force the last term in (11) to 0. This is also consistent with Remark 6 and 7. For instance, the gap is trivial when $n_u = 5$ regardless of the values of σ and $\|\mathbf{m}_2 - \mathbf{m}_1\|$, and almost disappears when $n_u \geq 10$. Exception happens when $\|\mathbf{m}_2 - \mathbf{m}_1\|$ approaches to 0 in the sense that the gaps do not exist, which is attributable to the fact that having two different reference locations with the identical mean RSS measurement from one AP (i.e., $p = 1$) results in half chance of successful localization.

Moreover, given finite and correlated samples, the probabilities of successful localization with respect to different values of σ, n_c and α are plotted in Fig. 5 based on Corollary 1 and Theorem 4. Similar results can be observed as in the uncorrelated case except that much more samples are required.

Thirdly, considering the case in Remark 9, namely $n_c = n_u$, the ratio $\frac{\sigma_c^2}{\sigma_u^2}$ with respect to α and sampling size n_c is plotted in Fig. 6 according to (25). As can be seen, with n_c increasing, the ratio generally rises; as a result, the degradation in localization performance due to sample correlation is gradually enlarged, but appears to be stable when n_c is above a threshold. In addition, given a specific sampling size, the higher is the value of α, the larger is the ratio, thus resulting a lower probability of successful

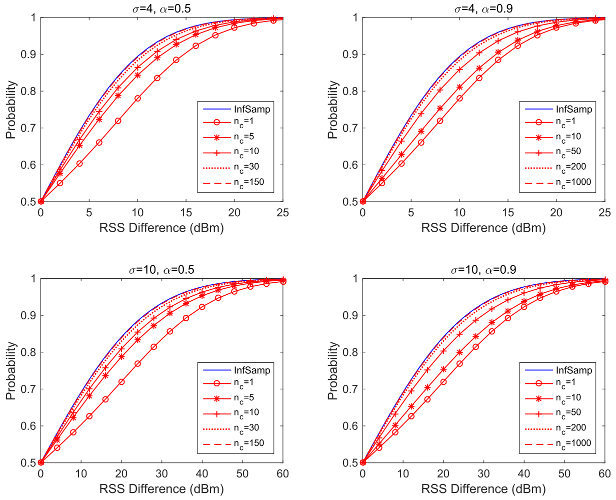

Fig. 5. The probability of successful localization with finite and correlated samples.

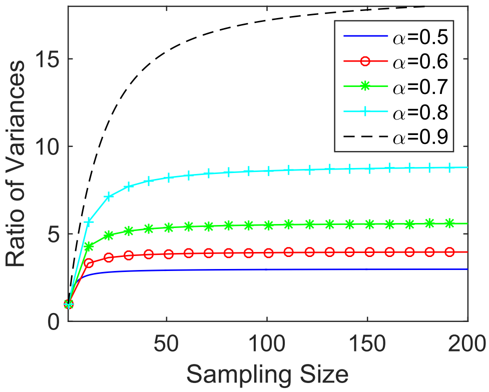

Fig. 6. The ratio $\frac{\sigma_c^2}{\sigma_u^2}$ calculated using (25) with respect to different values of α and different sampling sizes.

localization according to (19) in Corollary 1; that is to say, high correlation in RSS measurements does deteriorate localization performance.

Finally, we shall explain how Remark 10 works in view of the above simulation results. According to Fig. 4, the localization performance of using 10 uncorrelated RSS measurements to produce a fingerprint approaches to that with infinite sampling, namely $n_u = 10$. Supposing that $\alpha = 0.5$ (or $\alpha = 0.9$), it follows from Remark 10 that $n_c = \frac{1+\alpha}{1-\alpha} n_u$, indicating that using $n_c = 30$ (or 190) correlated RSS measurements is able to achieve similar localization performance as using $n_u = 10$ uncorrelated RSS measurements, which evidently conforms to Fig. 5 and suggests that the efficient sampling size should be around 30 (or 190).

C. Experiments

We assigned 24 uniformly distributed reference locations in our lab (see Fig. 7) with the size of 6 m ×12 m, installed four WiFi APs (which enable packet sniffing WiFi) nearby the four corners and finally collected RSS measurement samples during 5 minutes at each reference location. In the experiments, different fingerprint databases are produced based on different RSS sampling strategies and K nearest neighbors (KNN) with $k = 3$ is employed to fulfill the online localization.

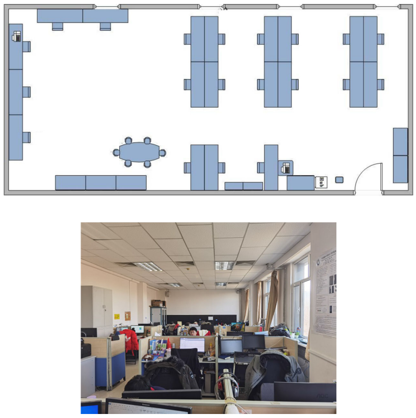

Fig. 7. The floor plan and the real scenario in our lab.

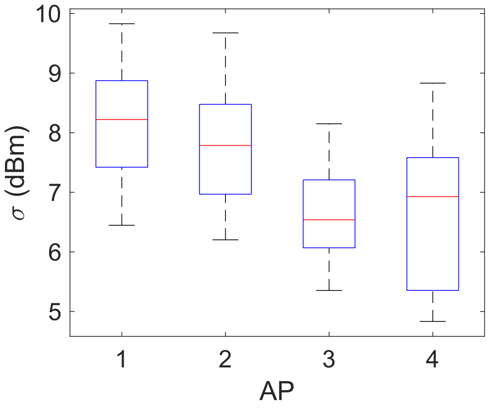

Fig. 8. The standard deviations of RSS measurements at different reference locations with respect to four APs.

To emulate sampling with different correlations, we actually select RSS samples from raw data with different intervals i.e., 1 s, 40 s and 100 s, are used, and as a result, the corresponding sample correlations (i.e., α) are respectively 0.98, 0.88 and 0.77 evaluated by using the Yule-Walker method.

In the first place, we shall investigate the assumptions we have made in this paper. To this end, we plot the standard deviations and histogram of RSS measurements collected at each reference location in Fig. 8 and Fig. 9 respectively. As can be seen, even if the standard deviations are not equal and the histograms are not perfectly normal, the standard deviations are in the same magnitude of order and the histograms are essentially similar to normal distributions, so that the theoretical analysis conducted in this paper is advisable and significative.

In the second place, we produce fingerprint databases by using different numbers of RSS samples (i.e., n_c) at each reference location given a specific α and plot the localization errors in Fig. 10 and Fig. 11. According to (26), a unique n_u is calculated given a pair of n_c and α, and is also depicted in both figures. Specifically, Fig. 10 illustrates that, given a specific α, the localization accuracy improves with the sampling size n_c as

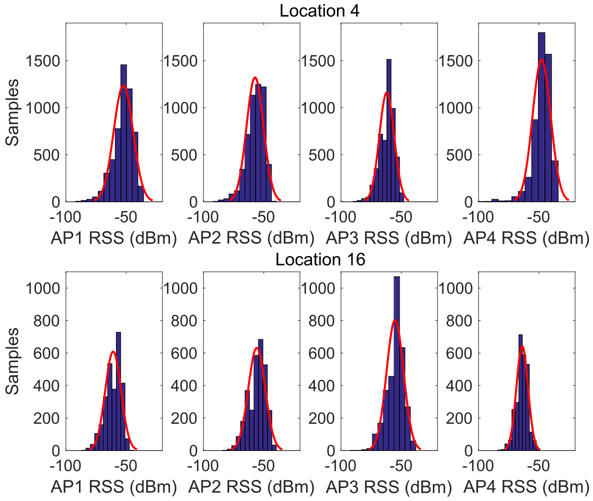

Fig. 9. The histograms of RSS measurements with respect to different AP at two reference locations.

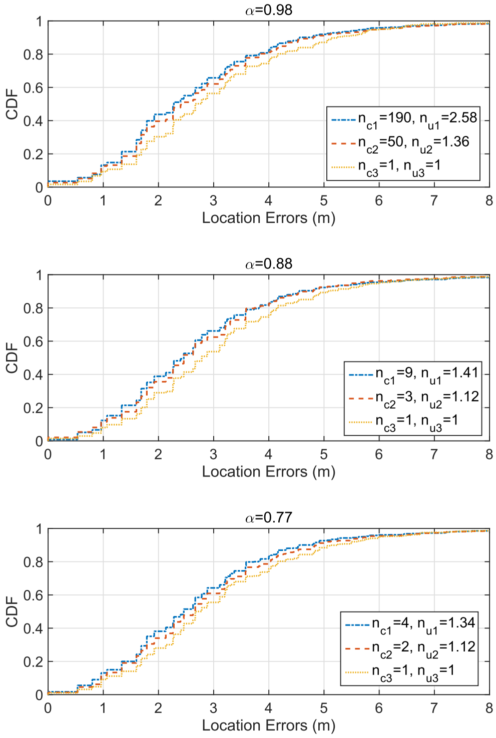

Fig. 10. The localization errors with respect to different sampling strategies.

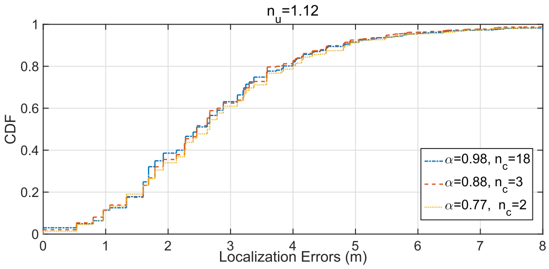

Fig. 11. The localization errors with respect to different sampling strategies given the same value of n_u.

well as n_u, which is consistent with Theorem 3 and Remark 7. Moreover, Fig. 11 shows that, regardless of the values of n_c and α, n_u is the key factor that determines the localization accuracy, which confirms the second part of Remark 10.

According to Fig. 4, when $n_u = 5$, the localization performance has been very close to the infinite sampling case, indicating that it is acceptable to collect 5 uncorrelated RSS samples to produce a fingerprint. However, even if we have made RSS measurements at each reference location for a relatively long period of time, which is far more than usual, the equivalent n_u is only as large as 2.58, implying that there is still room for improving the localization accuracy.

In summary, both the simulations and experiments validate the theoretical analysis and corresponding results in the paper.

VI. Conclusion

In this paper, a probabilistic framework was initially presented to evaluate the performance of WiFi fingerprint-based localization by involving only two fingerprints with infinite sampling, and was then extended to the more practical cases by gradually taking into consideration finite sampling and sample correlation. On these grounds, a theoretical analysis was conducted and revealed the fundamentals in relation to the performance of WiFi fingerprint-based localization. Particularly, the efficient sampling size for producing a fingerprint was thoroughly investigated based on a quantitative analysis. Extensive simulations and experiments were carried out, and confirmed the effectiveness of the proposed framework and the correctness of the corresponding performance analysis. The results in this paper not only help to understand the mechanism of WiFi fingerprint-based localization, but also provide insightful guidelines for efficiently building a fingerprint database.

Regarding future works, we would like to adopt the probabilistic framework to investigate other performance issues in relation to realistic WiFi fingerprint-based localization systems, e.g., how to calibrate a fingerprint database, how to design a superior localization algorithm, and etc.

References

[1] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” *IEEE Trans. Syst., Man, Cybern., Part C. (Appl. Rev.)*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.

[2] P. Bahl and V. N. Padmanabhan, “Radar: An in-building RF-based user location and tracking system,” in *Proc. IEEE INFOCOM Conf. Comput. Commun. 19th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, 2000, pp. 775–784.

[3] H. Zou, B. Huang, X. Lu, H. Jiang, and L. Xie, “A robust indoor positioning system based on the procrustes analysis and weighted extreme learning machine,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1252–1266, Feb. 2016.

[4] H. Zhao, B. Huang, and B. Jia, “Applying kriging interpolation for wifi fingerprinting based indoor positioning systems,” in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2016, pp. 1–6.

[5] K. Majeed, S. Sorour, T. Y. Al-Naffouri, and S. Valaee, “Indoor localization and radio map estimation using unsupervised manifold alignment with geometry perturbation,” *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2794–2808, Nov. 2016.

[6] B. Huang, G. Mao, Y. Qin, and Y. Wei, “Pedestrian flow estimation through passive wifi sensing,” *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1529–1542, Apr. 2021.

[7] P. Jiang, Y. Zhang, W. Fu, H. Liu, and X. Su, "Indoor mobile localization based on wi-fi fingerprint's important access point," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 4, 2015, Art. no. 429104.

[8] J. Luo and L. Fu, "A smartphone indoor localization algorithm based on WLAN location fingerprinting with feature extraction and clustering," *Sensors*, vol. 17, no. 6, p. 1339, 2017.

[9] B. Huang, Z. Xu, B. Jia, and G. Mao, "An online radio map update scheme for wifi fingerprint-based localization," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6909–6918, Aug. 2019.

[10] L. Li, X. Guo, N. Ansari, and H. Li, "A hybrid fingerprint quality evaluation model for wifi localization," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9829–9840, Dec. 2019.

[11] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proc. IEEE INFOCOM*, 2004, pp. 1012–1022.

[12] E. Elnahrawy, X. Li, and R. P. Martin, "The limits of localization using signal strength: A comparative study," in *Proc. 1st Annu. IEEE Commun. Soc. Conf. Sensor Ad Hoc Commun. Netw.*, 2004, pp. 406–414.

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

[14] B. Huang, M. Liu, Z. Xu, and B. Jia, "On the performance analysis of wifi based localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4369–4373.

[15] M. Zhou, Y. Wang, Y. Liu, and Z. Tian, "An information-theoretic view of wlan localization error bound in GPS-denied environment," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4089–4093, Apr. 2019.

[16] L. Kanaris, A. Kokkinis, G. Fortino, A. Liotta, and S. Stavrou, "Sample size determination algorithm for fingerprint-based indoor localization systems," *Comput. Netw.*, vol. 101, pp. 169–177, 2016.

[17] C. Li, Q. Xu, Z. Gong, and R. Zheng, "Turf: Fast data collection for fingerprint-based indoor localization," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2017, pp. 1–8.

[18] B. Jia, B. Huang, H. Gao, W. Li, and L. Hao, "Selecting critical wifi aps for indoor localization based on a theoretical error analysis," *IEEE Access*, vol. 7, pp. 36312–36321, 2019.

[19] S. H. Jung, B. C. Moon, and D. Han, "Performance evaluation of radio map construction methods for wi-fi positioning systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 880–889, Apr. 2017.

[20] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, and M. Youssef, "Wideep: Wifi-based accurate and robust indoor localization system using deep learning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2019, pp. 1–10.

[21] Z. Li, K. Xu, H. Wang, Y. Zhao, X. Wang, and M. Shen, "Machine-learning-based positioning: A survey and future directions," *IEEE Netw.*, vol. 33, no. 3, pp. 96–101, May/Jun. 2019.

[22] K. Kaemarungsi and P. Krishnamurthy, "Properties of indoor received signal strength for WLAN location fingerprinting," in *Proc. 1st Annu. Int. Conf. Mobile Ubiquitous Syst.: Netw. Serv.*, 2004, pp. 14–23.

[23] M. A. Youssef and A. Agrawala, "On the optimality of wlan location determination systems," in *Proc. Commun. Netw. Distrib. Syst. Model. Simul. Conf.*, 2003, pp. 205–218.

[24] N. Swangmuang and P. Krishnamurthy, "Location fingerprint analyses toward efficient indoor positioning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2008, pp. 100–109.

[25] J. Yang and Y. Chen, "A theoretical analysis of wireless localization using rf-based fingerprint matching," in *Proc. IEEE Int. Symp. Parallel Distrib. Process.*, 2008, pp. 1–6.

[26] Y. Jin, W. S. Soh, and W. C. Wong, "Error analysis for fingerprint-based localization," *IEEE Commun. Lett.*, vol. 14, no. 5, pp. 393–395, May 2010.

[27] B. Huang, G. Qi, X. Yang, L. Zhao, and H. Zou, "Exploiting cyclic features of walking for pedestrian dead reckoning with unconstrained smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 374–385.

[28] W. Liu, Y. Chen, Y. Xiong, L. Sun, and H. Zhu, "Optimization of sampling cell size for fingerprint positioning," *Int. J. Distrib. Sensor Netw.*, vol. 10, no. 9, 2014, Art. no. 273801.

[29] Q. Chen, B. Wang, X. Deng, Y. Mo, and L. T. Yang, "Placement of access points for indoor wireless coverage and fingerprint-based localization," in *Proc. IEEE Int. Conf. High Perform. Comput. Commun.*, 2013, pp. 2253–2257.

[30] M. Youssef and A. Agrawala, "Handling samples correlation in the horus system," in *Proc. IEEE INFOCOM*, 2004, pp. 1023–1031.

Baoqi Huang (Member, IEEE) received the B.E. degree in computer science from Inner Mongolia University (IMU), Hohhot, China, in 2002, the M.S. degree in computer science from Peking University, Beijing, China, in 2005, and the Ph.D. degree in information engineering from The Australian National University, Canberra, ACT, Australia, in 2012. He is currently a Professor with the College of Computer Science, IMU. His research interests include indoor localization and navigation, wireless sensor networks, and mobile computing. He was the recipient of the Chinese Government Award for Outstanding Chinese Students Abroad in 2011.

Runze Yang received the B.E. degree in 2017 and the M.S. degree in 2019 from Inner Mongolia University, Hohhot, China, where he is currently working toward the Ph.D. degree in computer science. His research interests include mobile computing, and indoor localization and navigation.

Bing Jia (Member, IEEE) received the Ph.D. degree from Jilin Univesity, Changchun, China, in 2013. She is currently an Associate Professor with the College of Computer Science, Inner Mongolia University, Hohhot, China. Her current research interests include indoor localization, crowdsourcing, wireless sensor networks, and mobile computing.

Wuyungerile Li received the Ph.D. degree from Shizuoka Univesity, Hamamatsu, Japan, in 2013. She is currently an Associate Professor with the School of Computer Science, Inner Mongolia University, Hohhot, China. Her research interests include wireless sensor networks and mobile computing. She was the recipient of some research grants from the National Science Foundation of China and Inner Mongolia as a Principal Investigator.

Guoqiang Mao (Fellow, IEEE) is currently a Distinguished Professor and the Dean with the Research Institute of Smart Transportation, Xidian University, Xi'an, China. Before that, he was with the University of Technology Sydney, Ultimo, NSW, Australia and with The University of Sydney, Camperdown, NSW, Australia. He has authored or coauthored more than 200 papers in international conferences and journals, which have been cited more than 9000 times. His research interest includes intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques, and network modeling and performance analysis. Since 2018, he has been the Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. From 2014 to 2019, he was the Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Since 2010, he has been the Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was the Co-chair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He was the Chair, Co-Chair and a TPC Member in a number of international conferences. He is a Fellow of IET. He was the recipient of the Top Editor Award for outstanding contributions to the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, in 2011, 2014, and 2015.