Hypothesis Testing Based Model for Fingerprinting Localization Algorithms

Arash Behboodi*, Filip Lemic†, Adam Wolisz‡
†Institute for Theoretical Information Technology, RWTH Aachen University
‡Telecommunication Networks Group (TKN), Technische Universität Berlin
Email: arash.behboodi@ti.rwth-aachen.de, {lemic,wolisz}@tkn.tu-berlin.de

Abstract—Despite the popularity of Fingerprinting Localization Algorithms (FPS), general theoretical frameworks for their performance studies have rarely been discussed in the literature. In this work, after setting up an abstract model for the FPS, we show that a fingerprinting-based localization problem can be cast as a Hypothesis Testing (HT) problem and therefore various results from the HT literature can be used to provide insights, guidelines, and performance bounds for the FPS. This includes the scaling limits of error probability in terms of the number of measurements and the precise characterization of localization error. The provided results hold for the general FPS. Additionally, Received Signal Strength (RSS)-based fingerprinting algorithms are particularly considered from the theoretical viewpoint due to their widespread practical usage. Simulations and experimental results characterize numerically the findings of the theoretical framework and demonstrate its consistency with realistic localization scenarios.

I. INTRODUCTION

Precise location of people and devices, both indoors and outdoors, is an essential information for future networks as an enabler of context-aware services, location-aware and pervasive computing, and ambient intelligence. For that reason, variety of localization solutions have been proposed and studied in the recent years. Radio-Frequency (RF)-based solutions are popular due to their low cost and availability. They leverage available technologies such as IEEE 802.11 (WiFi), IEEE 802.15.4 (ZigBee), IEEE 802.15 (Bluetooth), and Ultra-Wide Band (UWB). Different RF signal features can be used for localization including Time of Arrival, Angle of Arrival, Received Signal Strength, and the quality of RF transmission in digital communication channels (e.g. Link Quality Information, Bit-Error Ratio). Based on signal features processing, one can roughly distinguish three categories of localization algorithms, i.e. geometry-based, fingerprinting, and Bayesian-based ones [1]. RF fingerprinting algorithms are particularly attractive because they rely on available wireless infrastructures and therefore avoid a costly setup. These algorithms utilize specific signal features that are correlated with spatial characteristics of the environment. The idea is to use the observed feature at a location to construct an identification tag for that location which is called the fingerprint. A database is constructed, explicitly or implicitly, by gathering fingerprints of different locations. A pattern matching algorithm identifies the location by finding the most similar fingerprint in the database to the reported fingerprint from an unknown location.

There are various works in the literature studying different aspects of fingerprinting algorithms [2]. A survey of WiFi-based fingerprinting algorithms is provided in [3] and variants of such algorithms are studied in e.g. [4], [5]. The main theoretical work on fingerprinting algorithms is [6], where the authors provide an analysis of the effect of the number of visible Access Points (APs) and radio propagation parameters on the performance of fingerprinting algorithms. These results have been extended to a complexity analysis in [7]. The authors in [8] propose a probabilistic model RSS-based fingerprinting relating the location to RSS. The authors also discuss the performance of fingerprinting using likelihood-based detection algorithms and provide insights for fingerprinting design. The scalability of fingerprinting has been discussed in [9], where the authors suggest the scalability improvement by reducing a training database size. In [10], the authors discuss the robustness of fingerprinting to outliers and effects such as shadowing. However, various insights of fingerprinting, such as the scaling limits of errors probabilities or the characterization of localization error, have not been yet discussed from the theoretical perspective, despite the importance of such discussion for the optimal design of fingerprinting-based localization systems.

In this paper, the signal feature is assumed to be a random variable with a probability distribution that varies with the location. The problem of localization then boils down to finding the probability distribution underlying the observed measurements, which essentially is a hypothesis testing problem. Based on this observation, we prove series of theorems establishing fundamental limits of fingerprinting algorithms. This includes a proof for existence of a fingerprinting algorithm with completely accurate localization and also showing that the inaccuracy decays exponentially with the number of measurements. Kullback-Leibler (KL) divergence between probability distributions of a selected feature for fingerprinting at two different locations plays a central role in these results. Different consequences of these theorems for tuning the parameters of fingerprinting algorithms are discussed. The general framework is then instantiated for the RSS-based algorithms, as these are the most established instances of fingerprinting algorithms. Finally, we show that the implications of the developed theory hold for realistic scenarios that do not necessarily adhere to the assumptions made in the framework.

II. SYSTEM MODEL

A. Fingerprint

Fingerprinting algorithms are based on associating a fingerprint to a location, which is later used for the identification of that location. A specific feature of environment is chosen as
basis for creating a fingerprint. We chose the term environment to include multiplicity of possibilities whether a particular infrastructure is assumed for localization purpose or not. The signal feature is denoted by $S$ and belongs to the feature space $S$. $m$ consecutive observations of the signal feature $S = (S_1, \ldots, S_m) \in S^m$ are used as a random variable related to the location $u$ through the conditional probability $P_{\mathbf{S}|u}$. In general, the consecutive measurements of a signal feature can be statistically correlated. Fingerprints are then constructed based on observations $S$ at different locations. Fingerprints belong to the space $\mathcal{X}$, which may be different from $S$.

**Definition 1 (Fingerprint):** A fingerprint creating function $f$ is a mapping $S^m \rightarrow \mathcal{X}$ that assigns to observations $S$ an element $X$ called the fingerprint at the location $u$.

For example, consider a FPS measuring RSS from an anchor. Let the fingerprint be a Gaussian distribution fitted to the vector of $m$ measured RSS. The fingerprinting space is a space of Gaussian probability distributions and the signal feature is the received power. Since received power is a function of distance to the anchor, so are the fingerprints. There are some requirements for a general fingerprint to be useful for localization. In general if the probabilities $P_{X | u_1}$ and $P_{X | u_2}$ are close enough, i.e., if $d(P_{X | u_1}, P_{X | u_2}) \leq L$ where $d$ is a metric, then $u_1$ and $u_2$ should also be close which is $\|u_1 - u_2\| \leq s(L)$ where $s(L) \rightarrow 0$ whenever $L \rightarrow 0$. We call this spatial stability of the fingerprint. The quantitative characterization of stability depends on the metric chosen for measuring closeness of probabilities. Fingerprints should also be robust in time. This means that fingerprints should not change dramatically from training to measurement phase.

**B. Fingerprinting Algorithm**

After a signal feature and a fingerprint creating function have been selected, the next step is to design the fingerprinting algorithm. The first step called the training phase is a process of constructing a database consisting of pairs of locations and their fingerprints. Out of an uncountable set of locations in the localization space, only a finite number of locations can be chosen for direct measurement to construct the database. The training database can be constructed through extensive measurements, through simulation-based radio-map construction [11], [12] or the combination of both. The set of training locations, called a training grid, is denoted by $\Lambda \subset \mathbb{R}^d$. The region of nearest points to each training location $v \in \Lambda$ is called its Voronoi region denoted by $V_v$. The training locations in $\Lambda$ divide the localization space into regions. After choosing the training locations, a specific feature of environment is chosen for creating a fingerprint. The training database is created by measuring the signal feature $S$ at the training locations inside the training grid $\Lambda$. At each location $v$, multiple measurements are collected and fingerprints are constructed accordingly. The training database $\mathcal{D}$, which consists of pairs like $(v, X)$, is a subset of $\Lambda \times \mathcal{X}^n$. The next part consists of finding the location of a target node.

The fingerprinting algorithm reports the estimated location of the target node based on the observed signal feature. A target node placed at the location $u$ measures the selected feature $m'$ times and creates a fingerprint $X_u \in \mathcal{X}^{m'}$. In general, the number of measurements in the training phase and in the localization phase can be different, although we assume $n' = n$. After acquiring the fingerprint, a pattern matching function $g$ is used to estimate the target node’s location $\hat{u}$ based on the acquired fingerprint $X_u$ and fingerprints in the training database. The function $g$ is a mapping from $\mathcal{X}^n$ to the localization space $\mathcal{R}$. The pattern matching function can be regarded as a composition of multiple functions. For instance, a similarity kernel function can first find the most similar fingerprints in the database. Then, one can additionally use $k$–nearest neighbor methods to average between the $k$ closest training locations, rather than declaring single training location. The average can be weighted or not, adaptive or non-adaptive [13]. In any case, $k$–NN methods will have a set of estimated locations $\Lambda_e$, which is not equal to the original set of training locations $\Lambda$.

Based on the previous discussion, we can finally specify what a fingerprinting algorithm is.

**Definition 2 (Fingerprinting algorithm):** A fingerprinting algorithm for a localization space $\mathcal{R}$ consists of:

- A set of training locations $\Lambda$;
- A fingerprint creating function $f : S^m \rightarrow \mathcal{X}^n$, mapping a measured signal feature $S$ to a fingerprint $X$ in $\mathcal{X}^n$;
- A training database $\Lambda \times \mathcal{X}^n$ consisting of pairs of locations and their fingerprints;
- A pattern matching function $g : \mathcal{X}^n \rightarrow \mathcal{R}$ reports the final location by comparing the target node’s fingerprint to the ones from the training database.

The advantage of this abstraction, as it can be seen later, is its applicability in various scenarios. Usage of this abstraction for the RSS-based fingerprinting is discussed later in the paper.

**C. Performance of Fingerprinting Algorithms**

In this section, we introduce a framework for evaluating the performance limit of fingerprinting algorithms. The main performance metric for a localization algorithm is the localization error. Let $\hat{u}$ be the output of the FPS for a target node located at $u \in \mathcal{R}$. The localization error is defined as $\Delta(u) = \|\hat{u} - u\|$. Similar to the definition of error in information theory and statistics, it is possible to define the maximum and average localization error for all $u$’s denoted by $\Delta_{\text{max}}$ and $\bar{\Delta}$. Because we have assumed that the feature is a random variable, the localization error $\Delta(u)$ can also be a random variable and so can $\Delta_{\text{max}}$ and $\bar{\Delta}$. An important goal in localization is to reliably guarantee a given accuracy and the following definition formalizes this goal.

**Definition 3 (Achievable localization error):** The maximum localization error of $\delta$ is achievable with probability $1 - \epsilon$ if there is a FPS such that $\Pr(\Delta_{\text{max}} > \delta) \leq \epsilon$. In the same way the achievable average localization error can be defined. The definition assumes the theoretical limits $\delta$ and $\epsilon$. Next section establishes that $(\delta, \epsilon)$ can be made arbitrarily small by increasing the number of training points and measurements.
III. FINGERPRINTING AS HYPOTHESIS TESTING PROBLEM

In this section, we consider a general FPS and provide bounds on its performance. The only assumption for the conditional distribution \( P_{X|u_i} \) is that the measurements are independent and identically distributed. Let us assume a simple localization scenario where the target node is located either at \( u_1 \) or at \( u_2 \). Two kinds of errors can be defined and the aim is to minimize both of these probabilities. \( \alpha(u_1, u_2) \) and \( \beta(u_1, u_2) \) are the probabilities of incorrect identification.

\[
\alpha(u_1, u_2) = P(g(X) = u_2) \quad \text{Target node is at } u_1 \\
\beta(u_1, u_2) = P(g(X) = u_1) \quad \text{Target node is at } u_2.
\]

The problem of localization is then to decide between two probability distributions \( P_{X|u_1} \) and \( P_{X|u_2} \) based on the observation \( X \) and subject to constraints on \( \alpha(u_1, u_2) \) and \( \beta(u_1, u_2) \). The problem is related to statistical hypothesis testing where the goal is to decide between two conditional distributions as a hypothesis testing problem and benefit from abundant sample data. The errors \( \alpha(u_1, u_2) \) and \( \beta(u_1, u_2) \) are respectively called missed detection and false alarm. This connection enables us to regard a fingerprinting design problem as a hypothesis testing problem and benefit from abundant resources materials in that area. As the first step, the following theorem provides fundamental limits on the performance of two-points fingerprinting algorithms. \( D(\| \cdot \|) \) is KL-divergence.

**Theorem 3.1:** Consider a fingerprinting algorithm with fin-
ners prints and the location is announced as the location and otherwise, \( u_2 \) is announced. This basically shows that \( T_n \) is a good candidate for fingerprint and a comparison function for pattern matching. For a Neyman-Pearson test, there are lot of results characterizing the asymptotic and non-asymptotic behavior of the errors in (1) and (2). One particularly relevant parallel can be established when an a priori probability of the target node presence at different locations is known. This is an example of a map-aware localization. Knowing the map means that one knows a priori the probability that a target node is present at each location. The problem of localization is equivalent to finding the best Bayes probability of error defined as:

\[
P^{(e)}_{\alpha}(u_1, u_2) = P(u_1) \alpha(u_1, u_2) + P(u_2) \beta(u_1, u_2).
\]

Interestingly, Neyman-Pearson test with \( \alpha = 0 \) is the optimal test. It can be proven that there is a map-aware fingerprinting algorithm such that \( \liminf_{n \to \infty} \frac{1}{n} \log P^{(e)}_{\alpha} = -I_0(0) \) where \( I_0(0) \) is called the Chernoff information of the probabilities \( P_{X|u_1} \) and \( P_{X|u_2} \) and is defined as:

\[
I_0(0) = -\inf_{t \in \mathbb{R}} \frac{1}{n} \log \mathbb{E} \left( \exp \left( t \log \frac{P_{X|u_2}}{P_{X|u_1}} \right) \right)
\]

and the expectation is with respect to \( P_{X|u_1} \). The proof is an exact replication of the equivalent hypothesis testing problem and follows from large deviation theory analysis, as discussed.
in [15]. The important insight is that map-aware localization can be done using $T_n$ with the threshold zero and the error decays exponentially with $n$ and with the Chernoff information $I_0(0)$. These are some examples of interesting insights that can be derived from the hypothesis testing analogy. So far, a two-point localization scenario has been considered. This can be easily extended to a localization scenario with finite candidate locations. Basically, the FPS can reliably and accurately localize a target node within finite possible locations using enough large number of measurements. When possible locations are uncountable, the localization space is divided into multiple regions and the equivalent HT problem aims at finding in which region the target node is located. These regions are indeed related to the training grid. This formulation enables us to reformulate the above results. Mainly, the FPS can reliably locate the target node at its respective regions with overwhelming probability and with enough large number of measurements. The maximum error is then dependent on the size of regions. Although FPS can successfully find the target node’s region, the number of required measurements for reliable localization increases when the sizes of regions decrease. Intuitively, finer localization comes with the price of higher latency. Another issue is the strong assumption that an algorithm knows the conditional probability distribution for the path loss. In this work, it is assumed to account for small changes in RSS values. The fingerprint at the point $u$ is then $X_u = (X_u^{(1)}, \ldots, X_u^{(R)})$. Based on the results of previous section, the KL-divergence $D(P_{X|u_1}||P_{X|u_2})$ is adopted as the main metric of interest. This metric is inversely related to the latency of localization. Bigger KL-divergence indicates fewer measurements required for localization. The metric $D(P_{X|u_1}||P_{X|u_2})$ is evaluated as

$$
\sum_j \frac{(P_j^{(j)})^2}{2N_j} \left( \frac{1}{\|w_j-u_1\|^2} - \frac{1}{\|w_j-u_2\|^2} \right).
$$

From this, some insights can be directly derived. First, increasing the number of anchors can improve KL-divergence and thereby reduce the number of required measurements for localization, i.e. which can be interpreted as latency. This is subject to proper placement of anchor points. Note that the worst performance are obtained at the points with smallest KL-divergence and therefore new anchors should be placed in such a way to increase the KL-divergence exactly for those points. To see this, define $\ell(u, e) = D(P_{X|u}||P_{X|u+e})$ which is latency of distinguishing two points of distance $\|e\|$. Figure 1 shows level curves of $\ell(u, e)$ for $e = (0.1, 0.1)$. It can be seen that the small value of $\ell(u, e)$, which indicates bad localization performance, corresponds to an oval surrounding the anchors and particularly to points between the anchors. This observation suggests that a new anchor should be placed on those curves containing the localization area. One method of finding the position of new anchors is to consider Voronoi regions of current anchors and place a new anchor on the intersection of Voronoi regions. In this way, KL-divergence is increased by installing nearby anchors to compensate the effect of far anchors. An analytical explanation of this phenomenon can be given by using mean value theorem for several variables:

$$
D(P_{X|u_1}||P_{X|u_2}) \leq \sum_j \alpha^2(P_j^{(j)})^2 \left( \frac{1}{\|u_1-u_2\|^2} \right)^{\alpha^2+2}.
$$

where $w_j$ is a location on the line between $u_1$ and $u_2$. For enough far points from all anchors, all $w_j$ can be approximated as equal and one can use the function $\ell(u) = \sum_j \left( \frac{1}{\|u-w_j\|^2} \right)^{\alpha^2+2}$ for evaluating the effect of anchor placements on localization performance. First, for enough small or even large path loss exponent $\alpha$, the distance between two fingerprints is arbitrarily small. Moreover, the level curves of $\ell(\cdot)$ approximate well that of $\ell(\cdot)$ for far points from anchors.
IV. SIMULATION AND EXPERIMENTAL RESULTS

In this section, numerical simulations are used to characterize quantitatively the findings of the proposed theory within a model close to our theoretical assumptions and yet more complex with a finite number of training locations and measurements per location. Our experimental setup is then employed to show that our theoretical guidelines still hold for a realistic environment for the same algorithm. As a fingerprint at each location, we select the vector of average RSS values observed from different WiFi APs, which is a well-known and often used fingerprint selection method [13]. The pattern matching function is the Euclidean distance between vectors, which is also a standard method [18].

A. Simulation Results

In a simulation environment, we define a set of APs’ related parameters, i.e. their locations and transmit powers. RSS values obtained from each AP at a targeted node at an unknown location are modeled using the COST 231 multi-wall model for indoor radio propagation [19]. The applicability of the model has been demonstrated for localization purposes [20] and the model has been extensively used (e.g. [21]). The model accounts for the type and number of walls, floors or obstacles in an environment, as well as for the locations of APs. The first attenuation contribution in the model is a well-known one-slope term that relates the received power to distance. Two parameters influence the attenuation in this term: the constant $l_0$ (the path-loss at 1 m distance and at the center frequency of 2.45 GHz) and the path-loss exponent $\alpha$. The second attenuation contribution is a linear wall attenuation term. The number of walls in the direct path between an AP and a target node is counted and for each wall an attenuation contribution is assumed. The model outputs RSS values from the defined APs at a target node’s location.

For the simulation and later experimental examination environment we selected the TWIST testbed [22]. The TWIST testbed environment is an office building, with its outline given in Figure 6. In the model parameterization, we used measurements from the TWIST testbed and leveraged a least square fitting procedure that allows minimizing the cost function between the measured received power and the modeled one. The parameters that are given to the model are the constant $l_0$ related to the least square fitting procedure, the path-loss exponent $\alpha$, and the wall attenuation factor $l_w$. Additionally, a zero-mean Gaussian noise with standard deviation $\sigma$ has been added to the obtained RSS values, which a standard procedure in simulation of RSS-based localization systems. For deriving our simulation results we used $l_0 = 53.73$, $\alpha = 1.64$, $l_w = 4.51$, $\sigma = 2$. The transmit power of each AP was set to 20 dBm. In our simulation, we defined a set of 4 AP with their locations indicated in Figure 6. A target’s node location has been selected randomly, its location has been estimated using the selected fingerprinting algorithm, and the localization error, i.e. an offset from the true location has been calculated. The procedure has been repeated 10000 times and the results have been reported in a regular box-plot fashion.

First, we evaluate the statement given in Theorem 3.1 concerning the number of observations at both training locations and in the runtime phase of fingerprinting. Therefore, we fixed the training grid to a hexagonal one of 105 training locations. Furthermore, we increase the number of observations from 1 to 9 with a step of 3 in both phases of fingerprinting. The results are depicted in Figure 2. As visible from the figure, in comparison to the basic case where only one measurement is taken in both training and runtime phase, an increase in the number of measurements significantly reduces the localization error. Furthermore, the reduction of errors is higher for the increase in the number of measurements in both phases, in contrast to increasing this number in only one of them. For example, the error is in average reduced by roughly 15% in case of the increase in the number of measurements from 1 to 9 in one phase only. For the same increase, but in both phases, the average error is reduced by more than 25%.

Second, we evaluate the statement given in previous section, concerning the number and locations of APs. We start from the basic scenario with 4 APs (AP1, AP2, AP3, AP4) depicted in Figure 3. We then introduce two additional APs (AP5 and AP6 in Figure 3) in the environment, where their locations are selected either randomly or based on Voronoi vertices. Voronoi vertices-based selection places new APs at locations that are the farthest from the locations of the existing ones. Furthermore, based on the locations of, at this point, 6 APs, we investigated whether the localization error is higher for the increase in the number of APs generally notably improves the performance of fingerprinting. Secondly, in comparison to 5 different random sections, both in case of 6 and 10 APs, Voronoi vertices-based selection generally yields more than 15% better performance. In other words, selection of APs in a way that they are the farthest from the locations of existing APs significantly outperforms other selections. In conclusion, the results derived by simulation with a higher level of realism are consistent with the developed theory.

B. Experimental Results

The TWIST testbed is specifically designed for the performance evaluation of indoor localization solutions. It features automated experimentation capabilities, accurate ground-truth positioning, minimization and monitoring of external influences (e.g. interference), and immediate calculation and storage of the performance results [23] along the well-established guidelines for evaluation of localization solutions [24]. A training database has been created by collecting 40 RSS values from four APs at 41 locations, as indicated in Figure 6a). The used evaluation locations are shown in Figure 6b).

We evaluate the statement that the main benefit of using multiple APs is a reduction of the far anchors effect. In other words, localization errors increase as the distance between an AP and a target node increases. Figure 5 presents spatial distributions of errors for three situations distinguished by the number of APs used for localization. As visible in Figure 5a,
those locations farther away from AP 1 tend to have larger errors. This is because the inverse relation of RSS values with distance such that those points closer to AP 1 have a finer RSS granularity. It is interesting to see that fingerprinting localization algorithm performs acceptable in indoor environment even with one anchor due to shadowing and multipath effects. Furthermore, the APs are added according to the guidelines discussed in the paper. In the first step, AP 2 is deployed at the farthest location from AP 1. It can be seen in Figure 5b that such placement mainly decreases the errors at locations close to AP 2. In Figure 5c, four APs are deployed at four corners of the testbed, which again improves the localization error. Localization errors for different combinations of APs are given in Figure 7. As visible from the figure, an increase in the number of APs significantly reduces the error, e.g. in average the is reduced by roughly 50% in case the number of APs increases from 2 to 4. The results derived in a realistic environment where the assumptions from the theoretical framework do not necessarily hold are in line with the developed theory.

REFERENCES