Performance Analysis of RSS Fingerprinting Based Indoor Localization

Xiaohua Tian, *Member, IEEE*, Ruofei Shen, Duowen Liu, Yutian Wen, and Xinbing Wang, *Senior Member, IEEE*

Abstract—Indoor localization has been an active research field for decades, where received signal strength (RSS) fingerprinting based methodology is widely adopted and induces many important localization techniques, such as the recently proposed one building fingerprints database with crowdsourcing. While efforts have been dedicated to improve accuracy and efficiency of localization, performance of the RSS fingerprinting based methodology itself is still unknown in a theoretical perspective. In this paper, we present a general probabilistic model to shed light on a fundamental issue: how good the RSS fingerprinting based indoor localization can achieve? Concretely, we present the probability that a user can be localized in a region with certain size. We reveal the interaction among accuracy, reliability, and the number of measurements in the localization process. Moreover, we present the optimal fingerprints reporting strategy that can achieve the best localization accuracy with given reliability and the number of measurements, which provides a design guideline for the RSS fingerprinting based indoor localization system. Further, we analyze the influence of imperfect database information on the reliability of localization, and find that the impact of imperfect information is still under control with reasonable number of samplings when building the database.

Index Terms—Fingerpringting, localization, performance analysis

1 Introduction

Indoor localization has long been an active research field, which enables a vast range of mobile computing applications [1]. Various wireless techniques have been exploited to achieve accurate and efficient indoor localization, where the received signal strength (RSS) fingerprinting based methodology has been a seminal idea induces many indoor localization systems with different flavours [2]. Most of the RSS fingerprinting based localization systems are implemented in IEEE 802.11 wireless local area network (WLAN) environment, where the RSS measured for frames sent from different access points (APs) is utilized to infer the user's location. Specifically, the system first collects the RSS information from APs in the area of interest, where each piece of information is termed as a *fingerprint* and many such fingerprints result in a fingerprints database. During the localization phase, a user submits measured fingerprints to the system, which are compared with the fingerprints database so that the current location of the user can be estimated.

The fingerprints database can be built in many ways. The element in the database could be deterministic, which is just the RSS reading obtained from the wireless card's routine operation of RSS measurement [3]. The element could also be probabilistic, which is the RSS distribution that can be used for location determination in a probabilistic manner. As the RSS itself is a coarse characterization of radio propagation, which is influenced by many environmental factors, recent research turns to the finer-grained wireless feature, i.e., channel state information (CSI) [2], for a higher localization accuracy. Moreover, no matter if the RSS or the CSI is used for fingerprinting, building and updating the fingerprints database is expensive and laborious for any single entity, which spurs the recent active research on location determination with fingerprints collected with the crowdsourcing paradigm [4], [19].

While efforts have been dedicated to the RSS fingerprinting based indoor localization in order to improve the accuracy and efficiency, performance of the RSS fingerprinting based methodology itself is still unknown in a theoretical perspective. Results of empirical studies are highly dependent on experimental environment and implementation [5], [6], [23], [24], [25]. Theoretical analysis borrowed from ranging based cooperative localization in wireless sensor networks is based on ideal radio propagation model and unsuitable for fingerprinting based localization [25], [29], [30], [31]. The current lack of a theoretical insight into the RSS fingerprinting based methodology could incur the blindness for system designers: Can we further improve the performance of the localization system with better implementations or this has been the best we can achieve with the methodology?

In this paper, we present a general probabilistic model to shed light on the fundamental issue: how good the RSS

- *X. Tian is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and the National Mobile Communications Research Laboratory, Southeast University, Jiangsu 210018, China. E-mail: xtian@sjtu.edu.cn.*
- *R. Shen, D. Liu, Y. Wen, and X. Wang are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.*
  E-mail: {shenruofei, ldw123, nogerw, xwang8}@sjtu.edu.cn.

Manuscript received 31 Aug. 2015; revised 20 Nov. 2016; accepted 21 Dec. 2016. Date of publication 26 Dec. 2016; date of current version 29 Aug. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2016.2645221

fingerprinting based indoor localization methodology can achieve? Concretely, we first generalize the assumption of the widely used Log-Normal Path Loss (LNPL) model [2] to provide a more reasonable portrait of the RSS particularly in the indoor environment. We then construct a multi-dimensional probability space based on measure theory, in order to model all possible submitted RSS fingerprints in the location determination phase. Given the expected accuracy, the localization reliability calculation is transformed into the problem of integration over an event in the multi-dimensional sample space of the probability space.

Based on the problem formulation, we present domains of integration in the sample space for the location estimation in one-dimensional and two-dimensional physical indoor space, which are used to model localization process in corridors and ordinary rooms, respectively. We then derive reliability of location estimation in the two cases for any given accuracy requirement. Some interesting findings about the shape of the integration domains are presented, where skilful mathematical techniques are demonstrated.

Moreover, we provide an insight into the RSS fingerprinting based location determination, where we present the condition that there must be a function from a particular subspace of the entire sample space to the physical space. With such an insight, we present the optimal fingerprints reporting strategy that can achieve the best accuracy with given reliability requirement and the number of measurements, which provides a design guideline to the client side of the indoor localization system.

Further, we analyze the influence of imperfect information on performance of localization. The practical fingerprints database constructed in the training stage is unable to provide perfect distribution of fingerprints. We present the probability error of location estimation incurred by the imperfect information, with the relationship between the number of sampling in the training stage and probability error demonstrated.

2 Related Work

2.1 Probabilistic Models Used for Indoor Localization

The early indoor localization system in the context of WLAN is to infer the device's location using the technique of nearest neighbour(s) in signal space (NNSS) [3], where the idea is to compute the Euclidean distance between the measured RSSes and the recorded RSSes from APs strategically deployed at a set of locations. The system returns the location that minimizes the distance. One drawback of the nearest neighbour approach is that it does not fully utilize the opportunity of joint location determination from different APs [7], which leaves room for accuracy improvement.

In order to provide a model for fusing fingerprints from multiple APs, the probabilistic model has been used to estimate the user's location. The Nibble system utilizes Bayesian networks to infer the location of a mobile device [7], where the prior distribution probabilities about a location are obtained by performing sampling for the location over several days in the training phase. With the prior distribution and the Bayesian network, a posterior probability distribution over an estimated location given a set of fingerprints can be derived.

Besides the data fusion, the probabilistic model is also used to deal with the noisy features of the wireless channel, which incurs significant deviations of the sampled RSS fingerprints from those stored in the database thus impacts the accuracy. Youssef et al. propose a joint clustering technique [8], [9], which leverages the Bayes estimation theory addressing noisy wireless channel and reducing computational cost of searching through the fingerprints database. Battiti et al. propose a similar model for localization error caused by the variability of RSS measurements, which is utilized by the local search heuristic technique for improving the localization accuracy [16], [17]. A comparative study on the performance of the indoor localization is presented in [5], where many probabilistic techniques are briefly surveyed and evaluated with experiments.

While probabilistic models are used in the indoor localization system in an ad hoc manner, most of them focus on inferring the best location estimation in the tactical level. Kaemarungsi et al. develop a preliminary probabilistic model for a localization system based on the NNSS approach [6]. With simplified assumptions on the wireless channel feature in the indoor environment, essential properties of the RSS fingerprinting based methodology still remain unknown. The probabilistic model to be presented in this paper is used to analyze the fundamental limits of the general indoor localization technique based on the RSS fingerprinting based methodology. We have a very general assumption of the wireless channel and no assumption on the pre-deployment efforts. Many interesting theoretical findings are to be presented, which have not been shown to the best of our knowledge.

2.2 Crowdsourcing Based Indoor Localization

The indoor localization schemes above requires pre-deployment efforts: there must be some fixed APs whose locations are known for calibration. With more and more APs deployed by different operators, the indoor localization system designer faces a dilemma: information sources are not fully utilized if just using a limited number of pre-deployed APs; however, collecting the training data from all possible APs could be laborious and expensive for any single entity. Further, how fingerprints from APs in unknown locations can be utilized to achieve the most accurate location determination is a challenge.

Chintalapudi et al. propose the EZ localization scheme with limited pre-deployment efforts [18]. Mobile devices record and report RSS fingerprints perceived with respect to different APs at possible unknown locations in the training phase. The fingerprint is represented as the mean and standard deviation of the RSS seen from those APs. EZ only needs the mobile device to occasionally obtain an absolute location at the edge of the indoor environment through GPS, and users can move around at will in the indoor space in normal course.

The almost pre-deployment free service model could spare explicit efforts needed from indoor localization service providers for training data. Fingerprints collection can be performed with crowdsourcing, where any ordinary smart phone user without professional training can collect the fingerprints in an area when passing around. Rai et al. present a calibration zero-effort system Zee [19], which leverages the embedded sensors of mobile devices to track the device itself while simultaneously performing Wi-Fi scans as the carrier of the device traverses an indoor environment.

Wu et al. develop a crowdsourcing based indoor localization system LiFS to avoid the traditional site survey process [20], [21]. The basic idea is to first deploy some landmarks in the physical space, then leverage information derived from smartphone embedded sensors and user motions to construct a high-dimensional sample space with Multidimensional Scaling (MDS) algorithm, which is used to visualize similarities or dissimilarities in data [20]. Physical space can also be characterized using the high-dimensional space induced by MDS. The userās location can be estimated by comparing the high-dimensional sample space and physical space.

Shen et al. present a crowdsourcing based system *Walkie-Markie* [26] to generate indoor pathway maps from the user contributed data. The central idea of the system is to exploit Wi-Fi-Marks defined by Wi-Fi RSS features in the indoor space, so that crowdsourced data by dead reckoning [28] can be fused. Luo et al. propose a self-calibrating participatory indoor localization system [27], which requires no prior knowledge about the building and user intervention including the floor planning.

EZ, Zee and LiFS emphasize on the implementation of localization systems, and the work of Shen et al. [26] and Luo et al. [27] present interesting solutions to estimate the floor planning; however, fundamental issues about the crowdsourcing based indoor location determination are still unclear. With those crowdsourced RSS fingerprints, must there be a mapping from any combination of fingerprints to a location? Chintalapudi et al. mentioned that RSSes from different APs are unequally effective [18], but which APs are more valuable for location determination if the locations of APs are unknown? Our work in this paper will shed light on these fundamental issues.

3 System Model

Consider an indoor space denoted by S, where the long and narrow space such as a corridor can be modelled as an one-dimensional Cartesian space with $S \subset \mathbb{R}$, and the ordinary space such as a room can be modelled as a two-dimensional Cartesian space with $S \subset \mathbb{R}^2$. We use $\vec{r}$ to denote a location in S, where $\vec{r} = x_1$ and $\vec{r} = (x_1, x_2)$ in the one-dimensional and two-dimensional Cartesian coordinate system, respectively. For both fingerprints collection and location determination, the mobile device reports the RSS readings obtained by measuring signals sent from each AP. The measured result is a random variable with respect to a specific location denoted by $\mathcal{P}(\vec{r})$

$$\mathcal{P}(\vec{r}) = \mu(\vec{r}) + \sigma\mathcal{Y}, \tag{1}$$

where $\mu(\vec{r})$ represents how the mean of RSS readings varies with respect to locations. $\mathcal{Y}$ is the normalized Gaussian random variable with $\mathcal{Y} \sim \mathcal{N}(0, 1)$ and σ is a constant representing the standard variance of the received signal.

Equation (1) is a generalized model derived from the LNPL model [2], [18]. If we let $\mu(\vec{r}) = P_T - PL_0 - 10\gamma log_{10}\frac{d}{d_0}$, where P_T is the transmitted power, PL_0 is the path loss at the reference distance d_0, d is the distance between the location of the transmitter to $\vec{r}$ and γ is the path loss exponent, Equation (1) degenerates to the LNPL model, where $\sigma\mathcal{Y}$ factually represents the shadowing effect. Extensive practical measurements and studies in the literature have revealed that the value of σ can be regarded as a constant in a certain region if the location of the transmitter is given, which has been widely adopted in the industrial standardization documents for radio propagation modeling [32], [33].

The LNPL model above only considers the path loss and shadowing but ignores the small scale fading incurred by multi-path effect. The multi-path effect will result in changes in the received power, but it is extremely difficult to accurately predict how the multi-path effect will incur the change. To this end, we use a function $\mu(\vec{r})$ to denote the average aggregated effect of free space path loss and multi-path effect at location $\vec{r}$, thus we generalize the original LNPL model and obtain the radio propagation model as shown in Equation (1). Such modeling approach is also adopted by a number of work focusing on characterizing wireless communication channels such as in [13], [14], [15], and the Gaussian assumption is validated by a number of work on indoor localization [6], [9], [10], [11], [12], [18].

As the real wireless environment is very unpredictable, we are unable to know exactly what the mean value of the measured RSSes $\mu(\vec{r})$ is like. However, it is observed from many previous experiments [5], [6], [9], [18] that the mean of measured RSSes changes in a non-dramatic manner with the small change of locations, which makes it reasonable to assume that $\mu(\vec{r})$ is continuous. Thus we can have the following approximation:

$$\mu(\vec{r'}) \approx \mu(\vec{r}) + \nabla\mu(\vec{r})(\vec{r'} - \vec{r}), \tag{2}$$

where $\vec{r'}$ is an estimated location of the user given the actual location of the user $\vec{r}$. Note that all the analysis based on Eq. (1) can be applied to the scenario using the LNPL model, which itself has been widely adopted in many indoor localization systems [6], [9], [10], [18]. The experimental results to be shown in Section 8 also support the modeling above.

In the training phase, the mobile device is randomly assigned a point in the indoor space and the device randomly chooses an AP and measures the RSS fingerprint once. As the result of each measurement is a random variable as shown in Eq. (1), all possible outcomes for the measurement form a sample space Ω. We define a σ-algebra $\mathcal{F}$ that is the collection of all events, where each event is a set containing zero or more outcomes. Eq. (1) gives an assignment of probabilities to events $\mathcal{P} : \Omega \to [0, 1]$, thus we can construct a probability space (Ω, $\mathcal{F}$, $\mathcal{P}$). Suppose that the mobile device performs the measurement n times at a given assigned point, then the Cartesian product of probability spaces induced by RSS measurements forms an n-dimensional probability space $(\Omega^n, \mathcal{F}^n, \mathcal{P})$, where we abuse $\mathcal{P} : \Omega^n \to [0, 1]$ for the convenience of demonstration.

The location determination phase can be considered as a mapping $\mathcal{M} : \Omega^n \to S$, $\vec{r'} = \mathcal{M}(\vec{o})$, where $\vec{o}$ is an outcome of Ω^n. It means that the localization system outputs an estimated location $\vec{r'}$ for a series of measurements of RSSes from randomly chosen APs. Since RSS measurement results are independently and identically distributed, the induced sample spaces of RSS measurements are orthogonal to each other, and Ω^n is homeomorphic to n-dimensional Cartesian

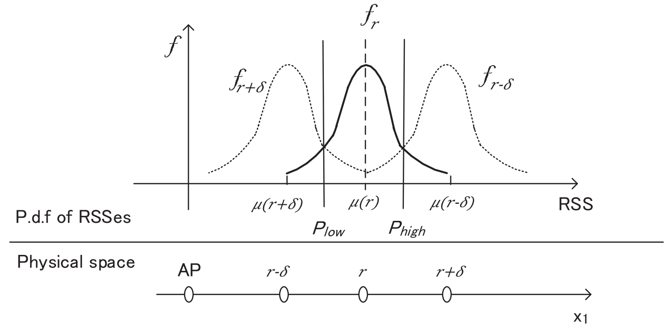


Fig. 1. Integration domain for one-dimensional localization with single measurement for single AP.

space. If applying the coordinate system of n-dimensional Cartesian space to Ω^n, we can obtain a presentation of $\vec{o}$ denoted as $\vec{P} = [P_1, P_2, \dots, P_n]^T$, in which P_i is the reading of each measurement result. Consequently, the original measurement $\mathcal{P} : \Omega^n \to [0,1]$ becomes $f : \mathbb{R}^n \to [0,1]$.

We use Q to denote the area that is centered at the user's actual location with radius δ in the physical space. $E(\delta)$ is used to denote the event in the sample space, which makes the localization system estimates the user's location to be in Q. We will use E to represent $E(\delta)$ in the following discussion to avoid tedious mathematical presentation. The size of δ determines the *accuracy* of the localization system. We use R to denote the probability that the user's estimated location is within the area Q, which is defined as the *reliability* of the localization system.

With the definition of accuracy and reliability, we can see that the probability of the event that the system correctly estimates the user's location is

$$R(E) = \int_E f(\vec{P})d\vec{P} = \int_Q g(\vec{r'}, \vec{r})d\vec{r'}, \tag{3}$$

where $f(\vec{P})$ is the possible measurement of the sample space in the n-dimensional Cartesian coordinate system, and $g(\vec{r'}, \vec{r})$ is the probability distribution function that the user is localized at $\vec{r'}$ given that the real location of the user is $\vec{r}$ in the physical space. Equation (3) indicates that the reliability can be interpreted as either the probability that the measurement falls into the event region E or the user is localized in an area that is centered at $\vec{r}$ and with radius δ. Based on such a model, we are to calculate the one-dimensional localization reliability for the case of one-time measurement for a single AP, and then extend the result to multiple-time measurements for multiple APs in the following section.

4 Localization In One-Dimensional Space

4.1 One-Time Measurement for Single AP

We set the origin of the spacial coordinate system at the location of the sole AP in the one-dimensional physical space; the corresponding probability density function (PDF) for each location in the sample space could be represented as shown in Fig. 1.

The location in one-dimensional physical space is a scalar, and the δ neighborhood of the user's actual location r is the line segment from $r - \delta$ to $r + \delta$. Note that the farther the location is from the AP, the smaller mean value could be observed at the location; therefore, $\mu(r + \delta)$ is less than $\mu(r - \delta)$. The probability the user is localized in the δ neighborhood of r (denoted by Q) is equivalent to that the reported RSSes fall within the range between P_{high} and P_{low} according to the principle of maximum likelihood estimation (MLE), which is the event E in this case. Due to the symmetry of the PDF according to Eq. (1), it is straightforward that $f_{r-\delta}(P_{high}) = f_r(P_{high})$ and $f_{r+\delta}(P_{low}) = f_r(P_{low})$. We have

$$\begin{cases} P_{high} = \frac{\mu(r-\delta)+\mu(r)}{2}, \\ P_{low} = \frac{\mu(r+\delta)+\mu(r)}{2}. \end{cases} \tag{4}$$

thus the reliability

$$R(\delta, r, \sigma) = \int_{P_{low}}^{P_{high}} f_r(P)dP = erf\left(\left|\frac{-\mu'(\vec{r})\delta}{2\sqrt{2}\sigma}\right|\right), \tag{5}$$

where $erf(\cdot)$ is the error function defined as: $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t$.

4.2 Multiple Measurements for Multiple APs

According to probability theory, the average of n i.i.d Gaussian variables is equivalent to a Gaussian variable with a standard deviation $\frac{\sigma}{\sqrt{n}}$. Two measurements to a single AP can be regarded as measurements for two identical APs located at the same place, which is to be confirmed by our result shown in Eq. (12). If several measurements are performed on a single AP, the RSS fingerprint is set to be a new random variable with the standard deviation $\frac{\sigma}{\sqrt{n}}$.

Similar to the situation in Section 4.1, the probability the user is localized in Q is equivalent to the probability the user's measurement of the RSS P_i falls into the event E

$$E = \left\{\vec{o} \middle| \prod_{i=1}^n f_r(P_i) \geq \prod_{i=1}^n f_{r+\delta}(P_i), \prod_{i=1}^n f_r(P_i) \geq \prod_{i=1}^n f_{r-\delta}(P_i)\right\}.$$

According to the radio propagation model Eq. (1), the outcomes in the event E satisfy the following inequality:

$$\prod_{i=1}^n \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(P_i - \mu_i(r))^2}{2\sigma^2}} \geq \prod_{i=1}^n \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(P_i - \mu_i(r\pm\delta))^2}{2\sigma^2}}.$$

After simplification, it is equivalent to

$$\sum_{i=1}^n \frac{\mu_i(r\pm\delta) - \mu_i(r)}{\sigma_i}\left(Y_i - \frac{\mu(r\pm\delta) - \mu(r)}{2\sigma_i}\right) \leq 0, \tag{6}$$

where $\mu_i(r)$ is the average RSS of AP i at r.

We normalize RSS readings P_i with respect to $Y_i = \frac{P_i - \mu_i}{\sigma_i}$, where σ_i might differ among different APs. For a given location of the receiver, the received signal propagated from different APs can be through different paths. A simple example is that one AP's signal is propagated through the line-of-sight (LOS) channel and another AP's signal could be propagated through the non-line-of-sight (NLOS) channel to the receiver; therefore, the observed values of σ for different APs can be different [32], [33]. The distance between two APs are usually hundreds of meters in practice to save the infrastructure investment, since one AP could cover an area of radius about 100 meters; moreover, densely deploying APs could incur interference. Consequently, it is reasonably to assume that the distance between two APs for

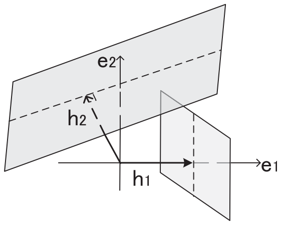


Fig. 2. Integration domain for one-dimensional localization with multiple measurements for multiple APs.

localization is larger than the dimension of δ neighborhood of a receiver's real location. This is why we assume the value of σ is constant for a given AP as shown in Equation (1) but the values of σ with respect to different APs are different. The experiment results to be shown in Section 8 also support such modeling.

We use $\vec{\mu}(r)$ to denote the mean RSS outcome at r. Now that it refers to the outcome itself, this notation does not depend on any coordinate system. Given a specific average RSS $\vec{\mu}(r)$, the set of $\{Y_i\}$ forms a coordinate basis for the sample space, where the origin is $\vec{\mu}(r)$, and each dimension is suppressed by a factor of σ_i. We use this coordinate system to characterize the event E and its probability in the following several sections due to its simplicity for the small scale analysis.

Apparently, the two constraints shown in Eq. (6) are two non-parallel hyper-planes in the sample space. Vectors $\vec{h}_1 = [\frac{1}{2\sigma_1}(-\mu_1(r) + \mu_1(r-\delta)), \frac{1}{2\sigma_2}(-\mu_2(r) + \mu_2(r-\delta)), \ldots, \frac{1}{2\sigma_n}(-\mu_n(r) + \mu_n(r-\delta))]^T, \vec{h}_2 = [\frac{1}{2\sigma_1}(-\mu_1(r) + \mu_1(r+\delta)), \frac{1}{2\sigma_2}(-\mu_2(r) + \mu_2(r+\delta)), \ldots, \frac{1}{2\sigma_n}(-\mu_n(r) + \mu_n(r+\delta))]^T$ together span a plane $\mathcal{W}$. As restrictions to the event E, Eq. (6) can then be rewritten in the vector form

$$\begin{cases} 2\vec{h}_1(\vec{o} - \vec{h}_1) \leq 0, \\ 2\vec{h}_2(\vec{o} - \vec{h}_2) \leq 0. \end{cases} \tag{7}$$

It is important to note that $\vec{h}_1$ and $\vec{h}_2$ can denote both the normal vectors to each hyperplane and the two points on each hyperplane that are closest to the origin. This fact will be helpful to deal with the two-dimensional issue. Now that the PDF and the constraint conditions are all normalized, we can rotate the coordinate system $\{Y_i\}$ to another orthonormal basis $\{\vec{e_i}\}, i = 1, 2, \ldots, n$, where $\vec{e_1}$ is parallel to $\vec{h_1}$ and $\vec{e_2} \in \mathcal{W}$. Consequently, $\mathcal{W}$ is spanned by only two coordinate axes. The rest coordinate axes are therefore all orthogonal to the plane. There exists an orthonormal basis for subspace $\bar{\mathcal{W}}$, i.e., $\{\vec{e_i}\}, i = 3, \ldots, n$. Any outcome $\vec{o}$ in the sample space can be decomposed into $\vec{o} = \sum_i c_i \vec{e_i}$, where coefficients c_i is determined and unique for any given vector $\vec{o}$ and orthonormal basis $\{\vec{e_i}\}$. Eq. (6) can then again be rewritten in the component form of the $\{\vec{e_i}\}$ basis

$$\begin{cases} 2|\vec{h}_1\vec{e_1}|(c_1 - |\vec{h}_1\vec{e_1}|) \leq 0, \\ 2|\vec{h}_2\vec{e_1}|(c_1 - |\vec{h}_2\vec{e_1}|) + 2|\vec{h}_2\vec{e_2}|(c_2 - |\vec{h}_2\vec{e_2}|) \leq 0. \end{cases} \tag{8}$$

Thus the probability that the system correctly estimates the user's location is

$$R(E) = \int_E f_r(\vec{P}) de^n \tag{9}$$

$$= \int_{-\infty}^{c_1 \leq |\vec{h}_1|} de_1 \int_{-\infty}^{\frac{|\vec{h}_2\vec{e_1}|^2 + |\vec{h}_2\vec{e_2}|^2 - |\vec{h}_2\vec{e_1}|c_1}{|\vec{h}_2\vec{e_2}|}} \frac{1}{2\pi} e^{-\frac{e_1^2+e_2^2}{2}} de_2. \tag{10}$$

Note that $f_r(\vec{P})$ is an n-variable Gaussian PDF. As of now, we successfully reduce the multiple integral to a much simpler two dimensional one. Multivariate Gaussian integral Eq. (11) is integrated on the area indicated in Fig. 2.

By Eq. (2), the dimension of E can be further reduced, for that $\vec{h}_1$ and $\vec{h}_2$ will now be parallel to each other, though in different directions

$$\begin{cases} \vec{h}_1 = [-\frac{1}{2\sigma_1}\mu_1'(r)\delta, \ldots, -\frac{1}{2\sigma_n}\mu_n'(r)\delta]^T, \\ \vec{h}_2 = [\frac{1}{2\sigma_1}\mu_1'(r)\delta, \ldots, \frac{1}{2\sigma_n}\mu_n'(r)\delta]^T. \end{cases}$$

Thus the two constraint conditions shown in Eq. (6) are parallel to each other. $R(E)$ can then be simplified as

$$R(E) = \int_{|\vec{h}_2|}^{|\vec{h}_1|} de_1 \int_{-\infty}^{-\infty} \frac{1}{(\sqrt{2\pi})^2} e^{-\frac{e_1^2+e_2^2}{2}} de_2 \tag{11}$$

$$\approx erf\left(\frac{\sqrt{\sum_{i=1}^{n}(\frac{\mu_i'(r)\delta}{2\sigma_i})^2}}{\sqrt{2}}\right). \tag{12}$$

4.3 Discussions

We can see that Eq. (12) is equivalent to Eq. (5) when $n = 1$ meaning that there is a single AP in the room, which corroborates our analysis. Moreover, the analysis above reveals some insight into the design of the indoor localization. First, the more data are reported to the system, the more reliable the location estimation is, since $\sqrt{\sum_i(\frac{\mu_i'(r)\delta}{2\sigma_i})^2} > \sqrt{(\frac{\mu_i'(r)\delta}{2\sigma_i})^2}$. Second, if $\frac{\mu_i'(r)}{\sigma_i} \geq \frac{\mu_j'(r)}{\sigma_j}, \forall i, j \leq n$, the reliability of the result can beimproved if the user reports AP_i's RSS rather than the other. This means that reporting a measurement of a cleaner channel (smaller σ) is more effective, and the sharper the signal varies around the user's location (greater $\mu'(r)$), the easier it is for the system to pinpoint the user's location.

5 Localization In Two-Dimensional Space

Finding the event E in the two-dimensional localization is more challenging. To simplify the modeling, we first present a mathematical expression of Q in the physical space, based on which we try to find the shape of the event E in the sample space. We are to prove that E is a hyper-cylinder, and then prove that the intersection between the hyper-cylinder and the cross-section plane is in the shape of an ellipse, which makes it possible to obtain the reliability through integration.

5.1 Multiple Measurements for Multiple APs

Fig. 3 illustrates how to represent a location in the two-dimensional physical space, where the location of the user is $\vec{r}$ and the location of any point on the boundary of the area Q is $\vec{r'}$. We use $\vec{\delta} = \vec{r'} - \vec{r}$ to denote a two-dimensional vector with the direction from the user's actual location to any point on the boundary of Q. We use θ to denote the angle between $\vec{\delta}$ and the horizontal axis, and use ϕ_i to

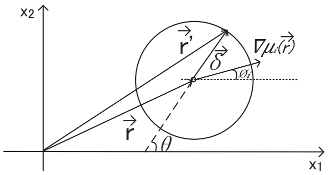


Fig. 3. Two-dimensional localization with multiple measurements for multiple APs.

denote the angle between $\nabla\mu_i(\vec{r})$ and the horizontal axis. By Eq. (2), we have

$$\mu_i(\vec{r'}) - \mu_i(\vec{r}) = \nabla\mu_i(\vec{r})\vec{\delta} = \delta|\nabla\mu_i(\vec{r})|cos(\theta - \phi_i), \qquad (13)$$

where $\delta = |\vec{\delta}|$.

We want to find the event that the user is localized within the area Q. According to MLE, this is equivalent to find E, where the probability density of the user's appearing on the boundary of Q is no greater than that of the user's appearing at $\vec{r}$

$$E = \left\{ \vec{\sigma}| \prod_{i=1}^{n} f_{\vec{r}}(P_i) \geq \prod_{i=1}^{n} f_{\vec{r}+\vec{\delta}}(P_i) \right\}. \qquad (14)$$

All outcomes in E follow the inequality

$$\sum_{i=1}^{n} \frac{\mu_i(\vec{r}+\vec{\delta}) - \mu_i(\vec{r})}{\sigma_i} \left(Y_i - \frac{\mu_i(\vec{r}+\vec{\delta}) - \mu_i(\vec{r})}{2\sigma_i} \right) \leq 0. \quad (15)$$

Substituting Eq. (13) into Eq. (15), we have the specific description of E

$$\sum_{i=1}^{n} \frac{\delta|\nabla\mu_i(\vec{r})|}{\sigma_i} cos(\theta - \phi_i) \left(Y_i - \frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i} cos(\theta - \phi_i) \right) \leq 0. \qquad (16)$$

Constraint condition Eq. (16) should hold true for any θ, thus there will be an infinite set of hyper-planes surrounding event E in the sample space, as shown in Fig. 4. We define the n-dimensional normal vector of the hyper-plane to be a function of θ:

$$\vec{h}(\theta) = \left[\frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} cos(\theta - \phi_1), \ldots, \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} cos(\theta - \phi_n) \right]^T$$

Theorem 1. *The orbit of $\{\vec{h}(\theta)\}$ and the origin are coplanar, i.e., on the same two-dimensional plane in the sample space.*

Proof. This is equivalent to prove that there exists a rank $n-2$ complementary subspace $\bar{\mathcal{W}}$ of $\mathcal{W}$, where $\mathcal{W}$ is spanned by $\{\vec{h}(\theta)\}$. Formally, $\forall \vec{g} \in \bar{\mathcal{W}}, \forall\theta$, $\vec{h}(\theta) \cdot \vec{g} \equiv 0$, that is

$$[\tfrac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} \{cos\theta cos\phi_1 + sin\theta sin\phi_1\}, \ldots,$$
$$\tfrac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} \{cos\theta cos\phi_n + sin\theta sin\phi_n\}][g_1, \ldots, g_2]^T \equiv 0.$$

Consequently, we need to prove

$$\begin{cases} cos(\theta)[\frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} cos\phi_1, \ldots, \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} cos\phi_n][g_1, \ldots, g_n]^T \equiv 0, \\ sin(\theta)[\frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} sin\phi_1, \ldots, \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} sin\phi_n][g_1, \ldots, g_n]^T \equiv 0. \end{cases}$$

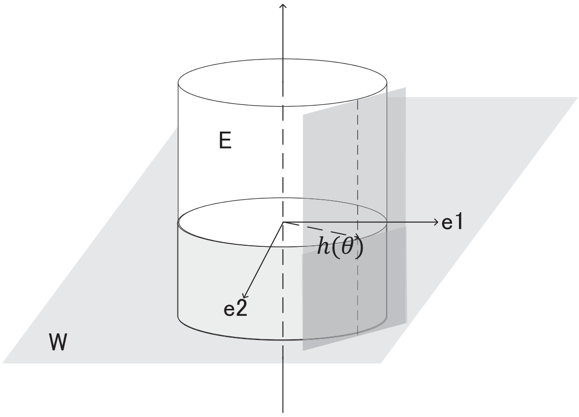


Fig. 4. Integral area of two-dimensional space.

The equations above hold true for all θ, thus

$$\begin{cases} [\frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} cos\phi_1, \ldots, \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} cos\phi_n][g_1, \ldots, g_n]^T \equiv 0, \\ [\frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} sin\phi_1, \ldots, \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} sin\phi_n][g_1, \ldots, g_n]^T \equiv 0. \end{cases} \qquad (17)$$

Adding $n-2$ lines of zero row vectors under the row vector in Eq. (17) makes an $n \times n$ square matrix $\mathcal{H}$

$$\mathcal{H} = \begin{pmatrix} \frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} cos(\phi_1) & \ldots & \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} cos(\phi_n) \\ \frac{\delta|\nabla\mu_1(\vec{r})|}{2\sigma_1} sin(\phi_1) & \ldots & \frac{\delta|\nabla\mu_n(\vec{r})|}{2\sigma_n} cos(\phi_n) \\ 0 & .. & 0 \end{pmatrix}.$$

Then $\bar{\mathcal{W}}$ is the solution space to the linear formula: $\mathcal{H}\vec{g} = 0$, where $rank(\mathcal{H}) \leq 2$, so $rank(\bar{\mathcal{W}}) \geq n-2$; therefore, $rank(W) = n - rank(\bar{\mathcal{W}}) \leq 2$. The straight line connecting $\vec{h}(\theta)$ and $\vec{h}(-\theta)$ will always come across the origin, thus the origin is also in plane $\mathcal{W}$. □

Equation (13) means that $\mathcal{W}$ is a tangent plane of surface M at $\vec{\mu}(\vec{r})$, where $M = \{\vec{\mu}(\vec{r'})|\vec{r'} \in S\}$ is the mean surface of RSS readings. Repeat the technique we used in Section 4.2, we will be again able to reduce the multivariate integral in the whole event to a two variable integral on a subset of plane $\mathcal{W}$. The next is to determine the domain of probability integration. By definition, it is the area inside the envelop of $\{\vec{h}(\theta)\}$.

Theorem 2. *The orbit of $\{\vec{h}(\theta)\}$ is an ellipse.*

Proof. Theorem (1) states that the orbit of $\{\vec{h}(\theta)\}$ is in a plane. To prove Theorem (2) is equivalent to prove that $\forall\theta, \exists\psi$,

$$\vec{h}(\theta) = \vec{U}cos(\psi) + \vec{V}sin(\psi), \qquad (18)$$

where $\vec{U}$ and $\vec{V}$ are constant vectors and $\vec{U}\vec{V} = 0$. If there exists such a constant $\alpha = \theta - \psi$ satisfing Eq. (18), then Theorem (2) is proven

$$\vec{h}(\theta) = \sum_{i=1}^{n} \frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i} \Big\{ cos(\psi + \alpha - \phi_i) \Big\} Y_i \qquad (19)$$

$$= \sum_{i=1}^{n} \frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i} \Big\{ cos(\alpha - \phi_i)cos\psi - sin(\alpha - \phi_i)sin\psi \Big\} Y_i. \qquad (20)$$

If $\vec{U}$ and $\vec{V}$ are assigned as following then Eq. (19) is satisfied

$$\begin{cases}\vec{U} &= \sum_i\left\{\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}cos(\alpha-\phi_i)Y_i\right\},\\ \vec{V} &= \sum_i\left\{-\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}sin(\alpha-\phi_i)Y_i\right\}.\end{cases} \tag{21}$$

Thus $\vec{U}\vec{V}=0$ is equivalent to

$$f(\alpha)=\sum_{i=1}^{n}\left(\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}\right)^2 sin(2\alpha-2\phi_i)=0. \tag{22}$$

Apparently, Formula (22) has four different solutions of α in the interval between 0 and 2π, because $f(\alpha)=-f(\alpha+\pi/2)$ and $f(\alpha)$ is a continuous function. There will be four zero points within each 2π period. The four solutions actually correspond to four different assignments of vector $\vec{U}$ to the semi-major axes and semi-minor axes. However, as we are only interested in the length of the semi-major axis and semi-minor axis, all four kinds of assignments are the same. We will use $\vec{U}$ as the semi-major axis in the following sections

$$\sum_i\left(\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}\right)^2(sin(2\alpha)cos(2\phi_i)-cos(2\alpha)sin(2\phi_i))=0, \tag{23}$$

where

$$tan2\alpha=\frac{\sum_{i=1}^{n}(\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i})^2 sin(2\phi_i)}{\sum_{i=1}^{n}(\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i})^2 cos(2\phi_i)}.$$

Thus we have

$$\begin{cases}|\vec{U}| &= \sqrt{\sum_{i=1}^{n}\{\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}cos(\alpha-\phi_i)\}^2},\\ |\vec{V}| &= \sqrt{\sum_{i=1}^{n}\{\frac{\delta|\nabla\mu_i(\vec{r})|}{2\sigma_i}sin(\alpha-\phi_i)\}^2}.\end{cases} \tag{24}$$

□

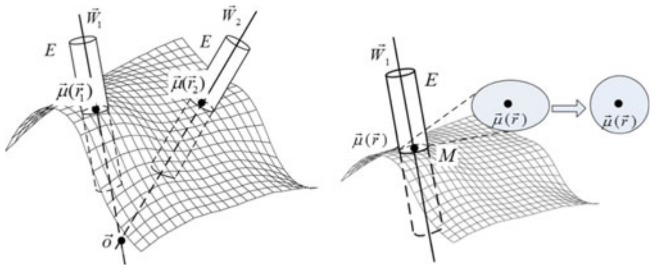
Consequently, the reliability of the location estimation in the two-dimensional space is

$$R(E) = \int_{\frac{e_1^2}{|\vec{U}|^2}+\frac{e_2^2}{|\vec{V}|^2}=1}\frac{1}{2\pi}e^{-\frac{e_1^2+e_2^2}{2}}de_1de_2 \tag{25}$$

$$= |\vec{U}||\vec{V}|\frac{1}{2\pi}\int_0^{2\pi}\frac{1-e^{-\frac{cos^2\psi|\vec{U}|^2+sin^2\psi|\vec{V}|^2}{2}}}{cos^2\psi|\vec{U}|^2+sin^2\psi|\vec{V}|^2}\,d\psi. \tag{26}$$

5.2 Discussions

Most conclusions in the one-dimensional situation still hold true in the two-dimensional case. More data will yield higher reliability; however, there are some distinguishable properties in the two-dimensional case worth of mentioning. First, if the $\nabla\mu_i(\vec{r})$ for all APs are the same, which means that $\phi_i=\phi_j,\forall i,j$, then it is impossible to determine the user's location. This is because $|\vec{V}|=0$ in this case thus $R(E)=0$. It means that there should be at least two APs and the corresponding directions of $\nabla\mu_i(\vec{r})$ are different from each other. Second, if the user observes that $\nabla\mu_i(\vec{r})$ and $\nabla\mu_j(\vec{r})$ for two APs i and j are either in the same direction or in the opposite direction, then it is just like in the one-dimensional case, thus if $\frac{|\nabla\mu_i(r)|}{\sigma_i}\geq\frac{|\nabla\mu_j(r)|}{\sigma_j},\forall i,j\leq n$, reporting the RSS reading from AP i is more effective than reporting that of AP j for location determination.



(a) Fundamentals of location determination

(b) Best RSS reporting strategy

Fig. 5. Sample space of RSSes.

6 Best Strategy for Location Determination

The analysis above shows that the utilities for reporting RSS fingerprints from different APs are different in the location determination process. A natural question to ask is: which fingerprints should the user report to the system so that the most accurate location estimation can be obtained? Before revealing the answer to such a question, we first present the fundamentals of the location determination.

6.1 Fundamentals of Location Determination

The fundamental issue of location determination is that: can every outcome in the sample space be mapped into a location in the physical space. The mean of RSSes $\vec{\mu}(\cdot)$ is a continuous mapping from the physical space to the mean surface of RSSes M. According to Eq. (2), each small area around $\vec{\mu}(\vec{r})$ can be approximated as a plane. Recall that the event E in the two-dimensional case is a hyper-cylinder. According to Theorem 2, the intersection of the hyper-cylinder and M forms an orbit, which is the same two-dimensional plane as $\vec{\mu}(\vec{r})$; therefore, if we shrink δ to zero, then the hyper-cylinder will shrink to an $n-2$ dimensional body $\bar{\mathcal{W}}$ and intersect with M at $\vec{\mu}(\vec{r})$. We use Fig. 5a to illustrate a simple example of 3-D sample space. $\bar{\mathcal{W}}$ is the event that the user is estimated to be most likely appearing at $\vec{r}$, because E is the event that the system estimates the user's location in the area with a radius no more than δ.

Consequently,

$$\vec{r}=\mathscr{M}(\bar{\mathcal{W}}), \tag{27}$$

where $\mathscr{M}:\Omega^{n-2}\rightarrow S$ is a mapping from the set of outcomes $\bar{\mathcal{W}}$ to the user's most likely location $\vec{r}$.

However, it is worth noting that we in fact abuse the notation $\vec{r}$ here, since the location obtained from the mapping $\mathscr{M}$ is not necessarily the actual location of the user. To see this, recall that $\bar{\mathcal{W}}$ is a $(n-2)$-dimensional body perpendicular to the tangent plane of M at $\vec{\mu}(\vec{r})$ in the $\{\frac{P_i}{\sigma_i}\}$ coordinate system, and M is a surface with curvature, thus it may happen that $\bar{\mathcal{W}}_1$ and $\bar{\mathcal{W}}_2$ that induced by two tangent planes of M intersects at an outcome, and this outcome can be mapped into two different points on M. This scenario is illustrated as in Fig. 5a, where the outcome $\vec{o}$ could be mapped into both $\vec{r_1}$ and $\vec{r_2}$. That is, the same set of RSSes can result in different localizations.

If this happens, we also want to use MLE to derive which location the user is more likely to appear. By the definition of MLE, the intersection outcome should be mapped to the point $\vec{r'}$ on M, if the inequality $f_{\vec{r'}}(\vec{P}) > f_{\vec{r''}}(\vec{P})$ is satisfied, where $f_{\vec{r'}}(\vec{P}) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\left(P_i-\mu_i(\vec{r'})\right)^2/2\sigma_i^2}, f_{\vec{r''}}(\vec{P}) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\left(P_i-\mu_i(\vec{r''})\right)^2/2\sigma_i^2}$, and $\vec{r'}$ and $\vec{r''}$ can be any points in the physical space. If $f_{\vec{r'}}(\vec{P}) > f_{\vec{r''}}(\vec{P})$, then $(\frac{P_i-\mu_i(\vec{r'})}{\delta_i})^2 < (\frac{P_i-\mu_i(\vec{r''})}{\delta_i})^2$. This means that in the $\{\frac{P_i}{\sigma_i}\}$ coordinate system, the Euclidean distance from the outcome $\vec{o} = [\frac{P_1}{\sigma_1}, \frac{P_2}{\sigma_2}, \ldots, \frac{P_n}{\sigma_n}]^T$ to the point $[\frac{\mu_1(\vec{r'})}{\sigma_1}, \frac{\mu_2(\vec{r'})}{\sigma_2}, \ldots, \frac{\mu_n(\vec{r'})}{\sigma_n}]^T$ should be less than that to $[\frac{\mu_1(\vec{r''})}{\sigma_1}, \frac{\mu_2(\vec{r''})}{\sigma_2}, \ldots, \frac{\mu_n(\vec{r''})}{\sigma_n}]^T$. It should be noticed that the latter two arrays are the $\{\frac{P_i}{\sigma_i}\}$ representation of the corresponding mean RSS outcomes $\vec{\mu}(\vec{r'})$ and $\vec{\mu}(\vec{r''})$, respectively, which are two arbitrary points on the mean RSS surface M; therefore, we should map the intersection outcome to the point with shortest Euclidean distance to M in the $\{\frac{P_i}{\sigma_i}\}$ coordinate system.

The analysis above indicates that if the user reports to the system the outcomes that are closer to M, it is more likely the user can be localized to the actual location. Consider the mean RSS surface M, no matter how great the curvature of M is, we can always find a very small space in Ω^n which is around $\vec{\mu}(\vec{r})$ on M, so that the part of M in such a small space can be approximated to its tangent plane $\mathcal{W}$ at $\vec{\mu}(\vec{r})$. Each $\bar{\mathcal{W}}$ for the given $\vec{\mu}(\vec{r'})$ is parallel to that for others. If we move $\vec{r}$ around on S, the point $\vec{\mu_i}(\vec{r})$ moves around correspondingly on $\mathcal{W}$. This is equivalent to say that $\bar{\mathcal{W}}$ scans the entire small space in Ω^n.

Every point within the small space is on a unique $\bar{\mathcal{W}}$, and there must be a mapping from Ω^n to S for every point in the sample space around M. It is interesting to find that if the space is very small, the tangent plane approximation is more accurate thus mapping the outcome into the surface is almost the same to find the Euclidean distance. If the outcome is far from M, it may happen that there is no such a $\bar{\mathcal{W}}$ so that the outcome can be mapped to a location.

In conclusion, if M is a plane, there must exist a function from Ω^n to S; if M is with curvature, the nearer the reported outcomes to M, the more likely the system will return a reliable location estimation with accuracy δ.

6.2 Best Strategy

With revealing the fundamentals of location determination, we now derive which APs users should measure so that they can be localized with the highest accuracy with the given reliability. Let $\mathbb{U} = \{AP_i\}, i = 1, \ldots, m$ be the set of all APs that can be sensed by the user's mobile device. A measurement strategy is defined as a sequence of measurements on APs and is denoted by $\mathcal{V}_n = (s^1, \ldots, s^n), s^j \in \mathbb{U}$. Note that the the superscript of s^j is the index of the measurement in the sequence, and it does not necessarily mean that the measurement is performed on AP_j. One AP can be measured more than once in the sequence. The whole set of strategies is denoted as $\mathbb{U}^n$, where the size of the set is m^n. The optimal strategy is denoted by $\mathcal{V}_n^*$, $\mathcal{V}_n^* \in \mathbb{U}^n$.

Recall that the event E is a hyper-cylinder in the sample space and the intersection between the hyper-cylinder and M is an ellipse centered at $\vec{\mu}(\vec{r})$, as shown in Fig. 5b. We now consider another event $\mathcal{E}(c)$, which is also a hyper-cylinder in the sample space; however, we let the intersection between such a hyper-cylinder and M be a circle centered at $\vec{\mu}(\vec{r})$ and with radius c, where $\vec{r}$ is the actual location of the user as shown in Fig. 5b. In another perspective, $\mathcal{E}(c)$ denotes the event that the outcomes for localizing a user at $\vec{r}$ fall in the newly defined hyper-cylinder. Thus the reliability of the location estimation is in fact the probability of the event $\mathcal{E}(c)$, which is similar to the previous analysis

$$R(\mathcal{E}(c)) = \int_0^{2\pi} \int_0^{c} \frac{1}{2\pi} e^{-\frac{\rho^2 cos^2\psi + \rho^2 sin^2\psi}{2}} \rho d\rho d\psi \tag{28}$$

$$= 1 - e^{-c^2/2}. \tag{29}$$

Let us switch our attention to the physical space. We consider the vicinity of $\vec{r}$, where each point on the boundary of the vicinity represents an outcome in Ω^n. The vicinity is denoted as $\mathcal{U}$ and it must satisfy that the outcomes for localizing those location points on the boundary of $\mathcal{U}$ just fall on the circle on M. The point on the circle is denoted as $\mu(\vec{r'})$. Thus

$$\sum_{i=1}^{n} \frac{(\mu_i(\vec{r'}) - \mu_i(\vec{r}))^2}{(2\sigma_i)^2} = c^2. \tag{30}$$

Put $\vec{r}$ and $\vec{r'}$ in the polar coordinate system with the origin at $\vec{r}$, then Eq. (30) can be transformed into $\sum_{i=n}^{n}(\rho(\theta)|\nabla\mu_i(\vec{r})|cos(\theta - \phi_i))^2/(2\sigma_i)^2 = c^2$, thus we have

$$\rho^2(\theta) = \frac{4c^2}{\sum p_i cos^2(\theta - \phi_i)}, \tag{31}$$

where $p_i = (|\nabla\mu_i(\vec{r})|/\sigma_i)^2$. Examining Eq. (31), and let $Q_1 = \sum p_i cos^2\phi_i, Q_2 = \sum p_i sin^2\phi_i, Q_3 = \sum 2p_i cos\phi_i sin\phi_i$. We have $Q_1\rho^2 cos^2\theta + Q_2\rho^2 sin^2\theta + Q_3\rho^2 cos\theta sin\theta = 4c^2$, which means that $\mathcal{U}$ is in fact an ellipse.

Define a complex parameter Z_i characterizing AP_i, where $Z_i = p_i e^{2i\phi_i}$, $\sum Z_i = \sum p_i e^{2i\phi_i}$, and $\sum Z_i^* = \sum p_i e^{-2i\phi_i}$. The area of $\mathcal{U}$ is denoted by u, where $u = 8\pi c^2/\sqrt{4Q_1Q_2 - Q_3^2}$. The area of the ellipse u profiles the accuracy of the localization. Recall that the event $\mathcal{E}(c)$ determines the area of $\mathcal{U}$ in the physical and $\mathcal{E}(c)$ is determined by outcomes the user has submitted. This means that which RSS fingerprints the user submitted determines the localization accuracy. To maximize the accuracy is equivalent to minimize u, thus the best strategy for the user is to adopt the measurement sequence $\mathcal{V}_n^*$, where

$$\mathcal{V}_n^* = \arg\max_{\mathcal{V}_n \in \mathbb{U}^n} \left\{ \left(\sum_{i \in \mathcal{V}_n} |Z_i| \right)^2 - |\sum_{i \in \mathcal{V}_n} Z_i|^2 \right\}. \tag{32}$$

It is indicated by Eq. (32) that the location determination system needs to search over the entire strategy profile $\mathbb{U}^n$ to find the optimal strategy. In the following discussion, we are to prove that we can narrow down the searching space by eliminating APs with small $|Z_i|$ from the set of all visible APs.

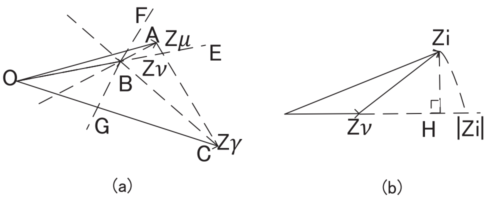


Fig. 6. Proof of Theorem 3.

Theorem 3. *Suppose that a user can choose to measure AP_ν, AP_μ and AP_γ, where the measurement for each AP is denoted by Z_ν, Z_μ and Z_γ, respectively. If Z_ν falls inside the $\triangle OZ_\mu Z_\gamma$ in the complex plane, then $Z_\nu \notin \mathcal{V}_n{}^*$.*

Proof. Fig. 6a shows an arbitrary Z_ν falls inside $\triangle OZ_\mu Z_\gamma$ in the complex plane. O is the origin. We use A and C to represent the corresponding point of Z_μ and Z_γ. We are to prove that any measurement such as Z_ν that falls in the area of the triangle must not be an element of $\mathcal{V}_n^*$ using contradiction.

If $Z_\nu \in \mathcal{V}_n^*$, let $T = \sum_{\mathcal{V}_n^*/\{Z_\nu\}} |Z_i|$, $G = \sum_{\mathcal{V}_n^*/\{Z_\nu\}} Z_i$, where $\mathcal{V}_n^*/\{Z_\nu\}$ stands for the difference sequence of $\mathcal{V}_n^*$ eliminating an arbitrary measurement to AZ_ν. If there are multiple measurements to AP_ν, it makes no difference to eliminate any one of them, since the measurement order does not matter. $\mathcal{V}_n^*/\{Z_\nu\} \cup \{Z_i\}$ stands for the strategy $\mathcal{V}_n^*/\{Z_\nu\}$ plus a measurement to AP_i. T is a real number while G could be a complex number. According to Eq. (32), we should have $u(\mathcal{V}_n^*) \le u(\mathcal{V}_n), \forall \mathcal{V}_n \in \mathbb{U}$, where $u(\mathcal{V}_n)$ is the area of the ellipse in the physical space given the chosen $\mathcal{V}_n$. Then the following equations should hold true for both μ and γ:

$$\begin{cases} \left(\frac{8\pi c^2}{u(\mathcal{V}_n^*)}\right)^2 \ge \left(\frac{8\pi c^2}{u(\mathcal{V}_n^*/\{Z_\nu\}\cup\{Z_\mu\})}\right)^2, \\ \left(\frac{8\pi c^2}{u(\mathcal{V}_n^*)}\right)^2 \ge \left(\frac{8\pi c^2}{u(\mathcal{V}_n^*/\{Z_\nu\}\cup\{Z_\gamma\})}\right)^2. \end{cases} \tag{33}$$

This is equivalent to prove $(T + |Z_\nu|)^2 - |G + Z_\nu|^2 \ge (T + |Z_i|)^2 - |G + Z_i|^2$, for $i = \mu$ and γ. According to Eq. (32), we should prove that

$$\frac{G}{T}(Z_i - Z_\nu) \ge |Z_i| - |Z_\nu|. \tag{34}$$

Let θ_i be the angle between Z_ν and $Z_i - Z_\nu$, then $|Z_i| - |Z_\nu| > |Z_i - Z_\nu| cos(\theta_i)$ as shown in Fig. 6b. This inequality still holds for the case $|Z_i| < |Z_\nu|$ or $\theta_i > \frac{\pi}{2}$, where the proof is straight and thus skipped due to the limitation of space. It is straightforward that $|G| < T$, therefore $|\frac{G}{T}| < 1$. Thus if Z_ν were to be an element of $\mathcal{V}_n^*$, the following two equations should both be true:

$$e(Z_\mu - Z_\nu) > |Z_\gamma - Z_\nu| cos(\theta_\mu), \tag{35}$$

$$e(Z_\gamma - Z_\nu) > |Z_\gamma - Z_\nu| cos(\theta_\gamma), \tag{36}$$

where e is a unit vector.

In Fig. 6a, OB and OE are collinear. We draw two lines BF and BG so that $\angle EBA = \angle ABF$, $\angle EBC = \angle CBG$. Eq. (35) indicates the range of direction for e is from BE to BF (counterclockwise); Eq. (36) indicates the range of direction for e is from BE to BG (clockwise); $\angle ABC < \pi$, which means that it is impossible for the two scopes to overlap, which means there is no such e that makes Eqs. (35) and (36) true at the same time, and the inequalities (33) can not hold simultaneously. Consequently, Theorem (3) is proved by the contradiction. □

Theorem 3 can be understood as following: If we use a point on the convex plane to represent the measurement Z_i, then there will be many points on the plane representing all possible measurements. Only those points on the convex hull of all points are possible candidates of the best strategy. It is worth mentioning that parameters used for determining the best strategy can be derived by analyzing the fingerprints collected for each AP in the database. There is no need for information about the location of APs, and no need for explicit efforts from users either. Finding the best strategy in a general case turns out to be non-trivial, and we present the details in another work [34].

7 Location Determination with Imperfect Information

7.1 Imperfect Information

The cornerstone underpinning our analysis on localization reliability above is the assumption: the distribution of the RSS at each location $\vec{r}$ is perfectly known. With such perfect information, we can construct an one-to-one mapping from the sample space to the physical space. In particular, we can always find a point on the mean surface of the RSS based on reported fingerprints, and the point on the mean surface $\vec{\mu}(\vec{r})$ corresponds to a location in the physical space $\vec{r}$.

Ideally, the mean of RSS readings at a given location can be perfectly known from the database, if the number of measurements at the location is large enough in the training phase; however, due to the cost of the training phase, the information recorded at the fingerprints database is usually imperfect, and the current crowdsourcing based fingerprints collection is unable to guarantee the quality of submitted fingerprints. In particular, crowdsourcing workers submit the current location and corresponding RSS fingerprints observed to the localization server in an opportunistic sensing manner [18], [21], [26], [27], where the location of the reporting worker is estimated with dead reckoning by utilizing the inertial measurement unit (IMU) of the worker's mobile device such as accelerometer, magnetometer and gyroscope [26], [27]. Due to the IMU error, the worker may incorrectly report the position where fingerprints are collected. As a result, the perfect information is usually unavailable in the database.

The consequence of the imperfect information is that the value of $\vec{\mu}(\vec{r})$ for each location in the physical space is inaccurate. A natural question to ask is: How the imperfect information will impact the reliability of location determination? In particular, what is the deviation from the true probability that a user can be correctly localized, which is incurred by the imperfect information? With limited number of measurements in the training phase, what is the best localization reliability can be obtained? These important issues are to be addressed in the following.

7.2 Impact on Localization in One-Dimensional Space

7.2.1 *One-Time Measurement for Single AP*

Recall our investigation of the simple case where the fingerprint is measured only once with respect to a single AP. The

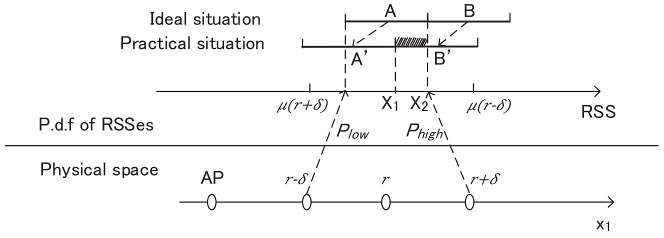


Fig. 7. One-dimensional localization with imperfect information.

domain E in the sample space corresponding to the δ neighborhood of $\vec{r}$ in the physical space is a line segment, where the endpoints of the segment are P_{high} and P_{low}, respectively. If the database has perfect information of fingerprints, sequential line segments in the physical space should be corresponding to sequential line segments in the sample space, as the ideal situation shown in Fig. 7, where A and B are midpoints of the two line segments, respectively.

However, the practical situation is that the information can be derived from the imperfect database, which means that the region E can migrate to somewhere else, as shown in the figure. We can use the midpoint to denote the line segment itself. If the values of submitted fingerprints fall in the shadow area as shown in Fig. 7, the server will determine that the user's physical location should be corresponding to the line segment B in the sample space with imperfect information; however, the user's actual location is in fact corresponding to the line segment A. The localization server can mistakenly determine the user's location due to the imperfect information.

Points on line segments in the practical situation could be regarded as points on line segments in the ideal situation after a random migration as shown in Fig. 7. For any point on the δ neighbourhood of r in the physical space, there is a corresponding point in the sample space. We assume that users appear on each point of the neighbourhood with the same probability. We use x_0 to denote the corresponding point in the sample space for a given point in the physical space, X_1 to denote the right boundary of line segment A after the random migration and X_2 the right boundary of line segment A before the random migration.

In the user's perspective, the probability deviation of correct localization is the absolute value of the difference between the probability that the user should be localized in certain location with perfect information and that with imperfect information. If we define such a probability deviation as the probability error, then the probability error caused by the imperfect database in this particular case is

$$p_{e_1} = \int_{X_1}^{X_2} \frac{1}{L_0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-x_0)^2}{2\sigma^2}} dx, \tag{37}$$

where $L_0 = 2\delta$.

We now consider a general case. Suppose that the length of the line segment A is L, and we set the origin of the horizontal axis to be at the left endpoint of line segment A, then $X_2 = L$, where $L = P_{high} - P_{low}$ as shown in Eq. (4). It is not straightforward to determine the coordinate of X_1, because points on the line segment A can migrate to anywhere in the sample space. A key observation is that X_1 is actually a P_{high} on the line segment A', thus $X_1 = L + \frac{r_{1x}+r_{2x}}{2}$, where r_{1x} and r_{2x} are deviations of the two midpoints A and B in the practical situation, respectively. Note that r_{1x} and r_{2x} are deviations along the horizontal axis, where deviating to the right is positive and to the left is negative. As a result, the probability error considering the general case as shown in the figure is

$$p_{e_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{e_1} \cdot \frac{N}{2\pi\sigma^2} e^{-\frac{N(r_{1x}^2+r_{2x}^2)}{2\sigma^2}} dr_{1x} dr_{2x}. \tag{38}$$

Since the error can also happen to the line segment A and the line segment left to A, the overall error probability is

$$P_e = 2 \int_0^{L_0} p_{e_2} dr, \tag{39}$$

where r denotes user's physical location in the coordinate system. Consider a very small δ, the mean of the RSS is not changing dramatically according to Eq. (2), thus we can apply local linearization to points in both sample space and physical space, which means that the length of a line segment in the sample space is proportional to that in the physical space

$$\begin{aligned} \mu(\vec{r'}) - \mu(\vec{r}) \approx& (\vec{r'} - \vec{r}) \cdot \nabla\mu(\vec{r}) \\ \approx& \left|(\vec{r'} - \vec{r})\right| \cdot |\nabla\mu(\vec{r})| \cdot \cos\varphi, \end{aligned} \tag{40}$$

where φ is the angle between two vectors: $\vec{r'} - \vec{r}$ and $\nabla\mu(\vec{r})$. In one dimensional case, the angle φ would be either 0 or π, leading to $|\cos\varphi| = 1$ and

$$\left|\mu(\vec{r'}) - \mu(\vec{r})\right| \approx \left|(\vec{r'} - \vec{r})\right| \cdot |\nabla\mu(\vec{r})|. \tag{41}$$

Then we have

$$P_e = 2 \int_0^{L} p_{e_2} \frac{L}{L_0} dx_0, \tag{42}$$

where every parameter can be obtained from the database in practice.

7.2.2 Multiple Measurements for Multiple APs

We now extend our analysis to the case where the database contains fingerprints measured multiple times with respect to multiple APs. Assume that the number of measurements is n, then the sample space is n dimensional. Suppose that nodes A and B are two points on the mean surface of the sample space. The challenge comes with the n-dimension is that: the two nodes can migrate to any positions in the space, which results in that the drifted points may not be on the mean surface thus making probability error analysis extremely complicated. This is illustrated in Fig. 8, where A' and B' denote the drifted means in the practical situation, respectively.

In Fig. 8, AB is an one-dimensional line segment and we could use a hyperplane to cut it in the middle. Since A and B denote means of two locations respectively, all reported fingerprints fall in the left side of the hyperplane should be determined to be at the location corresponding to A. All reported fingerprints fall in the right side of the hyperplane should be determined to be at the location corresponding to B. We use $\vec{r_1}$ and $\vec{r_2}$ to denote the deviation of A and B,

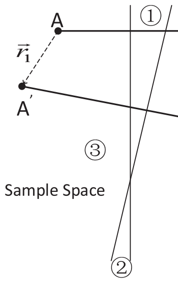


Fig. 8. Imperfect information with multiple measurements for multiple APs.

respectively. Similarly, we can have hyperplane cut line segment $A'B'$ in the middle, and each side of the hyperplane represents those fingerprints that can entail two different localization results in the practical situation.

It is straightforward that if reported fingerprints fall in area 1, the location determination result with imperfect database is different from that with perfect database, which incurs error. The location of the user should be determined to be corresponding to the area of B, but it is determined to be at the location corresponding to A. If reported fingerprints fall in area 2, the location of the user should be determined to be corresponding to the area of A, but it is determined to be at that corresponding to B. Localization errors happen when any of the events happening, because the imperfect information makes the serve believe that the boundary is the hyperplane intersecting with $A'B'$ while the real boundary is the one intersecting with AB.

The probability deviation that the user is correctly localized can be derived if area 1 and 2 can be mathematically characterized; however, the challenge is that it is difficult to imagine the shape of areas in the n-dimensional sample space. Line segment AB is one-dimensional, so its bisecting hyperplane is $n-1$ dimensional. Similarly, the bisecting hyperplane of line segment $A'B'$ is also $n-1$ dimensional. Although AB and $A'B'$ are not in the same hyperplane, their bisecting hyperplanes intersect with each other thus sharing $n-2$ dimensions. Consequently, we could always rotate the coordinate system, so that the projections of the two line segments are in the same plane while the rest of the shared $n-2$ dimensions orthogonal to the plane. That is, no matter how many dimensions the sample space has, we can always illustrate the situation in the sample space as shown in Fig. 8.

For any point in the δ neighbourhood of r in the physical space, there is a corresponding point in the sample space. The point in the sample space can also be mapped into a two-dimensional surface after the coordinate system rotation as described above. We use (x_0, y_0) to denote the coordinate of the point in the system after rotation. Then the probability error for the particular case as shown in Fig. 8 is

$$\begin{aligned} p = &\int_{area1} \frac{1}{L_0}\frac{1}{2\pi\sigma^2} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} dxdy - \\ &\int_{area2} \frac{1}{L_0}\frac{1}{2\pi\sigma^2} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} dxdy \\ = &\frac{1}{L_0}(p_1 - p_2), \end{aligned} \tag{43}$$

where $L_0 = 2\delta$. Note that fingerprints fall into area 3 will definitely make the system to localize the user at the location corresponding to A in both practical and ideal situation, thus the deviation is only incurred by the difference between area 1 and 2.

Consider the two hyperplanes bisecting AB and $A'B'$, we now study the angle between the two hyperplanes so that the specific expression of probability error can be derived. It is straightforward that

$$\begin{aligned} \cos\theta &= \frac{\vec{AB}\cdot\vec{A'B'}}{|AB||A'B'|} \\ &= \frac{(L,0)\cdot(r_{2x}-r_{1x}+L, r_{2y}-r_{1y})}{L\times\sqrt{(r_{2x}-r_{1x}+L)^2+(r_{2y}-r_{1y})^2}}, \end{aligned} \tag{44}$$

$$\sin\theta = \frac{|r_{2y}-r_{1y}|}{L}, \tag{45}$$

where r_{1x}, r_{1y}, r_{2x} and r_{2y} denote deviations of A and B in two dimensions, respectively. Assume that N times of measurements are performed independently in the training stage to build up the fingerprints database, then the drift distance of A and B follow the Gaussian distribution with mean 0 and variance value $\frac{\sigma}{\sqrt{N}}$, according to law of large numbers. That is, $r_{1x} \sim N(0, \frac{\sigma}{\sqrt{N}})$, $r_{1y} \sim N(0, \frac{\sigma}{\sqrt{N}})$, $r_{2x} \sim N(0, \frac{\sigma}{\sqrt{N}})$, $r_{2y} \sim N(0, \frac{\sigma}{\sqrt{N}})$. Consequently, values of r_{1x}, r_{1y}, r_{2x}, r_{2y} could appear in the range $\frac{-3\sigma}{\sqrt{N}} \sim \frac{3\sigma}{\sqrt{N}}$ with high probability. We are able to perform measurements many times so that r_{1x}, r_{1y}, r_{2x}, r_{2y} are all small polynomial terms compared with L, thus the following approximation can be obtained:

$$\cos\theta \approx \frac{|r_{2x}-r_{1x}+L|}{L}. \tag{46}$$

Since we have $\theta < \frac{|\vec{r_1}|+|\vec{r_2}|}{L}$, and we can also make N big enough to have the following approximation:

$$\sin\theta \approx \tan\theta \approx \theta. \tag{47}$$

Note that $\int_{area1} - \int_{area2} = \int_{area1+area3} - \int_{area2+area3}$. For the integration over area 1 and 3, we have

$$p_1 = \int_{-\infty}^{\frac{L/2-x_0}{\cos\theta}} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\pi\sigma^2}} dx, \tag{48}$$

$$p_2 = \int_{-\infty}^{\frac{L}{2}-x_0} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\pi\sigma^2}} dx. \tag{49}$$

Taking all possible situations for the data drift incurred by imperfect information into account, the probability error for location determination is

$$p = \left|\frac{1}{L_0}\left[erf\left(\frac{L/2-x_0}{\cos\theta\sqrt{2}\sigma}\right) - erf\left(\frac{L/2-x_0}{\sqrt{2}\sigma}\right)\right]\right|. \tag{50}$$

The error can also happen in the line segment AB and the line segment left to AB such as the line segment AC shown in Fig. 9. Consequently, the location determination error is determined by $area1 + area5 - area2 - area4$, which we could be put as $[(area1 + area3 + area4 + area5 + area6) - (area2 + area3 + area4 + area5 + area6)] + [(area5 + area1 + area2 + \quad area3 + area7) - (area4 + area1 + area2 + area3$

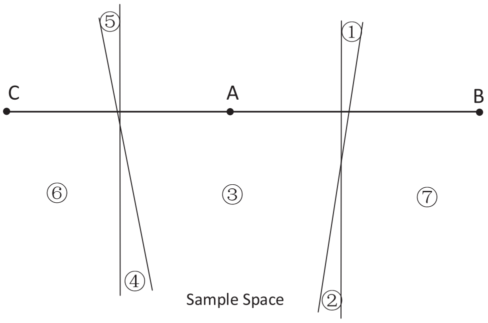


Fig. 9. One-dimensional localization with multiple measurements for multiple APs.

$+\ area7)]$. Thus we get the final probability error for the multiple measurements over multiple APs

$$\begin{aligned}P_e =&2\int_0^L\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{N^2}{4\pi^2\sigma^4}e^{-\frac{N(r_{1x}^2+r_{2x}^2+r_{1y}^2+r_{2y}^2)}{2\sigma^2}}p\\ &dr_{1x}dr_{1y}dr_{2x}dr_{2y}\frac{L}{L_0}dx_0\\ =&2\int_0^L\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{N^2}{4\pi^2\sigma^4}e^{-\frac{N(r_{1x}^2+r_{2x}^2+r_{1y}^2+r_{2y}^2)}{2\sigma^2}}\\ &\left[erf\left(\frac{L/2-x_0}{\cos\theta\sqrt{2}\sigma}\right)-erf\left(\frac{L/2-x_0}{\sqrt{2}\sigma}\right)\right]\\ &dr_{1x}dr_{1y}dr_{2x}dr_{2y}\frac{L}{L_0^2}dx_0.\end{aligned}\tag{51}$$

7.3 Impact on Localization in Two-Dimensional Space

The practical indoor localization system partitions the two-dimensional physical space into blocks [20], [21], such as the one shown in the left part of Fig. 10, where the center of each block represents the block itself. Let us first consider the ideal case with perfect information. Recall that the corresponding image in the sample space with respect to each point in the physical space is an point on the mean surface M as shown in the right part of Fig. 10. We use four hyperplanes to surround the point A on the mean surface M, where each hyperplane is orthogonally cutting the line segment between A and the neighbouring node in the middle. According to the principle of MLE, if reported fingerprints fall in the surrounded area, then the system will localize the user's location to be at block A. It is worth mentioning that we here do not adopt the hyper-cylinder discussed in Section 5 as the boundary in the sample space, because such a boundary can leave certain areas in the physical space uncovered.

We now provide the mathematical expression of such a surrounded area as shown in Fig. 10. Suppose that $A=[x_{1A},x_{2A},\dots,x_{nA}]^T$ and $B=[x_{1B},x_{2B},\dots,x_{nB}]^T$, then the bisector plane of line segment $\vec{AB}$ is

$$\vec{h_1}=\frac{\vec{AB}}{2}=\frac{(x_{1A}-x_{1B},x_{2A}-x_{2B},\dots,x_{nA}-x_{nB})}{2}.$$

The other three bisector planes could also be presented in a similar manner. If we use use area 1 to denote the surrounded area, then area 1 can be determined by: $\vec{h_1}\cdot\vec{r}\le|\vec{h_1}|^2$, $\vec{h_2}\cdot\vec{r}\le|\vec{h_2}|^2$, $\vec{h_3}\cdot\vec{r}\le|\vec{h_3}|^2$ and $\vec{h_4}\cdot\vec{r}\le|\vec{h_4}|^2$.

If we use area 0 to denote block A in the physical space, and assume that the user will appear in any point of block A with identical probability, then the reliability for the ideal situation is

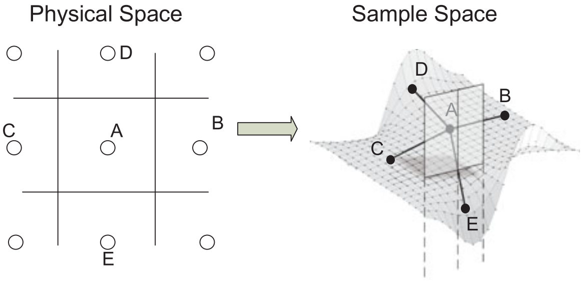


Fig. 10. Mapping from physical space to sample space.

$$\begin{aligned}P_{e_1}=&\iint_{area0}\mathrm{d}x_0\mathrm{d}y_0\int\cdots\int_{area2}\left(\frac{\sqrt{N}}{2\pi\sigma}\right)^n\\ &e^{-\frac{N[(x-x_1)^2+(x-x_n)^2]}{2\sigma^2}}\frac{1}{S_0}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n,\end{aligned}\tag{52}$$

where S_0 is the area of block A.

In the practical situation, the nodes A, B, C, D and E migrate to A', B', C', D' and E' in the sample space. We can also use four hyperplanes to surround a corresponding area 2, so that if reported fingerprints fall in area 2, the user is localized in block A in the physical space. Similarly, the reliability for the practical situation is

$$\begin{aligned}P_{e_2}=&\iint_{area0}\mathrm{d}x_0\mathrm{d}y_0\int\cdots\int_{area2}\left(\frac{\sqrt{N}}{2\pi\sigma}\right)^n\\ &e^{-\frac{N[(x-x_1)^2+(x-x_n)^2]}{2\sigma^2}}\frac{1}{S_0}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n.\end{aligned}\tag{53}$$

Consequently, the probability error is $P_e=\left|P_{e_1}-P_{e_2}\right|$.

It is extremely difficult to give a close-form expression of P_e; however, the probability error analysis inspires us to consider a very special case when determining the surrounded area in the sample space mentioned above, which could potentially incur large localization error. That is, what if the nodes A, C and D are on the same straight line, which means that the surrounded area is actually an open area. Although the general mathematical expressions also hold in the special case, the consequence in location determination is that large-scale localization error could happen. The physical meaning of the open area is that the user could be localized in physical locations corresponding to faraway areas in the sample space. In particular, if the reported fingerprint is $\mu(\vec{r'})$, based on which the user is most likely at location $\vec{r'}$, the system however still could localize the user to be at some location faraway from $\vec{r'}$.

Such a phenomenon can be avoided by utilizing the best fingerprints reporting strategy when constructing the database. We first illustrate why A, C and D could be on the same straight line, as shown in Fig. 11. The left part of the figure shows the setting of the physical space, and the right part shows an example of two-dimensional sample space, which means the number of measurements is two. In the fingerprints collection phase, a site surveyor standing at A could measure AP_1 and AP_2 once respectively, surveyors at C and D measure AP_1 and AP_2 twice respectively, then the corresponding nodes of these physical locations in the

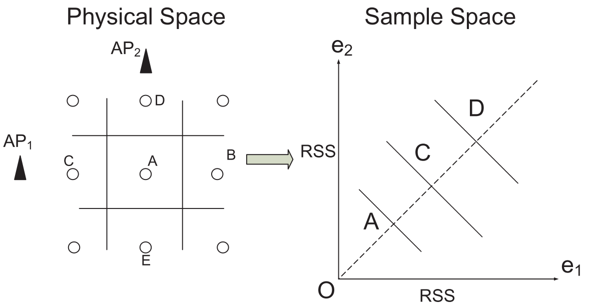


Fig. 11. Special case.

sample space are on the same straight line as shown in the right part of Fig. 11.

The way surveyors construct the database described above could be very possible if the best strategy is not considered. This is because surveyors usually prefer to measure APs with strongest signal strengths, such as AP_1 and AP_2 with respect to C and D, respectively. However, referring to the best strategy theory could reveal that such a survey can be of little avail for localization. Take location C for example, if the surveyor measures AP_1 twice, the corresponding complex parameter Z_1 are on the same straight line in the complex plane, which makes $(\sum_{i\in\mathcal{V}_n}|Z_i|)^2 - |\sum_{i\in\mathcal{V}_n} Z_i|^2 = 0$, where $\mathcal{V}_n = \{1, 1\}$ according to the surveying process. This could be better understood by reviewing Eq. (32). Although the best strategy presented earlier is for location estimation, it also provides guidance for the training phase.

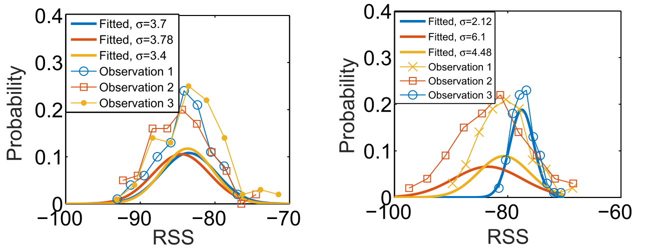
However, if the number of measurements is small, it may happen that the migrated points in the sample space are collinear, which will incur large deviation from the reliability. Fortunately, our numerical analysis shows that it is very difficult for such subtle migration of points in the sample space to happen, and deviations of reliability incurred by imperfect information is usually very small.

8 Experimental Results

We conduct experiments with the trace data collected by the EVARILOS testbed [35], in order to verify our theory. The data are collected in an unmanned utility room with many metal objects termed as "Zwijnaarde", where there is almost no outside interference and no persons are present in the environment. The trace data contains 144557 combinations of access point (AP) and reference point (RP), and each (AP, RP) tuple contains a number of RSS raw measurements. Detailed description of the testbed and data could be found in [36], [37], [38].

Verification of Main Assumptions. Although the rationale of our main assumption about the radio propagation has been explained in Section 3, where a number of work adopting the similar assumption is briefly surveyed, we still validate our assumption by performing analysis of the trace data from the EVARILOS testbed. We first filter out those unreliable measurements, where there is only 1 or 2 RSS readings recorded or all the RSS readings are exactly the same.

Fig. 12a shows three adjacent RPs' RSS observations and fitted curves. The RSSes observed at the three locations (1,290, 1,980), (1,290, 1,270) and (1,890, 1,270) are with respect to the same AP at (1,000, 1,712), and the RPs are round 3.5 meters from each other. There are totally 272, 146 and 98 observations at the three RPs, respectively. Observation curves represent the proportions of the observed RSS value. We fit the observation curves, and find they are approximately to be Gaussian distribution, with skewness and kurtosis less than $|0.5|$ and $|3.3|$ respectively. It can be seen that the observed RSS readings at adjacent RPs have similar value of σ, and the mean does not change dramatically.



(a) Adjacent locations W.R.T. the same AP.

(b) The same location W.R.T. different APs.

Fig. 12. Characteristics of radio propagation.

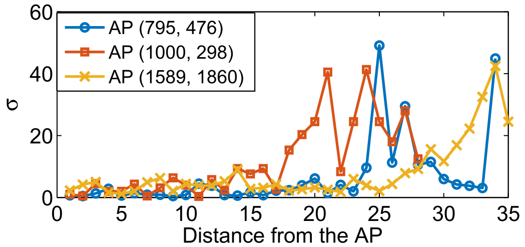
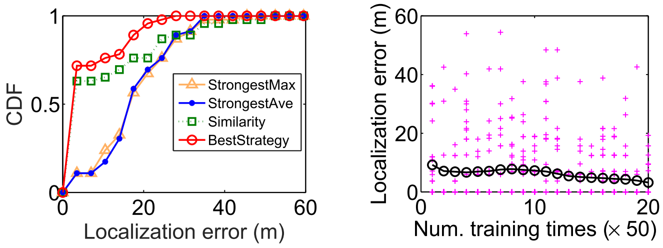


Fig. 13. Change of σ with the distance.

Fig. 12b shows the RSS values observed at the same RP (2,490, 1,270) with respect to three different APs at (2,644, 1,500), (3,998, 728), and (1,000, 600), respectively. The total numbers of RSS records with respect to each AP are 173, 208, and 196; the fitted curves are with the skewness and kurtosis less than $|0.31|$ and $|3.2|$, respectively. The observations corroborate our assumption that the RSS values observed at the same location with respect to different APs are with quite different values of σ.

Fig. 13 shows the change of σ's value with the distance from the AP. We show the trend of the change with respect to the three APs. We examine RSSes observed at all RPs that are less than 35 meters from the AP, and we calculate the corresponding σ value at each RP. As shown in the figure, 45 percent of the locations' values of σ vary less than 2, and 81 percent of the values of σ vary less than 5, if the RP is less than 17 meters from the AP. If the distance exceeds 23 meters, the change of the value of σ becomes dramatic. This observation corroborates the model in [32], [33], which supports our modeling assumption in Section 3.

Localization Performance. We use the data observed at a part of the RPs as the training set and that observed at the rest of the RPs as the test set to perform localization, with results illustrated in Fig. 14a. We compare performance of the proposed best strategy based on Eqs. (32) and (26) with other three reporting strategies widely used in the previous work [18]. The best strategy is the logical result of our modeling and analysis, thus if it outperforms other frequently used strategies, our modeling and analysis could be validated.



(a) Influence of AP selection.

(b) Influence of training times.

Fig. 14. Localization results.

With *StrongestAvg*, the user measures APs with the strongest average RSSes can be observed at the to-be-determined location in the online phase. With *StrongestMax*, the user measures APs with the strongest RSSes can be observed. With the *Similarity* strategy, APs are first clustered according to the similarity of their generated RSSes and the representative AP of each cluster is then selected. How to compute the similarity metric and how to select the representative AP in a cluster are described in [18]. As can be seen from Fig. 14a, it is obvious that the best strategy outperforms other three strategies. A more detailed statistical results are tabulated in Table 1. Although the best strategy has the largest maximum error due to the randomness of fingerprinting based localization approach, the overall performance is better than the other three strategies.

Fig. 14b illustrates the localization results influenced by the number of training times in the offline phase. We use reliable RSS records in all RPs to do the experiment in order to ensure that there are enough data for training. We use a part of RSS data for the offline phase and the rest of the data for testing. With the number of data used in the offline phase increasing, it is clear that the average localization error is decreasing as shown with the curve in the figure, which corroborates our theoretical analysis. Since for each RP, there are a part of RSS records used for training, if the data for testing are with the same RP, it is possible that the minimum localization error equals 0 as shown in the figure.

9 Conclusion

We have presented a general probabilistic model to shed light on a fundamental question: how good the RSS fingerprinting based indoor localization can achieve? Concretely, we have presented the probability that a user can be localized in a region with certain size, given the RSS fingerprints submitted to the system. We have revealed the reliability of location estimation. Moreover, we have shown that there exists an optimal fingerprints reporting strategy that can achieve the best accuracy for given reliability. Further, we have analysed the influence of imperfect database information on the reliability of localization, and found that the impact of imperfect information is still under control with reasonable number of samplings in the training phase.

TABLE 1
Localization Errors

Metric	StrgstMax	StrgstAvg	Similarity	BestStrategy
Avg. err [m]	19.3	19.3	18.4	16.4
Min. err [m]	6.0	3.8	3.6	3.5
Max. err [m]	42.6	37.6	48.0	55.8
Med. err [m]	17.1	18.4	14.0	11.4

Acknowledgments

The authors thank Dr. Filip Lemić and Dr. Eli De Poorter for sharing the EVARILOS trace data. This work is supported by NSF China (Nos. 61572319, U1405251, 61532012, 61325012, 61521062); National Mobile Communications Research Laboratory, Southeast University (No. 2014D07). A preliminary version of this article appeared in the *Proceedings of the 34th IEEE International Conference on Computer Communications* (IEEE INFOCOM 2015).

References

[1] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proc. 10th Annu. Int. Conf. Mobile Comput. Netw.*, 2004, pp. 70–84.

[2] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–32, 2013.

[3] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proc. IEEE INFOCOM*, 2000, pp. 775–784.

[4] K. Chintalapudi, A. P. Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw.*, 2010, pp. 173–184.

[5] 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. Sensor Ad Hoc Commun. Netw.*, 2004, pp. 406–414.

[6] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proc. IEEE INFOCOM*, 2004, pp. 400–408.

[7] P. Castro, P. Chiu, T. Kremenek, and R. Muntz, "A probabilistic room location service for wireless networked environments," in *Proc. 3rd Int. Conf. Ubiquitous Comput.*, 2001, pp. 18–34.

[8] M. A. Youssef, A. Agrawala, and A. U. Shankar, "WLAN location determination via clustering and probability distributions," in *Proc. 1st IEEE Int. Conf. Pervasive Comput. Commun.*, 2003, pp. 143–150.

[9] M. Youssef and A. Agrawala, "The horus WLAN location determination system," in *Proc. 3rd Int. Conf. Mobile Syst. Appl. Serv.*, 2005, pp. 205–218.

[10] H. Hashemi, "Impulse response modeling of indoor radio propagation channels," *IEEE J. Sel. Areas Commun.*, vol. 11, no. 7, pp. 967–978, Sep. 2006.

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

[12] M. Youssef, M. Abdallah, and A. Agrawala, "Multivariate analysis for probabilistic WLAN location determination systems," in *Proc. 2nd Annu. Int. Conf. Mobile Ubiquitous Syst.: Netw. Serv.*, 2005, pp. 353–362.

[13] C. Oestges, N. Czink, B. Bandemer, P. Castiglione, F. Kaltenberger, and J. Paulraj, "Experimental characterization and modeling of outdoor-to-indoor and indoor-to-indoor distributed channels," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2253–2265, Nov. 2010

[14] C. Fischione, F. Graziosi, and F. Santucci, "Approximation for a sum of on-off lognormal processes with wireless applications," *IEEE Trans. Commun.*, vol. 55, no. 10, pp. 1984–1993, Oct. 2007.

[15] F. Graziosi and F. Santucci, "A general correlation model for shadow fading in mobile radio systems," *IEEE Commun. Lett.*, vol. 6, no. 3, pp. 102–104, Mar. 2002.

[16] R. Battiti, M. Brunato, and A. Delai, "Optimal wireless access point placement for location-dependent services," Univ. Trento, Trento, Italy, Tech. Rep. DIT-03-052, 2003. [Online]. Available: http://eprints.biblio.unitn.it/489/1/DIT-03-052-withCover.pdf

[17] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless LANs," *Comput. Netw.*, vol. 47, no. 6, pp. 825–845, 2005.

[18] K. Chintalapudi, A. P. Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw.*, 2010, pp. 173–184.

[19] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw.*, 2012, pp. 293–304.

[20] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: Wireless indoor localization with little human intervention," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw.*, 2012, pp. 269–280.

[21] C. Wu, Z. Yang, and Y. Liu, "Smartphones based crowdsourcing for indoor localization," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2199–2214, Oct. 2013.

[22] M. A. Youssef and A. Agrawala, "On the optimality of WLAN location determination systems," 2003. [Online]. Available: http://www.cs.umd.edu/ moustafa/papers/cnds04.pdf

[23] H. Liu, et al., "Push the limit of WiFi based localization for smartphones," in *Proc. 18th Annu. Int. Conf. Mobile Comput. Netw.*, 2012, pp. 305–316.

[24] H. Liu, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Accurate WiFi based localization for smartphones using peer assistance," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2199–2214, Oct. 2013.

[25] G. Chandrasekaran, et al., "Empirical evaluation of the limits on localization using signal strength," in *Proc. 6th Annu. IEEE Commun. Soc. Conf. Sensor Mesh Ad Hoc Commun. Netw.*, 2009, pp. 1–9.

[26] G. Shen, Z. Chen, P. Zhang, T. Moscibroda, and Y. Zhang, "Walkie-Markie: Indoor pathway mapping made easy," in *Proc. 10th USENIX Conf. Networked Syst. Des. Implementation*, 2013, pp. 85–98.

[27] C. Luo, H. Hong, and M. C. Chan, "PiLoc: A self-calibrating participatory indoor localization system," in *Proc. IEEE 13th Int. Symp. Inf. Process. Sensor Netw.*, 2014, pp. 143–153.

[28] A. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU," in *Proc. IEEE Int. Symp. Intell. Signal Process.*, 2009, pp. 37–42.

[29] N. Patwari, J. Ash, S. Kyperountas, A. O. Hero, R. Moses, and N. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 55–69, Jul. 2005.

[30] 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.

[31] M. Angjelichinoski, D. Denkovski, V. Atanasovski, L. Gavrilovska, "Cramér-Rao lower bounds of RSS-based localization with anchor position uncertainty," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2807–2834, May 2015.

[32] International Telecommunication Union (ITU), "Guidelines for evaluation of radio interfacce technologies for IMT-Advanced," Tech. Rep. ITU-R M. 2135-1, Dec. 2009, pp. 30–33. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2135-1-2009-PDF-E.pdf

[33] WINNER II, "WINNER II Channel Models," Tech. Rep. IST-4-027756, Sep. 2007, pp. 44–45. [Online]. Available: http://www.cept.org/files/1050/documents/winner2%20-%20final%20report.pdf

[34] Z. Zhang, D. Liu, S. Zhu, S. Chen, and X. Tian, "Squeeze more from fingerprints reporting strategy for indoor localization," in *Proc. 13th Annu. IEEE Int. Conf. Sens. Commun. Netw.*, 2016, pp. 1–9.

[35] EVARILOS testbed. (2016). [Online]. Available: http://evarilos.intec.ugent.be/

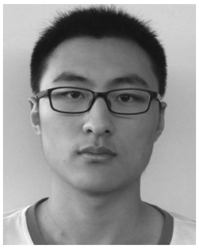
[36] T. V. Haute, et al., "Platform for benchmarking of RF-based indoor localization solutions," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 126–133, Sep. 2015.

[37] T. V. Haute, et al., "Comparability of RF-based indoor localization solutions in heterogeneous environments: An experimental study," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 23, no. 1/2, pp. 92–114, 2016.

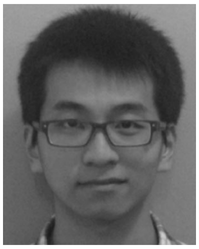
[38] F. Lemic, A. Behboodi, V. Handziski, and V. Wolisz, "Experimental decomposition of the performance of fingerprinting-based localization algorithms," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2016, pp. 695–704.

Xiaohua Tian (S'07-M'11) received the BE and ME degrees in communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2003 and 2006, respectively. He received the PhD degree from the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, in Dec. 2010. Since Mar. 2011, he has been in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, and now is an associate professor with the title of SMC-B scholar. He serves as the column editor of the *IEEE Network Magazine* and the guest editor of the *International Journal of Sensor Networks* (2012). He also serves as the TPC member of IEEE INFOCOM 2014-2017, best demo/poster award committee member of IEEE INFOCOM 2014, TPC co-chair of IEEE ICCC 2014-2016, TPC co-chair of the 9th International Conference on Wireless Algorithms, Systems and Applications (WASA 2014), TPC member of IEEE GLOBECOM 2011-2016, and TPC member of IEEE ICC 2013-2016, respectively. He is a member of the IEEE.

Ruofei Shen received the BE degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. His research interests include indoor localization and TDMA based protocol.



Duowen Liu is currently working toward the BE degree in computer science from Shanghai Jiao Tong University, Shanghai, China, and is expected to graduate in 2017. His research interests include indoor localization and compressive sensing.



Yutian Wen received the BE degree from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2015. His research interest includes crowdsourcing based indoor localization.

Xinbing Wang received the BS degree (with hons.) from the Department of Automation, Shanghai Jiaotong University, Shanghai, China, in 1998, and the MS degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001. He received the PhD degree, major from the Department of electrical and Computer Engineering, minor from the Department of Mathematics, North Carolina State University, Raleigh, in 2006. Currently, he is a professor in the Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, China. He has been an associate editor of the *IEEE/ACM Transactions on Networking* and the *IEEE Transactions on Mobile Computing*, and the member of the Technical Program Committees of several conferences including ACM MobiCom 2012, ACM MobiHoc 2012-2014, and IEEE INFOCOM 2009-2017. He is a senior member of the IEEE.