Random Fourier Feature Based Deep Learning for Wireless Communications

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

10-11-2021 / 12-11-2021

CITATION

Iltra, Rangeet; Kaddoum, Georges (2021): Random Fourier Feature Based Deep Learning for Wireless Communications. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16973650.v1>

DOI

[10.36227/techrxiv.16973650.v1](https://doi.org/10.36227/techrxiv.16973650.v1)

Random Fourier Feature Based Deep Learning for Wireless Communications

Rangeet Mitra and Georges Kaddoum

Abstract

Deep-learning (DL) has emerged as a powerful machine-learning technique for several problems encountered in generic wireless communications. Also, random Fourier Features (RFF) based DL has emerged as an attractive solution for several machine-learning problems; yet existing works lack rigorous analytical results to justify the viability of RFF based DL. To address this gap, we analytically quantify the viability of RFF based DL in this paper. Precisely, we present analytical proofs which show that the RFF based DL architectures have lower approximation-error and a lower probability of misclassification as compared to classical DL architectures for a fixed dataset-size. In addition, a new distribution-dependent RFF (DDRFF) is proposed to facilitate DL architectures with low training-complexity. The presented analytical contributions and the DDRFF are validated by relevant case-studies such as: a) line of sight (LOS)/non-line of sight (NLOS) classification, and b) message-passing based detection of low-density parity check (LDPC) codes over nonlinear visible light communication (VLC) channels. Especially, in the low training-data regime, the presented simulations depict significant performance gains for RFF based DL. Lastly, in all the presented simulations, it is observed that the proposed DDRFFs significantly outperform RFFs, which make them useful for potential machine-learning/DL applications for communication systems.

Index Terms

Random Fourier Features, Deep Learning, Wireless Communications, Distribution-dependent learning

I. INTRODUCTION

The capacity of classical machine learning methodologies are limited in terms of learning accurate representations from data and in generalizing over large datasets [1], [2]. On the other hand, deep-learning (DL) has emerged as a promising machine-learning paradigm to model a nonlinear/abstract

Rangeet Mitra and Georges Kaddoum are with the Resilient Machine learning Institute (ReMI) of the École de Technologie Supérieure (ÉTS), University of Quebec, Montreal, Canada. (Email: rangeet.mitra.1@ens.etsmtl.ca).

This work is under Major Revision in the IEEE Transactions on Cognitive Communications and Networking.

mapping from observations or learning representations from data. Furthermore, DL based algorithms have been successfully deployed in numerous sub-domains like computer-vision, speech processing, natural language processing, and time-series prediction. Recently, DL has also emerged useful for several applications in wireless communications [3], [4]. For various tasks/problems in these sub-domains, diverse DL-architectures have emerged, which include the multilayer perceptron, the convolutional neural network (CNN) for spatial data, and the recurrent neural network (RNN) for time-series data [5]–[7]. Especially for time-series data, long-short term memory (LSTM) has emerged as a promising DL architecture, as the LSTMs comprehensively address the issue of exploding/vanishing gradients [8], [9] encountered in the backpropagation algorithm for modelling/prediction. However, in spite of the widespread deployment of DL for complex machine-learning tasks, the existing deep neural networks (DNN) are known for their sensitivity to hyperparameters, such as, the number of layers, number of hidden nodes in each layer and the nature of the activation functions.

On the other hand, classical kernel based learning techniques are well-known for their capability of modeling high-dimensional representations, their generalization [10]–[12], and their hyperparameter-independence [13]; however, the classical kernel based methods require representation of the learning-parameter as an implicit inner-product in a reproducing kernel Hilbert space (RKHS) using Mercer kernels. Since the exact nature of the implicit feature-map is unknown, recent works approximate the feature map explicitly using sampling methods such as random Fourier features (RFF) [14], [15]. From the simulations presented in existing works, it is observed that the RFF based DL significantly enhances the performance of generic feedforward DNNs [16]. Furthermore, the RFFs lead to intrinsically regularized parameter updates which leads to improved generalization of [11], [17], [18] RFF based DNN architectures. This intuition has prompted the proposal of several RFF-DL methods [16], [19]. However, the existing works on RFF-DL motivate their results on intuition/simulation examples instead of rigorous analytical results to motivate their simulations/experiments on RFF based DL. Furthermore, RFFs require a large number of dimensions to gain an accurate approximation of an RKHS, which significantly increases the overall computational complexity, and creates a serious implementation bottleneck in the practical deployment of RFF based DL. Thus, with an aim to comprehensively address these issues with existing RFF based DL, we outline the following novelty points of this work:

- We quantify the viability of fixed kernel width based RFF-maps in the context of DNN rigorously. We present our claims in the form of two theorems and present analytical justifications for the benefits

of RFF-mapped observations for DL. For a fixed dataset-size, the derived theorems guarantee an improved likelihood of correct detection for time-series data for RFF-DL.

- To overcome the computational complexity incurred by RFF-mappings, a distribution dependent RFF (DDRFF) is proposed that outperforms classical RFF in scenarios with low/medium RFF dimensions, and delivers improved classification performance.

We validate the presented analytical guarantees for RFF based DL and the DDRFFs over the following case-studies: a) line of sight (LOS)/non-line of sight (NLOS) identification for wireless links using LSTM based DNN, and b) Message-passing based low-density parity check (LDPC) decoding over nonlinear visible light communication (VLC) channels. Next, we present a review of existing works the above two sub-domains.

A. LOS/NLOS classification for wireless links

Accurate inference of channel-state is critical for node-localization, and link-adaptation over ad-hoc tactical networks, which necessitates extracting accurate information of the channel-conditions and tracking the users' channels. However, in high mobility scenarios, inferring accurate information about the channel state is quite challenging, mainly due to the time-varying nature of the wireless channel, which significantly impairs localization/degrades the overall wireless link due to detrimental outages caused by NLOS scenarios [20]–[22]. Hence, it is essential to develop accurate signal processing algorithms that estimate and track the channel, and also infer the channel type, i.e. LOS vs NLOS such that suitable link-adaptation or network topology selection can be performed [23].

In this section, we focus on reviewing generic signal processing algorithms for LOS/NLOS detection. The existing DL based LOS/NLOS detection methods use receive signal strength indicator (RSSI) [24], [25] to train LSTM/hybrid CNN [26]. Besides, an LSTM trained with local temporal RSSI features is proposed for indoor localization in [27]. Furthermore, unsupervised approaches for LOS/NLOS identification include the use of a Gaussian mixture model for LOS/NLOS classification [28]. Also, the work in [21] suggests tracking V2V channels using IEEE 802.11p, and it is particularly highlighted that the NLOS components cause link-outages due to packet losses, and that accurate LOS/NLOS detection is needed to mitigate such losses a-priori by link-adaptation. Lastly, it should be noted that the task of LOS/NLOS detection is significantly non-trivial for outdoor scenarios due to typically longer ranges, high user-mobility and delay-spread, which necessitates fast and accurate LOS/NLOS identification. For

scenarios with user-mobility (that leads to time-series data based formulations), LSTMs are useful for LOS/NLOS prediction. In these scenarios, it is very much necessary to perform accurate LOS/NLOS labeling and tracking with a few data-points.

B. LDPC-decoding for static VLC channels

The LPDC codes are well-celebrated examples of capacity-achieving codes [29], have been widely deployed for both radio-frequency (RF) and VLC based communication systems [30], [31]. However, in the context of VLC, the the codewords experience distortion due to nonlinear transfer-characteristics of the light emitting diode (LED), which, if unmitigated, severely impairs message-passing based detection. In this work, the observed codewords from a nonlinear VLC channel are iteratively mapped to an approximate RKHS using RFF to mitigate the LED nonlinearity and to recover the codewords. In the derived RFF based message passing algorithm (MPA), the message-priors are iteratively updated based on a soft measure of RFF-approximation at each iteration. From the conducted simulations over nonlinear VLC channels, a significant gain in **bit-error rate (BER) performance** is observed for the proposd RFF based message-passing algorithm. Furthermore, a significant BER-performance gain is observed upon deployment of the proposed DDRFFs as compared to RFFs, which renders the proposed RFFs viable.

It may be noted that the presented case studies are used to illustrate the benefits promised by the RFF based maps in the context of DL. In details, the major contribution of this work, namely, the Theorem. 2 in Section. III are existence proofs to justify the utility of RFF maps for enhancing the performance of a weak DL, and this analytical guarantee is critical to motivate several existing works which use the RFF based DL based on intuition, rather than a strong analytical motivation. Notably Theorem. 2 is generic for all datasets (and not limited to the case studies considered in this paper).

II. OVERVIEW OF RFF BASED SIGNAL PROCESSING

In this section, we provide an overview of RFF based signal processing. Using an implicit feature map to RKHS (denoted by $\Phi : \mathbb{R}^n \rightarrow \mathcal{H}$), and invoking the Representer theorem [32], an arbitrary function $f(\cdot)$ may be represented as the following weighted combination:

$$f(\cdot) = \sum_{\forall j} \beta_j \kappa(\mathbf{x}_j, \cdot), \quad (1)$$

where $\mathbf{x}_j$ denotes the j^{th} observation, β_j **denotes** the approximation-weights, and $\kappa(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuous and shift-invariant Mercer kernel. Estimating the above representation of $f(\cdot)$ is compu-

tationally involved and requires expressing $f(\cdot)$ *only* in terms of Mercer kernels which prevents us from gaining intuitive insights (as opposed to insights provided by intermediate layers of DNN). To reap the benefits of RKHS based approaches (like regularization, generalization etc) by potentially deploying them as features in DNN, the Mercer kernel $\kappa(\cdot, \cdot)$ are well-approximated by RFF [33]. This approximation is motivated by the Bochner's theorem [34], which is restated as follows:

Theorem 1. *A continuous and shift-invariant kernel $\kappa(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is positive-definite iff it is the Fourier transform of a Borel measure $\rho(\cdot)$ on $\mathbb{R}^n$.*

Using Bochner's theorem, a positive-definite kernel can be expressed as

$$\kappa(\mathbf{a}, \mathbf{b}) = \int_{\omega} \exp(j\omega^T[\mathbf{a} - \mathbf{b}])\rho(\omega)d\omega. \quad (2)$$

Denoting $\hat{\Phi}_{\omega}(\mathbf{a}) = \exp(j\omega^T\mathbf{a})$, one can re-write $\kappa(\mathbf{a}, \mathbf{b})$ (where $\hat{\Phi}_{\omega} : \mathbb{R}^n \rightarrow \mathbb{R}^{n_G}$) as

$$\kappa(\mathbf{a}, \mathbf{b}) = \mathbb{E}_{\omega \sim \rho(\omega)}[\hat{\Phi}_{\omega}(\mathbf{a})\hat{\Phi}_{\omega}(\mathbf{b})]. \quad (3)$$

Further bounds on the error in kernel-approximation were derived in [35] using the RFF based approximation of feature-maps. In particular, for a real Gaussian kernel, an RFF (denoted here by $\hat{\Phi} : \mathbb{R}^n \rightarrow \mathbb{R}^{n_G}$) is obtained as

$$\hat{\Phi}(\mathbf{x}) = \sqrt{\frac{2}{n_G}} \begin{bmatrix} \cos(\omega_1^T \mathbf{x} + b_1) \\ \vdots \\ \cos(\omega_{n_G}^T \mathbf{x} + b_{n_G}) \end{bmatrix}, \quad (4)$$

where each $\{\omega_i\}_{i=1}^{n_G}$ is a Gaussian vector, with zero mean and covariance $\frac{1}{\sigma^2}\mathbb{I}_{n_G}$, with $\mathbb{I}_{n_G}$ denoting the identity matrix of size n_G , σ denotes the kernel-width, and $(\cdot)^T$ denotes the transpose operation.

It is noted that since an RKHS is closed, an exact representation exists for a wide class of functions in an RKHS [32], which makes RKHS based learning methods suitable for function-approximation. However, most RKHS techniques rely on learning a dictionary of observations [11], [36], [37], and hence are sensitive to inclusion of erroneous entries due to noise in the initial learning-stages, and also makes practical implementation complex. In this regard, RFFs provide an approximate explicit map to RKHS, which facilitates practical implementations with a finite memory budget whilst delivering equivalent performance as promised by RKHS based learning algorithms, which make them promising for RFF based DNN-

architectures.

III. PROOF OF VIABILITY OF RFF FOR TRAINING LSTM

In this section, an analytical guarantee is derived for the RFF based LSTMs in terms of classification accuracy/optimizing the hit-or-miss cost function. First of all, we enlist the considered assumptions as follows:

- We consider two kinds of sequences of observations: a) Independent and identically distributed (i.i.d) observations in $\mathbb{C}^n$ denoted by $\mathbf{s} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots)$ and b) Observations mapped to RKHS using RFF denoted by $\zeta = (\hat{\Phi}_1, \hat{\Phi}_2, \dots, \hat{\Phi}_i, \dots)$.
- We denote the linear inner-product space in $\mathbb{C}^n$ as $\mathcal{X}$. Furthermore, we denote the space spanned by Φ_1 as $\mathcal{H}$, which can be considered as an extension space of $\mathcal{X}$.
- c) The sequence of actual, and predicted labels are denoted by $\mathbf{g} = (g_1, g_2, \dots, g_i, \dots)$ and $\hat{\mathbf{g}} = (\hat{g}_1, \hat{g}_2, \dots, \hat{g}_i, \dots)$, respectively.
- Furthermore, we present analytical results for a learner with $\mathbf{s}$ and ζ as possible inputs and analytically compare the performance of this learner for each of the two cases.

Based on these assumptions, we proceed to formulate the following theorem¹:

Theorem 2. *For a given learner, the likelihood of correct detection of a sequence of labels given the mapped sequence ζ , is greater than the likelihood of the corresponding correct detection given $\mathbf{s}$ for some time-index $i > T$. Moreover, given B points in a dataset, n_G RFFs, and under an accurate kernel-width initialization, the RFF map improves the likelihood of correct detection by $\mathcal{O}\left(\frac{Bn_G}{2} \log \left[\frac{n_G}{2}\right]\right)$.*

Proof: We first assume the contrapositive, i.e.,

$$p(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s}) > p(\hat{\mathbf{g}} = \mathbf{g}|\zeta). \quad (5)$$

In other words,

$$\frac{p(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})}{p(\hat{\mathbf{g}} = \mathbf{g}|\zeta)} > 1. \quad (6)$$

¹It may be noted that, we can extend the results of this theorem to any weak DL, instead LSTMs alone. This “weakness” in the weak DL can be either due to an inexact architecture choice, or due to poor generalization due to small dataset-sizes.

This, in-turn, implies the following:

$$|\mathbf{J}| > 1, \quad (7)$$

where $\mathbf{J}$ denotes the Jacobian from $\mathbf{s} \rightarrow \boldsymbol{\zeta}$. However, from [18, eq. (16)], it is noted that $|\mathbf{J}| < 1$, which contradicts (5), and thus concludes the proof for the performance-gain of RFF-DL.

For the second part, we note from [18, eq. (16)] that an equivalent expression for the fixed-kernel width based RFF-DL considered in this paper is given by:

$$-\infty < \log(|\mathbf{J}|^B) < -\frac{Bn_G}{2} \log \left[\frac{n_G}{2} \right] + B \log[n_G h^2], \quad (8)$$

which can be refactored as follows:

$$-\infty < \log(|\mathbf{J}|^B) < -\frac{Bn_G}{2} \log \left[\frac{n_G}{2} \right] + B \log[n_G(h^{o^2} + \Delta h^2 + 2h^o \Delta h)], \quad (9)$$

where Δh denotes the deviation in the value of h from its optimal value h^o . For B data-points and under a reasonably accurate kernel-width initialization that typically leads to a small error in the kernel-width, we conclude that the likelihood of correct detection in (9) improves by $\mathcal{O}\left(\frac{Bn_G}{2} \log \left[\frac{n_G}{2} \right]\right)$. ■

Using Theorem. 2, we outline a Corollary below on the generic robustness of RFF-DL to errors in the value of the kernel-width, Δh . We prove below the robustness of RFF-DL over a large range of kernel-width Δh .

Corollary. 1 *The RFF map offers improved robustness to error in the kernel-width Δh . In details, the allowable error in this kernel-width is bounded as follows:*

$$0 < \Delta h < \frac{n_G^{\frac{\frac{n_G}{2}-1}{2}}}{2^{\frac{n_G}{4}}} \quad (10)$$

Proof: From (9) in Theorem. 1, we require the following condition in order to warrant the achievement of a positive gain by the RFF map:

$$-\frac{Bn_G}{2} \log \left[\frac{n_G}{2} \right] + B \log[n_G(h^o + \Delta h)^2] < 0. \quad (11)$$

Rearranging, the above equation is rewritten as follows:

$$B \log \left[\frac{2^{\frac{n_G}{2}} (h^o + \Delta h)^2}{n_G^{\frac{n_G}{2}-1}} \right] < 0. \quad (12)$$

To ensure the above condition, the following needs to be satisfied:

$$\frac{2^{\frac{n_G}{2}} (h^o + \Delta h)^2}{n_G^{\frac{n_G}{2}-1}} < 1, \quad (13)$$

which leads to the following condition on Δh :

$$0 < \left| \frac{\Delta h}{h^o} \right| < \frac{n_G^{\frac{n_G}{4}-\frac{1}{2}}}{2^{\frac{n_G}{4}} h^o} - 1, \quad (14)$$

It is noted that for large n_G , this allowable upper bound grows as $\mathcal{O}(n_G^{\frac{n_G}{4}})$, which indicates the robustness of the RFF map to a large range of relative error in the value of the kernel-width². ■

Next, we present the following theorem on the computational complexity of RFF-DL and compare it with classical DL.

Theorem 3. *Consider a training algorithm which operates over a dataset with batch-size B . Further, let the per-batch computational complexity of this training algorithm be an arbitrary function $r_{DL}(n, S, B, H, G)$ for the generic DL and $r_{RFF-DL}(n_G, S, B, H, G)$ for the RFF-DL, where H denotes the number of hidden-nodes and G denotes the number of layers. Then there always exists an $n_G > n$ for which the RFF-DL converges faster than generic DL.*

Proof: We assume that the RFF-DL based approach requires T batches for convergence, with the per-batch computational complexity of the generic DL given by $r_{DL}(n, S, B, H, G)$ and the RFF-DL being given as $r_{RFF-DL}(n_G, S, B, H, G)$. It is inferred from Theorem. 2 that after observing T batches of batch-size B , the RFF-DL improves the likelihood of correct detection at least by the following factor:

$$\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\zeta)] - \log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})] = \frac{TBn_G}{2} \log \left[\frac{n_G}{2} \right] - TB \log[n_G(h^o + \Delta h)^2] + C, \quad (15)$$

where we denote $\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\zeta)]$ assuming the sequence ζ consisting of T vectors for $C > 0$ and $\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\zeta)] \approx \log[p_{T'}(\hat{\mathbf{g}} = \mathbf{g}|\zeta)]$. Assuming both $\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\zeta)]$ and $\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})]$ to lie within an ϵ -neighborhood at $T' > T$, i.e., $|\log[p_{T'}(\hat{\mathbf{g}} = \mathbf{g}|\zeta)] - \log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})]| \leq \epsilon$, we have the

²In our simulations/case-studies in this paper and for other problems internally investigated, we have varied this kernel-width over few decades above/below its ideal value without experiencing any performance-degradation.

following inequality:

$$\left| \frac{T' B n_G}{2} \log \left[\frac{n_G}{2} \right] - T' B \log[n_G(h^o + \Delta h)^2] \right| < \epsilon, \quad (16)$$

which implies the following:

$$B T' \log \left[\frac{n_G^{\frac{n_G}{2}-1}}{2^{\frac{n_G}{2}} (h^o + \Delta h)^2} \right] < \epsilon, \quad (17)$$

Further, assuming convergence of both LLRs to similar values (approximately zero in an ideal case) after T' batches, and further assuming that $\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\zeta)]$ allows for the re-expression of (15) as follows:

$$\log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})] + \frac{(T' - T) B n_G}{2} \log \left[\frac{n_G}{2} \right] - B(T' - T) \log[n_G(h^o + \Delta h)^2] + C = \log[p_{T'}(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})], \quad (18)$$

which leads to the following:

$$(T' - T) = \frac{\log[p_{T'}(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})] - \log[p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})] - C}{\frac{B n_G}{2} \log \left[\frac{n_G}{2} \right] - B \log[n_G(h^o + \Delta h)^2]} \quad (19)$$

or

$$(T' - T) = \frac{\log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right] - C}{B \log \left[\frac{n_G^{\frac{n_G}{2}-1}}{2^{\frac{n_G}{2}} (h^o + \Delta h)^2} \right]}, \quad (20)$$

From (17), the above inequality is re-expressed as follows:

$$T' - T = \frac{T' \left\{ \log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right] - C \right\}}{\epsilon}, \quad (21)$$

which quantifies the gap between T' and T for corresponding values of $p_{T'}(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})$ and $p_T(\hat{\mathbf{g}} = \mathbf{g}|\mathbf{s})$. Thus, for a similar performance-level, the difference between their respective computational complexities after observing T' batches is bounded by: Thus, for a similar performance-level, the difference between their respective computational complexities after observing T' batches is given by:

$$\Delta r = r_{DL}(n, S, B, H, G) B T' - r_{RFF-DL}(n, S, B, H, G) B T, \quad (22)$$

which can be alternatively re-expressed as follows:

$$\Delta r = B(T' - T) r_{DL}(n, S, B, H, G) + B T (r_{DL}(n, S, B, H, G) - r_{RFF-DL}(n_G, S, B, H, G)). \quad (23)$$

From (21), it is concluded that:

$$\Delta r = \frac{BT' \left\{ \log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right] - C \right\}}{\epsilon} + BT(r_{DL}(n, S, B, H, G) - r_{RFF-DL}(n_G, S, B, H, G)). \quad (24)$$

Therefore, to achieve an ϵ -similar likelihood of correct detection after T' iterations within the same computational budget, n_G can be chosen to satisfy the following for a desired ϵ at convergence and for a given gain in LLR of³ $\log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right]$, we put $\Delta r = 0$:

$$\frac{T'}{T} \left\{ \log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right] - C \right\} = r_{RFF-DL}(n_G, S, B, H, G) - r_{DL}(n, S, B, H, G). \quad (25)$$

Noting as per assumptions that $T' > T$, $\log \left[\frac{p_{T'}(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})}{p_T(\hat{\mathbf{g}}=\mathbf{g}|\mathbf{s})} \right] - C > 0$, and infinitesimal ϵ makes the term in the L.H.S of (25) large, and allows for the existence of $n_G \gg n$ for a gain in convergence-speed of $\Delta T = T' - T$ within the same computational budget. In fact, from (19), it is possible to bound the allowable increase in the per-batch complexity as $\mathcal{O} \left(\frac{BT' \Delta T}{\epsilon T} \log \left[\frac{n_G^{\frac{n_G}{2}-1}}{2^{\frac{n_G}{2}} (h^o + \Delta h)^2} \right] \right)$. ■

IV. PROPOSED DDRFF

The RFF based DNNs require a large number of RFFs to gain an accurate approximation of an RKHS. In this context, a DDRFF is proposed in this section, which achieves lower approximation error as compared to classical RFFs for a given number of RFF dimensions, and hence significantly lowers the computational complexity required for achieving a given error floor.

Indexing the incoming observations by j , we can update the following Parzen estimate of the p.d.f of observations, denoted by $\hat{p}(\mathbf{x})$.

$$\hat{p}(\mathbf{x}) := \left(\frac{j-1}{j} \right) \hat{p}(\mathbf{x}) + \frac{1}{j} \mathcal{K}_\lambda(\mathbf{x} - \mathbf{x}_j), \quad (26)$$

where the spread parameter for kernel density estimation, λ , is drawn from Silverman's rule [38]. Consequently, the mean of the RFF can be readily derived, using a moving-average estimator as follows:

$$\boldsymbol{\mu}_{\hat{\Phi}} := \nu \boldsymbol{\mu}_{\hat{\Phi}} + \underbrace{\int_{\mathbf{x}} \hat{\Phi}(\mathbf{x}) \hat{p}(\mathbf{x})}_{\mathcal{S}}, \quad (27)$$

³Note that the values of $r_{RFF-DL}(n_G, S, B, H, G)$ and $r_{DL}(n, S, B, H, G)$ specific to the type of the DL and the kind of optimizer used (such as Adam, gradient-descent with/without momentum etc.)

where $\mu_{\hat{\Phi}}$ denotes the mean of the RFF, estimated by the moving-average estimator, and $\nu \in [0, 1]$ is the forgetting factor. However, from (26) one can adapt $\mathcal{S}$ as

$$\mathcal{S} := \frac{i-1}{i}\mathcal{S} + \frac{1}{i} \frac{1}{M} \int_{\mathbf{x}} \sum_{j=1}^M \mathcal{K}_{\lambda}(\mathbf{x} - \mathbf{x}_j) \hat{\Phi}(\mathbf{x}) d\mathbf{x}. \quad (28)$$

Denoting $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^{n_G}$ as a vector, we get:

$$\begin{aligned} \mathcal{S}_i := & \frac{i-1}{i}\mathcal{S}_i + \exp\left(-\frac{\lambda^2 \boldsymbol{\omega}_i^T \boldsymbol{\omega}_i}{2}\right) \times \\ & \frac{1}{i} \frac{\sqrt{2\pi\lambda^2}}{M} \sum_{j=1}^M \cos(\boldsymbol{\omega}_i^T \mathbf{x}_j + b_{n_G}), \end{aligned} \quad (29)$$

where M is the size of the considered batch. This gives us a new smoothed RFF that can present potentially useful features for low-complexity DNN architectures.

V. PROOF OF VIABILITY OF DDRFFS

In Section. III, we demonstrated the viability of RFF based LSTMs in terms of the mis-classification error/the ‘‘hit-or-miss’’ cost function. In this section, we prove that, compared to classical RFFs, the proposed DDRFFs provide a better approximation to the RKHS by reducing the approximation-error (which is a ‘‘soft’’ error-metric). We state this claim in the form of the following two theorems.

Theorem 4. *The mapping given by the DDRFF map is closed under the same RKHS $\mathcal{H}$ as the classical RFF. Furthermore, DDRFFs formulated in (29) deliver an improved approximation of RKHS.*

Proof: We begin by recalling that for a fixed number of n_G dimensions, the approximation error ϵ for classical RFF is given by $n_G \sim O(\epsilon^{-2} \log \epsilon^{-2})$ [39]. However, we also note that ϵ is not only a function of dimension n_G , but also a function of the set of $\boldsymbol{\omega}_i$. In other words, the approximation error depends on how much the samples of $\boldsymbol{\omega}_i$ deviate from a Gaussian distribution; particularly, when n_G is not high enough to converge to the desired Gaussian distribution. Hence, in the sequel, we denote the approximation error as $\epsilon_{n_G}(\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ denotes an approximate continuum of $\{\boldsymbol{\omega}_i\}_{i=1}^{n_G}$. This assumption of an approximate continuum enables the re-expression of the summation in (29) as an integral. Using this, the energy of $\epsilon_{n_G}(\boldsymbol{\omega})$ is expressed as follows:

$$\epsilon_{n_G}^{(1)} = \int_{-\infty}^{\infty} |\epsilon_{n_G}(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \quad (30)$$

From (29), we can approximate each component of the proposed RFF as follows:

$$\frac{\sqrt{2\pi\lambda^2}}{M} \sum_{j=1}^M \int_{\omega} \exp\left(-\frac{\lambda^2\|\omega\|^2}{2}\right) \cos(\omega^T \mathbf{x}_j + b) d\omega. \quad (31)$$

Noting the fact that the RFF can be expressed as the sum of an element in RKHS $\mathcal{H}$ with some error ϵ_{n_G} , the above equation can be re-expressed as

$$\frac{\sqrt{2\pi\lambda^2}}{M} \sum_{j=1}^M \int_{\omega} \exp\left(-\frac{\lambda^2\|\omega\|^2}{2}\right) [\mathcal{H}_j + \epsilon_{n_{G_j}}(\omega)] d\omega, \quad (32)$$

which can be simplified as follows:

$$\begin{aligned} & 2\pi\lambda^2\mathcal{H}_j + \\ & \underbrace{\frac{\sqrt{2\pi\lambda^2}}{M} \sum_{j=1}^M \int_{\omega} \exp\left(-\frac{\lambda^2\|\omega\|^2}{2}\right) \epsilon_{n_{G_j}}(\omega) d\omega}_{\epsilon_{n_G}^{(2)}}. \end{aligned} \quad (33)$$

From the above equation, we make the following observations:

- Applying Parseval's theorem, one can easily note that the energy of $\epsilon_{n_G}^{(1)}$ is greater than $\epsilon_{n_G}^{(2)}$, as the term $\exp\left(-\frac{\lambda^2\|\omega\|^2}{2}\right)$ performs a low-pass filtering, and attenuates higher-magnitude "frequencies" (or ω).
- It is noted that the desired component, which is apart from $\epsilon_{n_G}(\omega)$, lies in the RKHS $\mathcal{H}$ (though it is scaled by $2\pi\lambda^2$. To arrive at this result, the closure property of $\mathcal{H}$ over a scaling operation is invoked for the proposed DDRFF).

■

Theorem 5. *For the same error floor ϵ , the number of RFFs required by the DDRFFs is lower by a factor $\mathcal{O}\left(\left(\frac{\lambda^2}{2\pi}\right)^n\right)$.*

Proof: It is noted that upon using the DDRFFs, and upon invoking the Cauchy-Schwarz inequality, the number of RFF dimensions required for an error floor $\mathcal{K}\epsilon$, is written as follows:

$$n_G^1 = \mathcal{C}\mathcal{K}^{-2}\epsilon^{-2} \log(\mathcal{K}^{-2}\epsilon^{-2}), \quad (34)$$

where $\mathcal{C}$ is an arbitrary constant, and $\mathcal{K} = \left[\frac{\lambda^2}{2\pi}\right]^{\frac{n}{2}} < 1$ is a scaling constant making the error floors of RFF and DDRFF same to facilitate comparison. Similarly, for DDRFFs, the dimensions n_G is quantified

as follows:

$$n_G^2 = \mathcal{C}\epsilon^{-2} \log(\epsilon^{-2}). \quad (35)$$

Therefore, this ratio is re-expressed as:

$$\frac{n_G^1}{n_G^2} = \frac{\lambda^{2n}}{(2\pi)^n} \left[1 + \frac{\epsilon^2 n_G n}{4} \log\left(\frac{\lambda^2}{2\pi}\right) \right], \quad (36)$$

Noting that $\epsilon^2 \approx 0$ for sufficiently large RFF dimensions, we write the following:

$$\frac{n_G^1}{n_G^2} \approx \frac{\lambda^{2n}}{(2\pi)^n}. \quad (37)$$

From the above, it is noted that $\lambda < 1$, which leads to the desired result. ■

VI. ARCHITECTURE OF RFF BASED LSTM

In this section, we present the architecture of a typical RFF based LSTM based DL. For illustration, the proposed architecture is shown in Fig. 1. As detailed in Theorem 2, the input is mapped to an approximate RKHS using RFFs as outlined in (4) or (29). Next, there are optional fully connected layers with a subsequent RFF transformation. The cascade of consecutive mapping renders the overall mapping to a sequence of RKHSs and the result in Theorem 2, can be readily utilized by replacing $\mathcal{H}$ with the last RKHS. Lastly, the observations in RKHS are presented as an input to the LSTM layer for prediction of the NLOS/LOS labels. Notably, the benefits of the proposed mapping prior to presentation to the LSTM layer have been highlighted in Theorem 2, which indicates that the posterior is more “peaked” given RFFs as input to the LSTM, as compared to the indigenous observations. These steps can be summarized in the following set of equations:

$$\begin{aligned} \mathbf{f}_t &= \sigma_g(\mathcal{B}^{(1)}\hat{\Phi}(\mathbf{x}_t) + \mathcal{B}^{(2)}\gamma_{t-1} + \mathbf{b}_f) \\ \mathbf{i}_t &= \sigma_g(\mathcal{A}^{(1)}\hat{\Phi}(\mathbf{x}_t) + \mathcal{A}^{(2)}\gamma_{t-1} + \mathbf{b}_i) \\ \mathbf{o}_t &= \sigma_g(\mathcal{D}^{(1)}\hat{\Phi}(\mathbf{x}_t) + \mathcal{D}^{(2)}\gamma_{t-1} + \mathbf{b}_o) \\ \chi_t &= \chi_{t-1} \circ \mathbf{f}_t + \mathbf{i}_t \circ \sigma_c(\mathcal{F}^{(1)}\hat{\Phi}(\mathbf{x}_t) + \dots \\ &\quad \mathcal{F}^{(2)}\mathbf{h}_{t-1} + \mathbf{b}) \\ \gamma_t &= \mathbf{o}_t \circ \sigma_h(\chi_t), \end{aligned} \quad (38)$$

Fig. 1. Depiction of LSTM based prediction using RFF.

where $\circ$ denotes the Hadamard product, $\sigma_g(\cdot)$ is a sigmoid activation function, $\sigma_c(\cdot)$ and $\sigma_h(\cdot)$ denote the $\tanh(\cdot)$ activation function, and $\mathbf{f}_t, \mathbf{i}_t, \mathbf{o}_t, \mathbf{x}_t, \gamma_t \in \mathbb{R}^{n_h}$ where n_h denotes number of hidden nodes. Furthermore, $\mathbf{f}_t$ denotes the forget-gate of LSTM with weights $\mathcal{B}^{(1)} \in \mathbb{R}^{n_h \times n_G}, \mathcal{B}^{(2)} \in \mathbb{R}^{n_h \times n_h}$ and bias $\mathbf{b}_f \in \mathbb{R}^{n_h}$. $\mathbf{i}_t$ denotes the input gate with weights $\mathcal{A}^{(1)} \in \mathbb{R}^{n_h \times n_G}, \mathcal{A}^{(2)} \in \mathbb{R}^{n_h \times n_h}$ and bias $\mathbf{b}_i \in \mathbb{R}^{n_h}$. $\mathbf{o}_t$ denotes the output gate with weights $\mathcal{D}^{(1)} \in \mathbb{R}^{n_h \times n_G}, \mathcal{D}^{(2)} \in \mathbb{R}^{n_h \times n_h}$ and bias $\mathbf{b}_o \in \mathbb{R}^{n_h}$. $\mathbf{x}_t$ is the sum of the gating of the previous parameter-value with the forget-gate and the gating between input nodes and

input-layer (parameterized by weights $\mathcal{F}^{(1)} \in \mathbb{R}^{n_h \times n_G}$, $\mathcal{F}^{(2)} \in \mathbb{R}^{n_h \times n_h}$ and bias $\mathbf{b} \in \mathbb{R}^{n_h}$), which is passed through the activation function $\sigma_c(\cdot)$ and gated with the output gate $\mathbf{o}_t$ to obtain the next-state $\gamma_t \in \mathbb{R}^{n_h}$. The aforementioned weights and biases are optimized using the backpropagation algorithm; however, we do not encounter the vanishing/exploding gradient problem for time-series prediction. In addition, the mapping to an approximate RKHS $\hat{\Phi}(\cdot)$ (as derived in (38)), prior to input to LSTM increases the accuracy of predictions as inferred from Theorem. 2.

In the next section, we present simulations, which validate the viability of the proposed RFF based LSTM for two problems. First, LOS/NLOS classification over an outdoor WINNER II channel. Next, the suitability of the proposed DDRFF for LDPC decoding is demonstrated by realistic simulations over nonlinear VLC channels.

VII. SIMULATIONS

In this section, we present simulations to validate the paradigm of RFF based learning for LOS/NLOS based classification and message-passing based detection for LDPC decoding. From the simulations presented below, one can observe significant gains for various nonlinear classification problems when using RFF-approximations of an RKHS, compared to using the indigenous observations.

A. RFF-based LSTM for LOS/NLOS classification

In this subsection, simulations are presented for LOS/NLOS identification in outdoor communication systems. The simulation parameters are summarized in Table. I. We consider various outdoor WINNER II channel scenarios, wherein there is a single base-station, and the receiver moves from an initial location using a random-walk mobility model. The antenna height at the transmitter is assumed to be 4m, while the mobile stations were assumed to move along a trajectory in the horizontal plane drawn from a 2D random-walk mobility model. The OFDM standard assumed is IEEE 802.11ax with a guard band of 3.2 μ s. The complex channel-estimates at the receiver are transformed by concatenating real and imaginary components prior to RFF mapping.

In this [section](#), we present the following two comparison-cases: a) the case in which we present $\mathbf{x}$ directly to the LSTM layer and b) the case in which $\mathbf{x}$ is mapped to an RKHS using RFF layer(s) prior to the LSTM layer. The candidate neural networks were trained with the initial location of the receiver as mentioned in Scenario I of Table I⁴. It is worthwhile to mention that the initial location of the user for

⁴From our simulations, we also note that the proposed approach works for an arbitrary initial point; however, we show some "extreme" skewed cases for brevity.

Scenario I was chosen heuristically such that there are almost equal number of examples of LOS and NLOS observations, as the user moves along the aforementioned random-walk based trajectory. Consequent to training, the candidate neural networks were tested on the testing-observations derived from Scenario II (which has more LOS labels due to the initial location being near to the base station), and Scenario III (which has more NLOS labels due to the initial location being far away from the base station). The testing F_1 -score is plotted for the C1 and D1 outdoor scenarios. From Figs. 2 3, it is observed that the gains in F_1 -score performance are more prominent in the low training-data regime, which makes DDRFFs better suited for high-mobility outdoor scenarios with a small coherence-time. For example, it is observed from Figs. 2 and 3 for the C1 and the D1 scenarios respectively show that the RFF based LSTM achieve an F_1 -score of 0.88 with only 400 training data-points, while the corresponding LSTM based approach without RFF could only achieve a similar value of F_1 -score after observing 3000 data-points.

Also, from the simulated receiver operating characteristics (ROC) plotted in Fig. 4 and Fig. 5 for 400 training samples, we observe an improved performance from the RFF based LSTM than the generic LSTM. Notably, this improvement in ROC performance very much aligns with the gains promised by the analysis derived in the previous sections. Also, from Figs. 2 and 3 for the F_1 -score performance, and Figs. 4 and 5 for the ROC comparisons, it is observed that the DDRFF-LSTM slightly outperforms the RFF-LSTM.

For the proposed method, comparisons are presented with respect to the ground truth in Table. II and Table. III and the converged F_1 score is reported for 400 data-points and 3000 data-points. From Tables. II and III, it is observed that the proposed RFF based LSTMs deliver improved F_1 scores with 400 training data-points as compared to generic LSTMs. However, when all the compared classifiers are shown for 3000 data-points, it is observed that both the generic LSTM and the RFF based LSTMs achieve similar F_1 scores. The performance-gains observed in the low-data regime for the fixed kernel width based RFF-DL are certainly aligned with the analytical result derived in Theorem. 2. This fast convergence is particularly useful for scenarios with high user-mobility and fast fading.

TABLE I
SIMULATION PARAMETERS

Outdoor channel model	Winner-II
Mobility model	Random walk
Guard Band	3.2 μ s
OFDM Standard	IEEE 802.11ax
Antenna Height at Tx	4m
Coordinates of base-station	(50,150)
(Training): Initial Rx coordinates C2	(200,120)
(Training): Initial Rx coordinates C1	(60,200)
(Testing): Initial Rx coordinates C2	(300,250),(60,200)
(Testing): Initial Rx coordinates C1	(300,250),(450,500)
[Urban scenario]: LOS/NLOS probability for C1	[40, eq. (1)],[41, Table. 4.7]
[Suburban scenario]: LOS/NLOS probability for C2	[40, eq. (2)],[41, Table. 4.7]
Number of RFF layers	1
$\mathbf{x}$	Sequence of Channel-estimates
RFF dimensions for input layer	200
Number of LSTM layers	1
Number of Hidden nodes for LSTM layer	50
Considered WINNER II Scenarios (Suburban)	C1, D1
Test-data sequence length	6000
Bandwidth	20 MHz

TABLE II
COMPARISON OF CONVERGED F1 SCORE FOR THE SUBURBAN SCENARIO C1

Algorithm Name	Testing F1 Score with 400 data-points	Testing F1 Score with 3000 data-points
LSTM	0.8034	0.8907
RFF-LSTM	0.8802	0.8968
DDRFF-LSTM	0.8802	0.8968

TABLE III
COMPARISON OF CONVERGED F1 SCORE FOR THE RURAL MACROCELL SCENARIO D1

Algorithm Name	Testing F1 Score with 400 data-points	Testing F1 Score with 3000 data-points
LSTM	0.8031	0.8916
RFF-LSTM	0.8836	0.9006
DDRFF-LSTM	0.8836	0.9006

B. Case-study II: LDPC decoding for VLC

In this subsection, we describe our case-study for LDPC decoding in a nonlinear VLC channel. We assume LOS VLC channel modelled by a Lambertian radiation pattern [42], [43], with a memory Rapp LED nonlinearity (which is widely used for modelling a white LED [44], [45]). [The overall system model](#)

for an LOS VLC channel at the i^{th} time-instant can be written as follows:

$$\mathbf{y}_i = h_i f(\mathbf{x}_i + \alpha \mathbf{x}_{i-1}) + \mathbf{n}, \quad (39)$$

where $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2 \mathbb{I})$, with σ_n^2 denoting the overall variance of the additive noise which accounts for the overall effect of shot-noise and ambient noise at the photodetector. The static VLC channel-gain h_i is given as:

$$h_i = \begin{cases} \frac{(m+1)Ag \cos(\theta) \cos^m(\phi)}{2\pi d^2} & , 0 < \theta < \Theta \\ 0 & , \text{otherwise} \end{cases}$$

where A indicates the effective area of the photodetector (PD), θ is the angle of incidence, ϕ is the incident angle at n^{th} detector, Θ indicates the field of view (FOV) angle of the receiver, d is the horizontal separation between the LED and the PD, and m denotes the Lambertian order which is related to the LED semi-angle at half power, $\phi_{\frac{1}{2}}$, as follows:

$$m = \frac{-\ln(2)}{\ln(\cos \phi_{\frac{1}{2}})},$$

g is the gain of optical concentrator given by:

$$g = \frac{n_{ref}^2}{\sin^2(\Theta)},$$

where n_{ref} is the refractive index of gain of optical concentrator. Moreover, x_i denotes encoded independent and identically distributed (i.i.d) on-off keying (OOK) transmissions, and $f(\cdot)$ denotes the LED given as follows:

$$f(x) = \begin{cases} \frac{x}{\left(1 + \left|\frac{x}{x_{sat}}\right|^{2p}\right)^{\frac{1}{2p}}}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad (40)$$

where x_{sat} denotes the LED's saturation current and p is a parameter that controls the severity of the LED nonlinearity. The message is encoded according to a 802.11n LDPC generator-matrix. From (39), it is observed that the nonlinearity $f(\cdot)$ warps/distorts the transmitted codewords $\mathbf{x}$, alters their algebraic structure, and causes errors in LDPC decoding. Hence, based on the previous discussion, it is proposed to learn the bits based on a detector trained on $\hat{\Phi}(\mathbf{y})$, where the RFF dimensions are equal to the codeword-length⁵. In fact, using the Representer theorem, (39) can be re-written as

$$\mathbf{y} = \langle k_x, \mathbf{x} \rangle_{\mathcal{H}} + \mathbf{n}, \quad (41)$$

⁵It is also possible to up-convert to higher dimensions using an RFF and then down-convert using an autoencoder. Though this dual-conversion may have performance benefits, it is computationally complex, and hence we focus our attention on the single-layer case.

where k_x is an operator in RKHS $\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes an inner-product in RKHS $\mathcal{H}$. The fact that the RKHS, $\mathcal{H}$, is complete, guarantees the existence of an operator k_y such that

$$\langle k_y, \mathbf{y} \rangle_{\mathcal{H}} \approx \mathbf{x} + \langle k_y, \mathbf{n} \rangle_{\mathcal{H}}. \quad (42)$$

It is noted that the above system equation allows itself to for statistical characterization as an AWGN channel with noise-variance equal to

$$\text{var}[\langle k_y, \mathbf{y} \rangle_{\mathcal{H}}] = \langle k_y, k_y \rangle_{\mathcal{H}} \sigma_n^2. \quad (43)$$

Given the theme of this work, the following approximation of RKHS is utilized

$$\langle k_y, \mathbf{y} \rangle_{\mathcal{H}} \approx \mathbf{\Omega}^T \hat{\Phi}(\mathbf{y}). \quad (44)$$

Next, a hypothesis, denoted by $\mathbf{\Omega}$, is trained on $\hat{\Phi}(\mathbf{y})$, which optimizes the following quadratic loss-function:

$$\|\mathbf{\Omega}^T \hat{\Phi}(\mathbf{y}) - \text{sign}(\mathbf{y})\|_2^2 \quad (45)$$

The Channel-decoding is performed using message-passing over a Tanner graph representation of the parity-check matrix [46]. We denote the graph-neighborhood of node k as $\mathcal{B}_k$ (which is the set of points incident on node k in the Tanner-graph apart from k itself). Additionally, for the j^{th} bit, the log-likelihood-messages from the bit-nodes to codewords, and the message from codewords to bit-nodes are denoted as $m_b(j)$ and $m_c(j)$ respectively. Lastly, we denote the length of the bit-string $\mathbf{b}$ as B and the size of the encoded codeword as C . The proposed RFF based detector is summarized in Algorithm 1. [Especially, Theorem. 2 of the current manuscript presents a strong and generic motivation for the achievement of higher LLR values upon deploying RFF based MPA in comparison to generic MPA.](#) A typical LOS VLC scenario is considered as in [45], with Rapp LED nonlinearity, where the memory-parameter of the nonlinearity, α , is considered to be 0.2, and the saturation current of the LED is 0.4. The channel-parameters considered for this case-study are summarized in Table. IV.

To validate the achievement of improved LLR values claimed in Theorem. 2, we now present a case study for MPA based channel-decoding. The generator matrix for the LDPC code is taken from the IEEE 802.11n standard with a block length of 648 [47], [48]. [The simulation results at 4 iterations for \$p = 0.5\$, \$p = 1\$, and \$p = 2\$ are depicted in Figs. 6, 7, and 8 respectively. The kernel width is tuned for LED nonlinearity with \$p = 0.5\$ and the same value is used for \$p = 1, 2\$. As a result of this non-ideal](#)

Algorithm 1 Message-Passing in RKHS

```

1: Initialize MAXITER.
2: for  $k = 1, 2, \dots, C$  do
3:   Initialize LLR:  $m_c(k) = \frac{-2y(k)}{\sigma_n^2}$ .
4: end for
5: while cnt < MAXITER do
6:   repeat
7:     for  $j = 1, 2, \dots, B$  do
8:        $m_b(j) = \log \left[ \frac{1 + \prod_{k \in \mathcal{B}_j} \tanh(m_c(k))}{1 - \prod_{k \in \mathcal{B}_j} \tanh(m_c(k))} \right]$ 
9:     end for
10:    for  $j = 1, 2, \dots, B$  do
11:       $m'_b(j) = \sum_{j \cup j \in \mathcal{B}_j} m_c(j) - \frac{2y(k)}{\sigma_n^2}$ 
12:       $\hat{b}(j) = \text{sign}(m'_b(j))$ .
13:    end for
14:    Calculate syndrome  $e = \mathbf{H}\hat{\mathbf{b}}$ .
15:    for  $j = 1, 2, \dots, C$  do
16:      Update:  $m_b(j) = \sum_{j \in \mathcal{B}_j} m_c(j) - \frac{2y(k)}{\sigma_n^2}$ 
17:    end for
18:  until  $e = 0$ 
19:  Estimate a map  $\Omega$  such that minimizes  $\|\Omega^T \hat{\Phi}(\mathbf{y}) - \text{sign}(\mathbf{y})\|^2$ .
20:  Estimate variance:  $\hat{\sigma}^2 = \text{var}[\Omega^T \hat{\Phi}(\mathbf{y})]$ .
21:  for  $k = 1, 2, \dots, C$  do
22:    Re-initialize LLR:  $m_c(k) = m_c(k) + \frac{-2\Omega^T \hat{\Phi}(\mathbf{y})}{\hat{\sigma}^2}$ .
23:  end for
24:  Update regressors:  $\mathbf{y} = \Omega^T \hat{\Phi}(\mathbf{y})$ .
25: end while

```

Fig. 2. F1-score obtained in the LOS/NLOS classification for LSTM, RFF-LSTM, and the DDRFF-LSTM over suburban scenario C1.

Fig. 3. F1-score obtained in the LOS/NLOS classification for LSTM, RFF-LSTM, and the DDRFF-LSTM for the rural macrocell scenario D1.

Fig. 4. ROC curves obtained in the LOS/NLOS classification for LSTM, RFF-LSTM, and the DDRFF-LSTM over suburban scenario C1 with a training-size of 400 observations.

TABLE IV
SIMULATION PARAMETERS

Room-Size	5m × 5m × 3m
Modulation Technique	OOK
Half power LED semi angle ($\phi_{\frac{1}{2}}$)	60 ⁰
LED-Height	2.25m
Area of photodetector	10 ⁻⁴ cm ²
FOV	70 ⁰
x_{sat}	0.4
Considered values of p	0.5, 1, 2

Fig. 5. ROC curves obtained in the LOS/NLOS classification for LSTM, RFF-LSTM, and the DDRFF-LSTM over rural macrocell scenario D1 with a training-size of 400 observations.

Fig. 6. BER of LDPC decoding with message-passing over nonlinear VLC channels with: a) RFF and b) DDRFF for $p=0.5$.

initialization, at 4 iterations, a degraded BER performance is observed for the RFF based message passing algorithm for $p = 1, 2$ as compared to $p = 0.5$. At 4 iterations, the BER performance of the RFF based message passing is found to be similar (or even sometimes slightly degraded due to transients) compared to the generic MPA that does not rely on RFF maps. Besides, for the low signal to noise ratios (SNRs) (e.g. at 2-6 dB), a slightly degraded performance is observed for the RFF based message passing based decoders, which is attributed to the error in the estimation of Ω in Step. 19 of Algorithm. 1. However, with

Fig. 7. BER of LDPC decoding with message-passing over nonlinear VLC channels with: a) RFF and b) DDRFF for $p=1$.

Fig. 8. BER of LDPC decoding with message-passing over nonlinear VLC channels with: a) RFF and b) DDRFF for $p=2$.

an increase in the number of MPA iterations, a significant gain in the BER-performance is observed for the RFF based MPA decoder. Lastly, it is also observed that the DDRFF based MPA decoder outperforms the generic RFF based MPA decoders. These benefits at these SNRs are very much feasible for typical VLC links, which are well known for their high signal to noise ratio (SNR) levels [49].

Computational Complexity: It is noteworthy that the claimed benefits on computational complexity by Theorems. 2-4 may be interpreted as the number of operations/RFF-sizes needed to achieve a given performance metric with the same number of data-points. From the presented ROC curves, it is observed

that for the same architecture, the RFF-LSTM and the DDRFF-LSTM enhances the performance of a weak LSTM with significantly lower training data-size. In this regard, it is noted that the perceived additional computational complexity the RFF map is actually less if the computational complexity is measured in terms of number of training data-points required to obtain a given pre-fixed performance-level (e.g., an F1 score of 0.88).

VIII. CONCLUSION

In this work, analytical results are derived to motivate RFF based DL. Further, a DDRFF is proposed to lower the training computational complexity for given a performance metric and a dataset-length. The performance-gains promised by the presented analysis and the proposed DDRFF are verified through case-studies, which include the LOS/NLOS identification in presence of outdoor user-mobility, and LDPC decoding over nonlinear VLC channels. In details, simulations were presented for typical WINNER II outdoor scenarios with user-mobility to validate the derived analytical proofs in the context of LSTMs for LOS/NLOS identification, and the viability of MPA based DL was demonstrated for message-passing based LDPC decoding. In alignment with the derived theoretical results, the simulations indicate significant performance-gains for RFF-DL, and for the proposed DDRFF based DL, which makes these approaches promising for enhancing generic DL applications.

ACKNOWLEDGEMENT

This work was supported by the ULTRA TCS research chair on intelligent tactical wireless networks for challenging environments, and by the grant number CRDPJ 538896-19 from the National Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- [1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [2] Y. Bengio *et al.*, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [3] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep learning for wireless communications: An emerging interdisciplinary paradigm," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 133–139, 2020.
- [4] T. Erpek, T. J. O'Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, "Deep learning for wireless communications," in *Development and Analysis of Deep Learning Architectures*. Springer, 2020, pp. 223–266.
- [5] S. Li, C. Wu, H. Li, B. Li, Y. Wang, and Q. Qiu, "FPGA acceleration of recurrent neural network based language model," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2015, pp. 111–118.

- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [7] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.
- [8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [10] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [11] W. Liu, J. C. Principe, and S. Haykin, *Kernel adaptive filtering: A Comprehensive Introduction*. John Wiley & Sons, 2011, vol. 57.
- [12] B. Chen, Y. Zhu, J. Hu, and J. C. Principe, *System parameter identification: information criteria and algorithms*. Newnes, 2013.
- [13] R. Mitra, G. Kaddoum, and V. Bhatia, "Hyperparameter-free transmit-nonlinearity mitigation using a kernel-width sampling technique," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2613–2627, 2020.
- [14] Z. Qin, B. Chen, Y. Gu, N. Zheng, and J. C. Principe, "Probability density rank-based quantization for convex universal learning machines," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2019.
- [15] F. Liu, X. Huang, L. Shi, J. Yang, and J. A. K. Suykens, "A double-variational bayesian framework in random fourier features for indefinite kernels," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2019.
- [16] S. Mehrkanoon and J. A. Suykens, "Deep hybrid neural-kernel networks using random Fourier features," *Neurocomputing*, vol. 298, pp. 46–54, July 2018.
- [17] S. Theodoridis, *Machine learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.
- [18] R. Mitra, G. Kaddoum, and G. Poitou, "Analytical guarantees for hyperparameter free RFF based deep learning in the low-data regime," *IEEE Trans. Circuits Syst. II: Express Briefs*, 2021.
- [19] S. Mehrkanoon, "Cross-domain neural-kernel networks," *Pattern Recognition Letters*, vol. 125, pp. 474–480, July 2019.
- [20] C. Mao, K. Lin, T. Yu, and Y. Shen, "A probabilistic learning approach to UWB ranging error mitigation," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.
- [21] J. Joo, M. C. Park, D. S. Han, and V. Pejovic, "Deep learning-based channel prediction in realistic vehicular communications," *IEEE Access*, vol. 7, pp. 27 846–27 858, Feb. 2019.
- [22] I. Guvenc, C.-C. Chong, and F. Watanabe, "NLOS identification and mitigation for UWB localization systems," in *2007 IEEE Wireless Communications and Networking Conference*. IEEE, Mar. 2007, pp. 1571–1576.
- [23] I. Güvenç, C.-C. Chong, F. Watanabe, and H. Inamura, "NLOS identification and weighted least-squares localization for UWB systems using multipath channel statistics," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, p. 271984, Aug. 2007.
- [24] J.-S. Choi, W.-H. Lee, J.-H. Lee, J.-H. Lee, and S.-C. Kim, "Deep learning based NLOS identification with commodity WLAN devices," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3295–3303, Dec. 2017.
- [25] V.-H. Nguyen, M.-T. Nguyen, J. Choi, and Y.-H. Kim, "NLOS identification in WLANs using deep LSTM with CNN features," *Sensors*, vol. 18, no. 11, p. 4057, Nov. 2018.
- [26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [27] Z. Chen, H. Zou, J. Yang, H. Jiang, and L. Xie, "WiFi fingerprinting indoor localization using local feature-based deep LSTM," *IEEE Systems Journal*, vol. 99, p. PP, June 2019.
- [28] J. Fan and A. S. Awan, "Non-line-of-sight identification based on unsupervised machine learning in ultra wideband systems," *IEEE Access*, vol. 7, pp. 32 464–32 471, Mar. 2019.

- [29] S. Lin and D. Costello, "Error correcting coding: Fundamentals and applications," 1983.
- [30] W. Ryan and S. Lin, *Channel codes: classical and modern*. Cambridge university press, 2009.
- [31] C. Tang, M. Jiang, H. Shen, and C. Zhao, "Analysis and optimization of P-LDPC coded RGB-LED-based VLC systems," *IEEE Photon. J.*, vol. 7, no. 6, pp. 1–13, Dec. 2015.
- [32] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *International Conference on Computational Learning Theory*. Springer, Sept. 2001, pp. 416–426.
- [33] P. Bouboulis, S. Chouvardas, and S. Theodoridis, "Online distributed learning over networks in RKH spaces using random Fourier features," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1920–1932, Apr. 2017.
- [34] S. Bochner, *Harmonic analysis and the theory of probability*. Courier Corporation, 2005.
- [35] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.
- [36] P. Bouboulis and S. Theodoridis, "Extension of Wirtinger's calculus to reproducing kernel hilbert spaces and the complex kernel LMS," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 964–978, Mar. 2011.
- [37] B. Chen, S. Zhao, P. Zhu, and J. C. Príncipe, "Quantized kernel least mean square algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 22–32, Jan. 2012.
- [38] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.
- [39] Z. Qin, B. Chen, and N. Zheng, "Random fourier feature kernel recursive least squares," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2881–2886.
- [40] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, Jan. 2017.
- [41] Y. d. J. Bultitude and T. Rautiainen, "Ist-4-027756 winner ii d1. 1.2 v1. 2 winner ii channel models," *EBITG, TUI, UOULU, CU/CRC, NOKIA, Tech. Rep., Tech. Rep*, 2007.
- [42] Z. Ghassemlooy, L. N. Alves, S. Zvanovec, and M.-A. Khalighi, *Visible light communications: theory and applications*. CRC press, 2017.
- [43] R. Mitra and V. Bhatia, "Precoded Chebyshev-NLMS-based pre-distorter for nonlinear LED compensation in NOMA-VLC," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4845–4856, Nov. 2017.
- [44] H. Elgala, R. Mesleh, and H. Haas, "An LED model for intensity-modulated optical communication systems," *IEEE Photon. Technol. Lett.*, vol. 22, no. 11, pp. 835–837, June 2010.
- [45] S. Jain, R. Mitra, and V. Bhatia, "KLMS-DFE based adaptive post-distorter for visible light communication," *Optics Communications*, vol. 451, pp. 353–360, Nov. 2019.
- [46] S. J. Johnson, *Iterative error correction: Turbo, low-density parity-check and repeat-accumulate codes*. Cambridge university press, 2010.
- [47] T. J. Richardson and R. L. Urbanke, "Efficient encoding of low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 638–656, Feb. 2001.
- [48] M. M. Mansour and N. R. Shanbhag, "High-throughput LDPC decoders," *IEEE Trans Very Large Scale Integr VLSI Syst*, vol. 11, no. 6, pp. 976–996, Dec. 2003.
- [49] S. Al-Ahmadi, O. Maraqa, M. Uysal, and S. M. Sait, "Multi-user visible light communications: State-of-the-art and future directions," *IEEE Access*, vol. 6, pp. 70 555–70 571, 2018.