Hyperparameter Free MEE-FP based Learning for Next Generation Communication Systems

Rangeet Mitra*, Georges Kaddoum*, and Daniel Benevides Da Costa[†]

Abstract

Information theoretic learning (ITL) criteria have emerged useful for mitigating degradations caused by unknown non-Gaussian noise processes in future wireless communication systems. Specifically, the reproducing kernel Hilbert space (RKHS) based approaches relying on ITL based learning criteria are envisioned to provide near-optimal mitigation of unknown hardware impairments and non-Gaussian noises. Among several ITL criteria, the recent works find the minimum error entropy with fiducial points (MEE-FP) promising due to its guarantee of unbiased estimation and generalization over generic noise distributions. However, MEE-FP based learning approaches are known to depend on an accurate kernel-width initialization. Also, the optimal value of this kernel-width is well-known to vary temporally and across deployment scenarios. To remove the dependency on kernel-width, a hyperparameter-free MEE-FP based adaptive algorithm is derived using random-Fourier features with sampled kernel widths (RFF-SKW). In addition, a detailed convergence analysis is presented for the proposed hyperparameter-free MEE-FP, which promises a near-optimal error-floor independent of step-size and guarantees convergence for a wide range of step sizes. The promised hyperparameter-independence and improved convergence for the proposed hyperparameter-free MEE-FP are validated by computer simulations considering different case studies.

Index Terms

Deep-learning, Information theoretic learning, Random Fourier features, RKHS, Hyperparameter independence.

[†]Rangeet Mitra and Georges Kaddoum are with the École de Technologie Supérieure (ÉTS), Université du Québec, Montréal-H3C1K3, Canada, Canada (e-mail: {rangeet.mitra.1@ens.etsmtl.ca, georges.kaddoum@etsmtl.ca}).

[‡]Daniel Benevides Da Costa is with the Future Technology Research Center National Yunlin University of Science and Technology (YunTech) Yunlin, Taiwan (e-mail: {danielbcosta@ieee.org}). This paper was submitted to the IEEE Transactions on Communications on 3rd June, 2021.

I. INTRODUCTION

The next-generation communication systems must be capable of delivering disruptively high data-rates, offering higher bandwidths, and supporting massive connectivity [1], [2]. However, the degradation caused by hardware impairments and generic non-Gaussian noise processes are found to present performance-bottlenecks for these emerging communication systems [3], [4]. Generally, the nature of these hardware impairments and non-Gaussian processes are known to vary from one ecosystem to another and to change temporally. To mitigate these impairments and to support high data-rates, deep-learning (DL) based receivers have emerged as a viable solution [1]. In existing works, several emerging DL paradigms, such as federated-learning [5], meta-learning for low-pilot training [6], reinforcement-learning [7], are derived for several of the relevant sub-tasks of the next-generation communication systems [1].

In parallel, the reproducing kernel Hilbert space (RKHS) based algorithms promise near-optimal mitigation of arbitrary channel-nonlinearities. Specifically, in the context of visible light communications (VLC) [8], [9], initial RKHS based methods focused mostly on dictionary based learning for post-distortion. However, recent works promote the explicit Monte-Carlo approximations of the RKHS, called random Fourier features (RFF), which significantly improves convergence and reduces implementation complexity compared to the pre-existing dictionary based approaches [10]–[13]. In addition, other important works have extended the scope of RKHS based approaches to other applications, such as, detection for massive MIMO (m-MIMO) [13]–[15], parameter-estimation for radar [16], [17], and detection for ultra-violet communications [18]. Further, other works have proposed the RFF based DL (RFF-DL) [19] and hyperparameter-independent RFFs [13], [20], and have carried out a rigorous analysis of RFF-DL in the low-data regime. Under additive white Gaussian noise (AWGN), several of the works reviewed above conclude that for unknown channel-nonlinearities, the RKHS based methods demonstrate a performance which is equivalent to a hypothetical ideal AWGN channel. Thus, the above discussions clearly motivate the optimality of RKHS based methods for generic nonlinear channels.

Non-Gaussian noise processes, such as impulsive-noise (IN) [4] and multi-modal non-Gaussian noises due to user-mobility [21], [22], were found to significantly degrade the performance of classical receivers. These noise-processes arise due to ambient processes, such as ice-cracking, switching transients, and electrical fluctuations [23]. Moreover, these noise processes are known to impair the performance of the internet of things (IoT), and smart-grid [4]. Among several existing methods for non-Gaussian noise-mitigation, the information theoretic learning (ITL) based methods have emerged as attractive for

generalized mitigation of unknown noise-processes [24], [25]. Specifically, ITL is found relevant for applications such as localization [26], [27], detection for sparse code multiple access (SCMA) [28], and for tracking in scenarios with user-mobility [29], [30]. It is further noted that these ITL based approaches have the ability to adapt according to generic non-Gaussian noises, without requiring explicit knowledge of their noise statistics. However, these ITL based methods depend on appropriate choice of hyperparameter-values, whose optimal values are known to vary from one dataset to another. Also, certain ITL based approaches are known for their convergence to biased estimators. However, a recent ITL based approach, namely, the minimum error entropy with fiducial points (MEE-FP), was found to mitigate this drawback and ensure unbiasedness [31]. Besides, most cost-functions for ITL (including the MEE-FP) involve a Gaussian function with a spread parameter, and several works analytically interpret these cost-functions as optimizers of high-dimensional correlation-coefficients [25]. In this context, it is noted that while the recent works aim to approximate these Gaussian functions using RFFs [32], the existing works still depend on the choice of a valid spread parameter for their performance.

To alleviate dependence on the spread parameter and for obtaining an unbiased and accurate MEE-FP based approximation of Gaussian kernels, this work proposes a hyperparameter free MEE-FP using random Fourier features with sampled kernel widths (RFF-SKW) [13]. Additionally, this work derives detailed convergence analysis of the proposed hyperparameter-free MEE-FP and presents relevant case studies. Computer simulations indicate excellent generalization over statistically diverse noise-processes. It is noted that for all the case studies considered in this paper, no hyperparameter-initialization/estimation is performed for either the kernel-width of the Mercer-kernel or the spread-parameters of the ITL cost-functions. In addition, no scenario-specific information on the channel-nonlinearity or any a-priori insight on the distribution/statistics of non-Gaussian noise-processes is assumed at the receiver.

The remainder of this paper is organized as follows: Section-II briefly reviews MEE-FP based learning. Section-III presents analytical theorems for hyperparameter-free MEE-FP, its convergence analysis. Further, Section-IV presents relevant case studies to validate the analytical results. Finally, conclusions are drawn in Section-V.

TABLE 1
LIST OF GENERIC NOTATIONS

$(\cdot)^T$	Transpose Operation
Training Dataset	$\mathcal{D} = \{(d_i, \mathbf{r}_i)\}_{i=1}^{ \mathcal{D} }$
i	Index for dataset $\mathcal{D}$
Cardinality of a set	$ \cdot $
ξ	Regularization parameter
Ω	Weights
η	Step-size
λ	Forgetting-factor
n_G	Number of RFFs
σ_n^2	Variance of the underlying noise-process
$\hat{\Phi}_{\alpha, \beta}$	RFF with sampled kernel-width [13]
ν	Vector of Gamma distributed random variable
$\mathbf{W}$	Matrix consisting of n_G standard Normal rows
$\mathbf{b}$	Uniformly distributed vector
Value of α	$\frac{1}{2}$
Value of β	$\frac{\sigma_n^2}{4\pi^2\alpha}$
$\text{diag}[\cdot]$	Diagonal matrix consisting of elements in the vector $[\cdot]$
ξ	Regularization parameter for MEE-FP
Δ	First-order difference operator

Notations: In this paper, vectors in $\mathbb{C}^n$ or $\mathbb{R}^n$ are denoted by bold-lowercase characters, such as $\mathbf{x}$ with n denoting the number of dimensions, matrices are denoted by bold-uppercase characters, such as $\mathbf{X}$, and the number of random Fourier features is denoted as n_G . In addition, the optimal value of a quantity $(\cdot)$ is denoted by $(\cdot)^o$, and $(\tilde{\cdot})$ denotes the deviation of $(\cdot)$ from its optimal value. Lastly, throughout the paper, the suffix $(\cdot)_i$ denote the value of $(\cdot)$ at the i^{th} sample/instant. Also, the generic notations used in this paper are summarized in Table. I.

II. REVIEW OF MEE-FP BASED LEARNING

In this section, we present a review of MEE-FP based learning for mitigating the non-Gaussian noise processes. In this regard, it is noted that classical minimum mean-square based receivers are well known for their optimality under additive Gaussian noise; however, their performance is found to severely degrade in the presence of non-Gaussian additive noises. It may be noted that while the error-energy may be a sufficient statistic for Gaussian noise processes, the error-energy alone is sub-optimal for the non-Gaussian noise processes that require generalized order statistics for their mitigation [25]. In this regard, the ITL based cost-functions are well-suited for mitigating generalized non-Gaussian processes for the following reasons:

- ITL based methods use implicit estimates of the underlying probability density function (PDF) of the underlying noise-process. In this regard, ITL based methods are well-known to “adapt” to the underlying noise-distribution, without requiring any knowledge or insight on the actual noise-PDF.
- ITL based methods aim to minimize the randomness in the error-processes, rather than some specific error-statistics (as with classical squared-error or cumulant based solutions). In details, these criteria are motivated from statistical mechanics, and the optimal-points of ITL based optimization can be interpreted as “thermal equilibrium” points.
- Some ITL based methods are also interpreted as maximizers of mutual-information and as generalized higher dimensional inner-products.

While ITL based approaches are known for their generalization to arbitrary noise processes, such approaches are prone to an unknown bias, and also significantly depend on a spread-parameter, whose optimal value could vary temporally and from one deployment scenario to another. The recent MEE-FP based criterion is known to mitigate the issue of estimator-bias.

We consider a generalized supervised scenario with a dataset, $\mathcal{D} = \{(d_i, \mathbf{r}_i)\}_{i=1}^{|\mathcal{D}|}$, where $|\cdot|$ denotes cardinality. Without loss of generality, we proceed with the definition of the MEE-FP cost function as:

$$J_{\text{MEE-FP}}(e_i) = \frac{\xi}{|\mathcal{D}|^2} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \kappa_h(e_i - e_j) + \frac{1 - \xi}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \kappa_h(e_i), \quad (1)$$

where, the error, $e_i = d_i - \Omega^T \mathbf{r}_i$, with $\Omega \in \mathbb{R}^{n_G}$ denoting the weights, $\mathbf{r}_i$ indicating the i^{th} observation, $\xi \in [0, 1]$ being a regularization parameter, and d_i denoting the labels. Additionally, $\kappa(\cdot)$ denotes the Gaussian function, which is given by:

$$\kappa(\cdot) = \exp(-h(\cdot)^2). \quad (2)$$

In this regard, the gradient-expression for the optimization of $J_{\text{MEE-FP}}(e_i)$ is given by:

$$\nabla J_{\text{MEE-FP}}(e_i) = \frac{\xi}{|\mathcal{D}|^2} \sum_{i=1}^{|\mathcal{D}|} \kappa'_h(e_i - e_j)(\mathbf{r}_i - \mathbf{r}_j) + \frac{1 - \xi}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \kappa'_h(e_i) \mathbf{r}_i, \quad (3)$$

where, $\mathbf{r}_i$ denotes the regressors at the i^{th} time-instant, and

$$\kappa'_h(\cdot) = -h\kappa_h(\cdot)(\cdot), \quad (4)$$

and the weights, Ω , are updated as:

$$\Omega := \Omega + \eta \nabla J_{\text{MEE-FP}}(e_i), \quad (5)$$

where, η denotes the step-size and $:=$ denotes assignment operator. It is noted that the objective in (1) does not rely on any statistical assumption on the underlying distribution of e_i . Rather, instead of any specific error-statistic, the MEE-FP objective in (1) optimizes the information content of e_i [31], which leads to distribution-invariance. However, we note the dependence of $J_{\text{MEE-FP}}$ on the hyperparameter h , whose optimal value is well known to vary across different datasets.

In this work, we aim to specifically use the generalized inner-product interpretation of the MEE-FP cost-function and re-express the gradients of the $J_{\text{MEE-FP}}$ cost-function as inner-products of the hyperparameter-free RFFs proposed in [13]. For the hyperparameter-free MEE-FP based learning, we present derivations and analytical results in the next section.

III. HYPERPARAMETER-FREE MEE-FP

In this section, we derive the hyperparameter-free ITL. From (1), we note the instantaneous approximation of $J_{\text{MEE-FP}}(e_i)$, $\hat{J}_{\text{MEE-FP}}(e_i)$, as follows:

$$\nabla \hat{J}_{\text{MEE-FP}}(e_i) = \xi \kappa'_h(e_i - e_{i-1})(\mathbf{r}_i - \mathbf{r}_{i-1}) + (1 - \xi) \kappa'_h(e_i) \mathbf{r}_i. \quad (6)$$

Using the random Fourier features with sampled kernel-widths derived in [13], we have the following approximations for $\kappa(e_i)$ and $\kappa(e_i - e_j)$ ¹:

$$\begin{aligned} \kappa(e_i) &\approx \langle \hat{\Phi}_{\alpha,\beta}(d_i), \hat{\Phi}_{\alpha,\beta}(\Omega_i^T \mathbf{r}_i) \rangle, \\ \kappa(e_i - e_j) &\approx \langle \hat{\Phi}_{\alpha,\beta}(e_i), \hat{\Phi}_{\alpha,\beta}(e_{i-1}) \rangle. \end{aligned} \quad (7)$$

where, $\hat{\Phi}_{\alpha,\beta}(\cdot)$ denotes the hyperparameter-free RFF based approximation for the feature map to RKHS. In this work, we use the distribution dependent feature map, as used in [13], and $\langle \cdot, \cdot \rangle$ denotes the inner-product in $\mathbb{R}^{n_G}$, with n_G denoting the number of RFFs. The hyperparameter-free RFF is expressed as:

$$\hat{\Phi}_{\alpha,\beta}(\cdot) = \sqrt{\frac{2}{n_G}} \cos(\text{diag}[\boldsymbol{\nu}] \mathbf{W}(\cdot) + \mathbf{b}), \quad (8)$$

¹It is noted that these approximations are very accurate whose errors fall as $\mathcal{O}\left(\frac{1}{n_G}\right)$.

where, for an input $(\cdot)$ of compatible dimensions, $\text{diag}[\boldsymbol{\nu}]$ denotes a diagonal matrix with the diagonal entries, $\boldsymbol{\nu} \in \mathbb{R}^{n_G}$, being drawn from an inverse-gamma distribution with shape parameter α and scale-parameter β , $\mathbf{W}$ denotes a normal Gaussian random-matrix having n_G rows of unit-variance, and $\mathbf{b} \in \mathbb{R}^{n_G}$ denotes a uniform random vector.

Using the above approximations, we find that (6) is equivalently expressed as:

$$\hat{J}_{\text{MEE-FP}}(e_i) = \xi \langle \hat{\mathbf{\Phi}}_{\alpha,\beta}(d_i), \hat{\mathbf{\Phi}}_{\alpha,\beta}(\boldsymbol{\Omega}_i^T \mathbf{r}_i) \rangle e_i \mathbf{r}_i + (1 - \xi) \langle \hat{\mathbf{\Phi}}_{\alpha,\beta}(e_i), \hat{\mathbf{\Phi}}_{\alpha,\beta}(e_{i-1}) \rangle \Delta e_i \Delta \mathbf{r}_i, \quad (9)$$

where, $\Delta \mathbf{r}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$ and $\Delta e_i = e_i - e_{i-1}$. To estimate the expected gradient of $J_{\text{MEE-FP}}(e_i)$ in (9), we consider a moving average mean-gradient estimator $\mathbf{g}_i$ with forgetting-factor λ :

$$\mathbf{g}_i = \lambda \mathbf{g}_{i-1} + \xi \langle \hat{\mathbf{\Phi}}_{\alpha,\beta}(d_i), \hat{\mathbf{\Phi}}_{\alpha,\beta}(\boldsymbol{\Omega}_i^T \mathbf{r}_i) \rangle e_i \mathbf{r}_i + (1 - \xi) \langle \hat{\mathbf{\Phi}}_{\alpha,\beta}(e_i), \hat{\mathbf{\Phi}}_{\alpha,\beta}(e_{i-1}) \rangle \Delta e_i \Delta \mathbf{r}_i. \quad (10)$$

Using the above gradient-expression, we adapt the weights as:

$$\boldsymbol{\Omega}_{i+1} = \boldsymbol{\Omega}_i + \eta \mathbf{g}_i. \quad (11)$$

Next, we present the convergence analysis of this approach, and provide relevant insights on the convergence-dynamics.

A. Convergence Analysis

In alignment with existing analytical methods for adaptive filters [33], we assume that the $\mathbf{g}_{i-1}$ is uncorrelated from $\mathbf{r}_i, e_i$, and $\Delta \mathbf{r}_i$ in (10), and constants $C_1, C_2 \ll n_G$ we have:

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}_i\|^2] &= \lambda^2 \mathbb{E}[\|\mathbf{g}_{i-1}\|^2] + \xi^2 \left(\mathbb{E}[\exp(-2h^o e_i^2) e_i^2] + \frac{C_1}{n_G} \right) \mathbb{E}[\|\mathbf{r}_i\|^2] + \\ &(1 - \xi)^2 \mathbb{E} \left[\left(\exp(-2h^o (e_i - e_{i-1})^2) + \frac{C_2}{n_G} \right) (e_i - e_{i-1})^2 \right] \mathbb{E}[\|\mathbf{r}_i - \mathbf{r}_{i-1}\|^2]. \end{aligned} \quad (12)$$

Below, we proceed to derive simplifications for $\mathbb{E}[\|\mathbf{g}_i\|^2]$ and $\mathbb{E}[\tilde{\boldsymbol{\Omega}}_i^T \mathbf{g}_i]$.

Simplification for $\mathbb{E}[\|\mathbf{g}_i\|^2]$: Inspecting (12), we aim to first simplify the cross-terms $\mathbb{E}[\exp(-2h^o e_i^2) e_i^2]$ and $\mathbb{E}[\exp(-2h^o (e_i - e_{i-1})^2) (e_i - e_{i-1})^2]$. Using the Bussgang theorem [34], these cross-terms are replaced by the Bussgang correlation coefficient, α_1 , for the function, $f(x^2) = \exp(-2x^2)$. Thus, (12) is expressed

as:

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}_i\|^2] &= \lambda^2 \mathbb{E}[\|\mathbf{g}_{i-1}\|^2] + \\ \xi^2 \mathbb{E}\left[\left(\alpha_1 e_i^2 + \frac{C_1}{n_G}\right) \mathbb{E}[\|\mathbf{r}_i\|^2]\right] &+ (1 - \xi)^2 \mathbb{E}\left[\alpha_1 (e_i - e_{i-1})^2 + \frac{C_2}{n_G}\right] \cdot \mathbb{E}[\|\mathbf{r}_i - \mathbf{r}_{i-1}\|^2]. \end{aligned} \quad (13)$$

Further, we denote $\mathbb{E}[\|\mathbf{r}_i\|^2] = \sigma_r^2$, and assume $\mathbf{r}_i$ as independent and identically distributed (i.i.d) with zero mean, which leads to the following simplification:

$$\mathbb{E}[\|\mathbf{g}_i\|^2] = \lambda^2 \mathbb{E}[\|\mathbf{g}_{i-1}\|^2] + \xi^2 \mathbb{E}\left[\left(\alpha_1 e_i^2 + \frac{C_1}{n_G}\right) \mathbb{E}[\|\mathbf{r}_i\|^2]\right] + (1 - \xi)^2 \left(2\alpha_1 e_i^2 + \frac{C_2}{n_G}\right) 2\mathbb{E}[\|\mathbf{r}_i\|^2], \quad (14)$$

which is further rearranged as:

$$\mathbb{E}[\|\mathbf{g}_i\|^2] \approx \sum_{j=0}^i \lambda^{2(i-j)} [\alpha_1 \xi^2 \sigma_r^2 + 4\alpha_1 (1 - \xi)^2 \sigma_r^2] e_j^2 + \frac{C_3 \mathbb{E}[e_i^2]}{n_G}, \quad (15)$$

where,

$$C_3 = \frac{4(C_1 \xi^2 \sigma_r^2 + 2C_2 (1 - \xi)^2 \sigma_r^2)}{n_G}. \quad (16)$$

Using the summation rule for geometric progression, (15) is further approximated as:

$$\mathbb{E}[\|\mathbf{g}_i\|^2] \approx \sum_{j=0}^i \lambda^{2(i-j)} \left[\alpha_1 \xi^2 \sigma_r^2 + 4\alpha_1 (1 - \xi)^2 \sigma_r^2 \right] \left(\mathbb{E}[\|\tilde{\mathbf{\Omega}}_j\|^2 \sigma_r^2 + \frac{\sigma_n^2}{1 - \lambda^2}] + \frac{C_3 \mathbb{E}[e_i^2]}{n_G} \right), \quad (17)$$

with σ_n^2 denoting the variance of the underlying noise process. Considering a large n_G , $\mathbb{E}[\|\mathbf{g}_i\|^2]$ is approximated as:

$$\mathbb{E}[\|\mathbf{g}_i\|^2] \approx \left[\alpha_1 \xi^2 \sigma_r^2 + 4\alpha_1 (1 - \xi)^2 \sigma_r^2 \right] \left(\mathbb{E}\left[\sum_{j=0}^i \lambda^{2(i-j)} \|\tilde{\mathbf{\Omega}}_j\|^2\right] \sigma_r^2 + \frac{\sigma_n^2}{1 - \lambda^2} \right), \quad (18)$$

Simplification for $\mathbb{E}[\tilde{\mathbf{\Omega}}_i^T \mathbf{g}_i]$: Noting the dependence on $\tilde{\mathbf{\Omega}}_i$ only on the latest update to the term of $\mathbf{g}_i$, we have:

$$\mathbb{E}[\tilde{\mathbf{\Omega}}_i^T \mathbf{g}_i] = (\xi \alpha_2 + (1 - \xi) \alpha_2) \sigma_r^2 \mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2] (\mathbb{E}[\|\mathbf{r}_i\|^2] + \mathbb{E}[\|\mathbf{r}_i - \mathbf{r}_{i-1}\|^2]), \quad (19)$$

with α_2 denoting the corresponding Bussgang coefficient for $f(x) = \exp(-h^o x^2)$. Since $\mathbf{r}_i$ is assumed as zero-mean i.i.d, the above expression is further simplified as:

$$\mathbb{E}[\tilde{\mathbf{\Omega}}_i^T \mathbf{g}_i] = 3\alpha_2 \sigma_r^4 \mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2]. \quad (20)$$

From (11), the dynamical evolution of $\mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2]$ is expressed as:

$$\mathbb{E}[\|\tilde{\mathbf{\Omega}}_{i+1}\|^2] = \mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2] - 2\eta\mathbb{E}[\tilde{\mathbf{\Omega}}_i^T \mathbf{g}_i] + \mathbb{E}[\|\mathbf{g}_i\|^2]. \quad (21)$$

From (18) and (20), and for constants $C_4, C_5 \ll n_G$, we arrive at the following dynamical equation for $\mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2]$:

$$\begin{aligned} \mathbb{E}[\|\tilde{\mathbf{\Omega}}_{i+1}\|^2] = \mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2] & \left[1 + \frac{1}{i} + \frac{2C_5\eta\alpha_2}{n_G} - 6\eta\alpha_2\sigma_r^4 + \left(\frac{\xi^2 + 4(1-\xi)^2}{1-\lambda^2} \right) \eta^2\alpha_1\sigma_r^4 \right] + \\ & \frac{\alpha_1\eta^2(\xi^2 + 4(1-\xi)^2)\sigma_r^2\sigma_n^2}{1-\lambda^2} + \sum_{j=0}^{i-1} \lambda^{2(i-j)} \mathbb{E}[\|\tilde{\mathbf{\Omega}}_j\|^2] \left[\xi^2 + 4(1-\xi)^2 \right] \alpha_1\eta^2\sigma_r^4. \end{aligned} \quad (22)$$

Upon re-arranging (22), it can be re-expressed as:

$$\begin{aligned} \mathbb{E}[\|\tilde{\mathbf{\Omega}}_{i+1}\|^2] = \mathbb{E}[\|\tilde{\mathbf{\Omega}}_i\|^2] & \underbrace{\left[1 + \frac{1}{i} - \frac{2C_5\eta\alpha_2}{n_G} - 6\eta\alpha_2\sigma_r^4 + \left(\frac{\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)}{1-\lambda^2} \right) \eta^2\alpha_1\sigma_r^4 + \frac{C_4\alpha_1\eta^2}{n_G} \right]}_v \\ & + \frac{\alpha_1\eta^2(\xi^2 + 4(1-\xi)^2)\sigma_r^2\sigma_n^2}{1-\lambda^2}. \end{aligned} \quad (23)$$

B. Insights from the Convergence Analysis:

From (23), some insights are highlighted below:

- *Simplification of α_1 and α_2 :* By the Bussgang theorem in [34], we note that the correlation-coefficient, α_1 , is given by:

$$\alpha_1 = \frac{\int_0^\infty \exp(-2h^\circ x) x dx}{\sigma_e^2}, \quad (24)$$

which simplifies to the following value:

$$\alpha_1 = \frac{1}{4\sigma_e^2 h^{\circ 2}}. \quad (25)$$

From the Bussgang theorem, it is similarly noted that:

$$\alpha_2 = \frac{\int_{-\infty}^\infty \exp(-h^\circ x^2) x^2 dx}{\sigma_e^2}, \quad (26)$$

which is simplified as:

$$\alpha_2 = \frac{\sqrt{\pi}}{4\sigma_e^2 h o^{\frac{3}{2}}}. \quad (27)$$

- *Step-size range for convergence:* To guarantee the convergence of (22), we obtain the following condition on the evolution-parameter v (defined below the underbrace in (23)):

$$0 < |v| < 1, \quad (28)$$

which leads to the following range for the step-size when the ideal kernel-width is known:

$$0 < \eta < \frac{6\sqrt{h^o}}{\left(\frac{\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)}{1-\lambda^2}\right)}. \quad (29)$$

Upon choosing $\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)$, this range becomes infinite, which indicates the robustness of the proposed approach to the step-size in the hypothetical scenario when the kernel-width is exactly known. However, since the kernel-width is sampled in practice, the range is rendered finite, and is given by:

$$0 < \eta < \frac{6\sqrt{h^o} + \frac{C_5}{n_G}}{\frac{C_4}{n_G}}. \quad (30)$$

This range is noted to be much higher than the typical ranges for the existing gradient-descent based approaches, since the above range scales with n_G .

- *Mean-squared error:* Setting $i \rightarrow \infty$ in (22) and rearranging, the expression for the steady-state misadjustment is expressed as follows:

$$\mathbb{E}[\|\tilde{\mathbf{\Omega}}_\infty\|^2] = \frac{\alpha_1 \eta (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)) \sigma_n^2}{6\alpha_2 \sigma_r^2 (1-\lambda^2) - \eta \alpha_1 \sigma_r^2 (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2))}, \quad (31)$$

and the steady-state mean squared error as:

$$\mathbb{E}[e_\infty^2] = \frac{\alpha_1 \eta (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)) \sigma_n^2}{6\alpha_2 (1-\lambda^2) - \eta \alpha_1 (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2))} + \sigma_n^2, \quad (32)$$

which, under the low step-size regime, is expressed as:

$$\mathbb{E}[e_\infty^2] = \frac{\alpha_1 \eta (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2)) \sigma_n^2}{6\alpha_2 (1-\lambda^2) - \eta \alpha_1 (\xi^2 + 4(1-\xi)^2 - (1-\lambda^2))} + \sigma_n^2. \quad (33)$$

- *Choice of λ :* From (23), it is noted that on choosing $\lambda = \sqrt{(1-\xi^2) + 4(1-\xi)^2}$, the expression for

the steady-state error in (33) reduces to the following:

$$\mathbb{E}[e_\infty^2] = \sigma_n^2. \quad (34)$$

Therefore, it is concluded that the proposed algorithm approaches the noise variance-floor of the underlying noise-process irrespective of the noise-distribution, and does so with a time-constant of $\frac{1}{6\eta\alpha_2\sigma_r^4}$. Since α_2 is typically very large for the medium/high SNR range, the convergence of this approach is significantly faster than for known first-order methods. Notably, this approach did not require any unknown hyper-parameters to achieve the noise-variance floor.

In the next section, we present case studies to highlight the generalization of the proposed approach to different communication scenarios.

IV. SIMULATION RESULTS AND DISCUSSIONS

Based on the analysis presented above, we present case studies in this section to demonstrate the universal applicability of the above approach. We consider two case studies: a) Channel estimation for visible light communication (VLC) systems that are jointly impaired by user-mobility and imperfect channel-state information (CSI), and b) detection for m-MIMO systems with impulsive noise (IN), and nonlinear characteristics due to the transmit-side power-amplifier (PA). For the presented case studies, it is noted that the MEE-FP based adaptive algorithms do not rely on prior information on the underlying noise-statistics, and do not require any insight on the statistics of the impairments.

A. Case study I: Channel Estimation for VLC Systems with User-Mobility and Imperfect CSI

In this section, we present a case study on VLC systems impaired by user-mobility and with channel estimation errors at the receiver. For this case study, we consider a random-waypoint (RWP) based mobility model [35]. We detail the system model for this case study next, which is summarized in Fig. 1.

System Model: We consider an on-off keying (OOK) transmission consisting of symbols, x_i at the i^{th} time-instant. These symbols are offset by a DC bias Γ . The corresponding observations at the receiver, y'_i , are denoted as:

$$y'_i = h(x_i + \Gamma) + n_i, \quad (35)$$

Fig. 1. System Model for case study I.

where, $n_i \sim \mathcal{N}(0, \sigma_n^2)$, denotes an additive white Gaussian noise (AWGN) process with zero mean and variance σ_n^2 and h denotes the channel-gain. Moreover, the RWP mobility model admits the following probability density function (PDF) for h [35]:

$$p(h) = \sum_{i=1}^4 K_i h^{-a_i}, \quad (36)$$

where, the coefficients for K_i and a_i are derived from the VLC system-model parameters in [35, Eq. (4)]. At the receiver, we assume an erroneous estimate of h , denoted by [22], [36], [37, Eq. (21)]:

$$\hat{h} = h + \delta, \quad (37)$$

with δ denoting the channel estimation error, which is drawn from a zero mean Gaussian distribution with variance σ_δ^2 . At the receiver, upon removal of the DC bias and subsequent maximal ratio combining (MRC) with this erroneous estimate of h , we have the equivalent system model:

$$y_i = \hat{h}x_i + \delta x_i + n_i. \quad (38)$$

Notably, this system has a noise floor due to the term δx_i and the degradation caused by a saturation in the bit-error rate (BER) performance due to this noise-floor was quantified in [36]. The saturation in

the BER performance calls for generalized channel estimation methods to improve the BER performance. It may be noted that although minimum bit error-rate based channel estimation/combining was proposed in [36], the algorithm requires a detailed insight on the underlying parameters of the VLC system-model, such as exact knowledge of the estimation-error level, σ_δ^2 , mobility model parameters $h_{\min}$ and $h_{\max}$, etc. The list of notations for this case study are summarized in Table. 2.

TABLE 2
LIST OF NOTATIONS FOR CASE STUDY-I:

h	VLC channel-gain
x_i	Transmitted symbols
ζ	Random variable modeling DC-bias instability
κ_h	Gaussian-kernel with hyperparameter h
ξ	Regularization parameter
Γ	DC bias
δ	Channel estimation error
n_i	additive white Gaussian noise (AWGN)
$\hat{h}$	Channel estimate
Training Dataset	$\mathcal{D} = \{y_i, x_i^{\text{pilots}}\}_{i=1}^{N_P}$
N_P	Number of pilots
x_i^{pilots}	Pilot-symbols at the i^{th} time-instant

Hyperparameter-free MEE-FP based channel estimation: In order to perform accurate channel estimation without relying on scenario-specific parameters, the proposed hyperparameter-free MEE-FP based channel estimation aims to gain an accurate channel estimate by considering the following dataset of $\{y_i, x_i^{\text{pilots}}\}_{i=1}^{N_P}$, with N_P denoting the number of pilots. For each iteration, we calculate the error-term as follows:

$$e_i = y_i - \hat{h}_i x_i^{\text{pilots}}, \quad (39)$$

Using this error-term for every iteration, we estimate the gradient according to (10), as follows:

$$g_i = \lambda g_{i-1} + \xi \langle \hat{\Phi}_{\alpha,\beta}(d_i), \hat{\Phi}_{\alpha,\beta}(\hat{h}_i x_i^{\text{pilots}}) \rangle e_i x_i + (1 - \xi) \langle \hat{\Phi}_{\alpha,\beta}(e_i), \hat{\Phi}_{\alpha,\beta}(e_{i-1}) \rangle \Delta e_i \Delta x_i^{\text{pilots}}, \quad (40)$$

Finally, for every iteration, we adapt the channel estimate as:

$$\hat{h}_{i+1} = \hat{h}_i + \eta g_i, \quad (41)$$

where, η is the step-size. The proposed algorithm is summarized in Algorithm 1, and the architecture of the estimator is depicted in Fig. 2.

Fig. 2. Illustration of the Hyperparameter-free MEE-FP based channel estimator.

Algorithm 1 Hyperparameter-free MEE-FP based Channel Estimation.

- 1: Initialize: $\hat{h} = 0$, $e_1 = y_1$, $\Delta e_1 = 0$, $\Delta x_1 = 0$, $\alpha = \frac{1}{2}$, $\beta = \frac{\sigma_n^2}{4\pi^2\alpha}$, MAXITER = N_P , Step-size η .
- 2: Choose λ, ξ such that $(1 - \lambda^2) = \xi^2 + 4(1 - \xi)^2$.
- 3: **while** $i \leq \text{MAXITER}$ **do**
- 4: Estimate gradient as per (40);

$$g_i = \lambda g_{i-1} + \xi \langle \hat{\Phi}_{\alpha,\beta}(d_i), \hat{\Phi}_{\alpha,\beta}(\Omega_i^T x_i) \rangle e_i x_i^{\text{pilots}} + (1 - \xi) \langle \hat{\Phi}_{\alpha,\beta}(e_i), \hat{\Phi}_{\alpha,\beta}(e_{i-1}) \rangle \Delta e_i \Delta x_i^{\text{pilots}}$$
- 5: Update channel estimates:

$$\hat{h}_{i+1} = \hat{h}_i + \eta g_i$$

6: **end while**

Simulations: Next, we present computer simulations for the performance evaluation of the hyperparameter-free MEE-FP channel estimation algorithm and for the validation of its convergence-dynamics. We consider a VLC setup, with a maximum mobility-radius $r_{\max} = 5\text{m}$, i.e. at a given instant the user's radius $r \in [0, r_{\max}]$. The results are shown for various LED semi-angles, and a 5% imperfect CSI. The simulation parameters are summarized in Table. 3. The results for the proposed channel estimation method are displayed in Figs. 3, 4, and 5 for LED half angles of 40° , 50° , and 60° respectively. From the BER results in Figs. 1-3, it is observed that the BER performance of the proposed hyperparameter free MEE-FP based algorithm approaches that of the MBER combiner. It may be additionally noted that though the normalized least mean squares (NLMS) based channel estimation algorithm converges to the same error-level as the

TABLE 3
SIMULATION PARAMETERS

Room-Size	5m × 5m × 3m
Modulation Technique	OOK
Half power LED semi angle ($\phi_{\frac{1}{2}}$)	40 ⁰ , 50 ⁰ , 60 ⁰
LED-Height	2.25m
Area of photodetector (A)	10 ⁻⁴ cm ²
$r_{\max}$	5m
α	$\frac{1}{2}$
β	$\frac{\sigma_n^2}{4\pi^2\alpha}$

proposed hyperparameter-free MEE-FP based channel estimator, the proposed channel estimator, for all the half-angle values, delivers improved convergence as seen from the presented mean-squared error (MSE) plots. Moreover, the proposed hyperparameter-free MEE-FP does not use any unknown values, e.g., the kernel-width σ and does not require heuristic based initialization or estimation by rule of thumb. On the other hand, the proposed estimator “adapts” to the noise-distribution.

Fig. 3. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility for 40° LED half-angle.

To further build on the hyperparameter-independence and the generalization of the achieved benefits to arbitrary noise shapes, the simulations were repeated for the same VLC scenario, where an additional assumption of DC-bias instability was considered. The effect of DC bias instability was modeled by a uniformly distributed random-variable $\zeta \sim U([- \epsilon, \epsilon])$. Incorporating the term corresponding to the DC-bias instability, the corresponding expression for (38) is expressed as:

$$y_i = \hat{h}x_i + \zeta\Gamma + \Delta x_i + n_i. \quad (42)$$

Repeating the same channel estimation algorithm for $\epsilon = 0.2$, and for half-angles 40°, 50°, and 60°, we conclude from Figs. 6, 7, and 8 respectively, that the proposed hyperparameter-free MEE-FP based

Fig. 4. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility for 50° LED half-angle.

Fig. 5. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility for 60° LED half-angle.

channel estimator presents negligible degradations to the “clairvoyant” MBER detector (called as such due to its requirement for scenario-specific details, like the estimation-error levels, half-angle values, and other parameters specific to the underlying mobility-model), and demonstrates faster convergence compared to the NLMS based channel estimator. It is also noted that due to the bias from the uniformly distributed DC bias instability, the performance of the NLMS algorithm is found to degrade compared to other approaches, as seen from the BER plots and the MSE characteristics. These simulations demonstrate the generalized benefits of the proposed hyperparameter-free MEE-FP algorithm over different system models. Next, to validate the steady-state error expression presented in (33), the step-size η was varied between $6e-3$ and $1.2e-2$, and without forsaking generality, the simulated steady-state error was plotted in Fig. 9 for SNR values of 18dB, 20dB, and 23dB. Further, from Fig. 9 and as per (33), we find no dependence of the steady-state error on the step-size η , since we choose $\xi^2 + 4(1 - \xi)^2 = 1 - \lambda^2$. Rather, the steady-state

Fig. 6. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility, and DC bias-instability for 40° LED half-angle.

Fig. 7. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility, and DC bias instability, for 50° LED half-angle.

error floor is mostly flat for various values of the step-size, and is given as follows (upto an added factor of $\frac{1}{n_G}$):

$$\mathbb{E}[e_\infty^2] = \sigma_n^2 + \frac{\Xi}{n_G}, \quad (43)$$

where, from (17), Ξ denotes a Lipschitz constant, which, for fitting, was 0.1 for $\text{SNR} = 18\text{dB}$, 1.5 for $\text{SNR} = 20\text{dB}$, and 2.9 for $\text{SNR} = 23\text{dB}$. One may note that all the above values of Ξ are relatively small as compared to n_G , and that the monotonicity of the converged MSE is maintained with respect to the SNR, regardless of the step-size. This SNR independent convergence is guaranteed only when η lies within a certain range, which is derived in (30).

Fig. 8. BER performance and convergence comparison for the hyperparameter-free MEE-FP channel estimator under RWP based user-mobility, and DC bias instability, for 60° LED half-angle.

Fig. 9. Analytical validation of the MSE Floor for various step-sizes, η , and various SNR values.

B. Parallel m -MIMO Detection Under PA Nonlinearity and IN:

In this section, we present a case study for uplink m -MIMO detection in the presence of impairments such as IN and PA nonlinearity. For the channel-model for this case study (similar to the ones in [13], [14], [38]), we consider U users each equipped with a single antenna. The channel-matrix at lag τ is given by $\{\mathbf{H}_\tau\}_{\tau=1}^T$, with T denoting the channel-memory, while the vector of user-symbols for time index τ is denoted as $\mathbf{s}_\tau \in \mathbb{C}^U$. Furthermore, B denotes the number of antennas at the base-station, and each $\mathbf{H}_\tau \in \mathbb{C}^{B \times U}$ is modeled to the correlated Rayleigh channel matrix with correlation coefficient ρ for $\tau \neq 0$. For $\tau = 0$, a line of sight (LoS) component with a Rice factor of 6dB is added to the correlated Rayleigh channel. The received signal, $\mathbf{u}_j$, at the base station at the j^{th} time-instant, $\mathbf{u}_j \in \mathbb{C}^B$, is denoted as [14],

[15]:

$$\mathbf{u}_j = \sum_{\tau=0}^T \mathbf{H}_\tau F(\mathbf{s}_{j-\tau}) + \mathbf{n}_j, \quad (44)$$

where, $\mathbf{n}_j$ is a two component IN vector with zero componentwise-means, covariance matrices $\sigma_n^2 \mathbf{I}$ and $1000\sigma_n^2 \mathbf{I}$, and probabilistic weights 0.99, and 0.01 [39], with $\mathbf{I}$ denoting the identity matrix. Several popular models exist for the nonlinear PA characteristics, such as Rapp, Saleh, Ghorbani etc. [40]. However the proposed approach does not consider the knowledge of their characteristics to be known at the receiver. Without loss of generality, we consider an AM-AM Rapp nonlinearity for our simulations, which is mathematically expressed as follows [40]:

$$F(x) = \frac{x}{\left(1 + \left|\frac{x}{x_{\text{sat}}}\right|^{2p}\right)^{\frac{1}{2p}}}, \quad (45)$$

where, p is a parameter that controls the severity of the nonlinearity. Without loss of generality, we consider a Rapp model for $F(\cdot)$, and the RFF-SKW features for its generic hyperparameter-free mitigation [13]. In this context, it noted that the RFF-SKW in [13] were successful in mitigating the impairments due to PA nonlinearities for AWGN noise, without the requirement of an accurate estimate of the kernel-width.

The considered system model is summarized in Fig. 10, and a summary of notations for this case study is provided in Table 4.

TABLE 4
LIST OF NOTATIONS FOR CASE STUDY-II:

$\mathbf{s}_j$	Transmitted symbol-vector
$\mathbf{H}_\tau$	m-MIMO channel-matrix for the time-instant τ
$F(\cdot)$	PA Nonlinearity
$(\cdot)^{(l)}$	Quantity $(\cdot)$ corresponding to the l^{th} block
T	Channel-memory
$\mathbf{u}_j$	Received regressor at the j^{th} time-instant
L	Number of blocks

Besides the PA nonlinearities, the m-MIMO systems are known to be impaired by IN. For systems impaired by IN, ITL based approaches are well known to be better suited [25]. However, the well-known ITL based criteria are well-known to depend on a spread parameter, whose optimal value is problem-specific, and acquiring a-priori insight on the value of this hyperparameter is non-trivial. Therefore, in this case study, we utilize the proposed hyperparameter-free MEE-FP based method for symbol-detection (as

Fig. 10. System model for case study II.

Fig. 11. Block-diagram of the proposed parallel detector.

mentioned previously, it is an ITL based approach that combines the benefits of the MEE and maximum correntropy based approaches and ensures unbiasedness), which has the potential to offer significantly improved convergence as compared to the squared-error based approach in [13].

For generic m-MIMO based parallel detection in RKHS [13] over AWGN, the complexity due to high-

dimensional regressors (i.e. the large n , which occurs when the base-station is equipped with a large number of antennas) can be mitigated using a parallel RKHS based detector proposed in [14], [15]. Notably, the formulation in [14] was dictionary-based. Due to the reliance on an online dictionary, this approach required extra computations for constructing the dictionary and was found to exhibit instability in early iterations due to concurrent updation of the dictionary and the parameter. To address these shortcomings, an RFF based formulation of the parallel RKHS based detector was derived in [15], and using the RFF-SKW features, a detector was proposed in [13], which could generalize across several nonlinearity-types under AWGN. However, the approach in [13] was a squared error based approach, which causes performance to degrade in the presence of IN.

In this case study, we use an RFF-SKW based parallel detector, which as far as nonlinearity-mitigation is concerned, is well-known to alleviate hyperparameter-dependence. First, the observations are transformed using a mapping $\mathbf{x}(\mathbf{u}) : \mathbb{C}^B \rightarrow \mathbb{R}^{2B}$ defined as $\mathbf{x}(\mathbf{u}) = [\Re(\mathbf{u})^T \Im(\mathbf{u})^T]^T$. For notational convenience, $\mathbf{x}(\mathbf{u})$, which denotes the regressors in this particular study, is simply denoted as $\mathbf{x}$. Due to the large dimensionality of the regressors, the inference is split into L blocks between the compute nodes. The l^{th} regressor-block is denoted as $\mathbf{x}^{(l)} \in \mathbb{C}^{\frac{2B}{L}}$, which is mapped at each compute node to $\hat{\Phi}_{\alpha,\beta}(\mathbf{x}^{(l)}) \in \mathbb{C}^{n_G}$, denoted as $\hat{\Phi}_{\alpha,\beta}^{(l)}$ for simplicity. At each node, the local estimate of $\mathbf{s}$ is denoted as $\hat{\mathbf{s}}^{(l)} = \mathbf{\Omega}^{(l)T} \hat{\Phi}_{\alpha,\beta}^{(l)}$, where $\mathbf{\Omega}^{(l)} \in \mathbb{R}^{n_G \times U}$ is the local parameter at the l^{th} compute node. For a modulation-order M , the best values of α' and β' for the SKW based regressors are given by [13]:

$$\alpha' = \frac{\sqrt{M}}{2}, \beta' = \frac{L\mathbb{E}[\|\mathbf{u}_j\|^2]}{4\pi^2}. \quad (46)$$

At each compute node, the parameter $\mathbf{\Omega}^{(l)}$ is learnt using the proposed MEE-FP, with the appropriate values of α and β to adapt the criterion, and the corresponding values of α' and β' for mapping incoming regressors to the RFF. Notably, for the proposed approach, we assumed knowledge of neither the ideal value of the kernel-width nor any particular value of the spread parameter of the MEE-FP. Finally, over all the L compute-units, majority voting based fusion is performed over all the block wise estimates of the user-symbols. The proposed algorithm is summarized in Algorithm 2, with the demodulation operation denoted as $\mathcal{Q}(\cdot)$. In addition, the proposed parallel detection algorithm is pictorially depicted in Fig. 11. Next, in Fig. 12, we perform a comparison between the KLMS based RFF-SKW proposed in [13], and the RFF-SKW based hyperparameter free MEE-FP proposed in this work. The simulation parameters for this case study is summarized in Table 5. Similar to the setup in [13] for the KLMS based RFF-SKW,

Algorithm 2 Parallel MEE-FP with RFF-SKW

```

1: Initialize:  $\Omega^{(l)} = \mathbf{0}$ ,  $N = \text{NUMPILOTS}$ , step-size  $\eta$ .
2: Choose  $\lambda, \xi$  such that  $(1 - \lambda^2) = \xi^2 + 4(1 - \xi)^2$ . % Training-Phase
3: for  $l = 1$  to  $L$  do
4:   Initialize:  $\mathbf{e}_1 = s_{1-D}, \mathbf{e}_1 = \mathbf{0}$ ,  $\mathbf{g}_i = \mathbf{0}$ 
5:   while  $j \leq N$  do
6:      $\hat{\mathbf{s}}_j^{(l)} = \Omega_{j-1}^{(l)T} \hat{\Phi}_h^{(l)}(\mathbf{x}_j)$ 
7:     Calculate error:  $\mathbf{e}_j = \mathbf{s}_{j-D} - \hat{\mathbf{s}}_j^{(l)}$ .
8:     Update gradient-estimate:

$$\mathbf{g}_j = \lambda \mathbf{g}_{j-1} + \xi \langle \hat{\Phi}_{\alpha,\beta}(d_j), \hat{\Phi}_{\alpha,\beta}(\Omega_j^T x_j) \rangle e_j \hat{\Phi}_{\alpha',\beta'}^{(l)}(\mathbf{x}_j) + (1 - \xi) \langle \hat{\Phi}_{\alpha,\beta}(e_j), \hat{\Phi}_{\alpha,\beta}(e_{j-1}) \rangle \Delta e_j \Delta \hat{\Phi}_{\alpha',\beta'}^{(l)}(\mathbf{x}_j)$$

9:     Update weights:

$$\Omega_j^{(l)} = \Omega_{j-1}^{(l)} + \eta \mathbf{g}_j$$

10:   end while
11: end for
12: % Testing-Phase
13: while  $j > 0$  do
14:   for  $l = 1$  to  $L$  do
15:      $\hat{\mathbf{s}}_j^{(l)} = \mathcal{Q}(\Omega^{(l)T} \hat{\Phi}_{\alpha',\beta'}^{(l)}(\mathbf{x}_j))$ .
16:   end for
17:   Output:  $\hat{\mathbf{s}}_j = \text{mode}_{\forall l=1:L} \{\hat{\mathbf{s}}_j^{(l)}\}$ .
18: end while

```

Fig. 12. BER performance of the proposed hyperparameter-free MEE-FP based parallel detector.

we assume a 4-quadrature phase shift keying (QPSK) transmission with $U = 4$, $B = 320$ antennas at the base-station, and a step-size $\eta = 0.7$. It is noted that for this study for RFF-SKW in [13], reported delayed convergence due to the approximation error caused by kernel-width sampling (which is well known to be non-Gaussian in general). Further, as opposed to the AWGN additive-distortion in [13], this work evaluates the performance of the MEE-FP based parallel detector in IN environments.

TABLE 5
SUMMARY OF SIMULATION-PARAMETERS FOR THE CASE STUDY ON PARALLEL M-MIMO DETECTION

Number of users: U	4
Number of antennas: B	320
Step-size: η	0.5
Number of parallel splits considered: L	2,4,8
Channel-memory: T	4 [38]
Distribution for LOS m-MIMO channel: $\mathbf{H}_{\tau=0}$	Rice-distribution with Rice factor $6dB$ [38]
Distribution for NLOS m-MIMO channel: $\mathbf{H}_{\tau \neq 0}$	Rayleigh-distribution [38]
n_G	200
IN Model	Two component Middleton-A [41]
Probabilistic weights for IN, and corresponding variances	$[0.99, 0.01], [\sigma_n^2, 300\sigma_n^2]$

From the simulations presented in Fig. 12, it is observed that the proposed MEE-FP shows improved adaptability to the errors caused by the sampled kernel width, and IN. These non-Gaussian processes are found to cause degradations in the BER performance and the MSE convergence characteristics for the considered values of L . Specifically, from Fig. 12, it is observed that the convergence of the RFF-SKW in [13] is slower and the RFF-SKW achieves a degraded BER performance for the split values $L = 2, 4, 8$. Notably, these performance-improvements for the proposed hyperparameter free MEE-FP are obtained without any knowledge of the kernel-width parameter or the value of the spread-parameter, which makes the proposed hyperparameter-free MEE-FP suitable for improving the performance of learning systems over unknown non-Gaussian noises.

V. CONCLUSIONS

In this paper, a hyperparameter-free formulation of an ITL based learning criterion, namely the MEE-FP, was derived using the RFF-SKW. The proposed hyperparameter-free MEE-FP was found to self-adapt according to different noise-distributions, and a scenario-specific initialization of the spread parameter was found to be unnecessary. An elaborate convergence-analysis was presented, in which step-size range was derived to ensure the convergence of the proposed approach and an expression for steady-state error was quantified. This enables the proposed approach to generalize across various deployments/datasets which are impaired by generic non-Gaussian noises. This hyperparameter-independence and distribution-invariance of the proposed approach was highlighted through various case studies.

REFERENCES

- [1] S. Ali *et al.*, “6G white paper on machine learning in wireless communication networks,” *arXiv preprint arXiv:2004.13875*, 2020.
- [2] D. A. Awan, R. L. Cavalcante, M. Yukawa, and S. Stanczak, “The roadmap to 6G: AI empowered wireless networks,” *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

- [3] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Commun. Surveys Tut.*, vol. 21, no. 4, pp. 3109–3132, 2019.
- [4] B. Selim, M. S. Alam, J. V. Evangelista, G. Kaddoum, and B. L. Agba, "NOMA-based IoT networks: Impulsive noise effects and mitigation," *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 69–75, 2020.
- [5] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, 2020.
- [6] S. Park, H. Jang, O. Simeone, and J. Kang, "Learning to demodulate from few pilots via offline and online meta-learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 226–239, 2021.
- [7] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, 2019.
- [8] R. Mitra and V. Bhatia, "Adaptive sparse dictionary-based kernel minimum symbol error rate post-distortion for nonlinear LEDs in visible light communications," *IEEE Photonics Journal*, vol. 8, no. 4, pp. 1–13, 2016.
- [9] R. Mitra, F. Miramirkhani, V. Bhatia, and M. Uysal, "Mixture-kernel based post-distortion in RKHS for time-varying VLC channels," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1564–1577, 2018.
- [10] P. Bouboulis, S. Pougkakiotis, and S. Theodoridis, "Efficient KLMS and KRLS algorithms: A random fourier feature perspective," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2016, pp. 1–5.
- [11] R. Mitra, S. Jain, and V. Bhatia, "Least minimum symbol error rate based post-distortion for VLC using random Fourier features," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 830–834, 2020.
- [12] R. Mitra, F. Miramirkhani, V. Bhatia, and M. Uysal, "Low complexity least minimum symbol error rate based post-distortion for vehicular VLC," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 800–11 810, 2020.
- [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, 2021.
- [14] R. Mitra and V. Bhatia, "Kernel-based parallel multi-user detector for massive-MIMO," *Computers & Electrical Engineering*, vol. 65, pp. 543–553, 2018.
- [15] V. Chhangani, R. Mitra, and V. Bhatia, "RFF based parallel detection for massive MIMO," in *European Conference on Networks and Communications (EuCNC)*. IEEE, 2020, pp. 1–5.
- [16] U. K. Singh, R. Mitra, V. Bhatia, and A. K. Mishra, "Kernel LMS-based estimation techniques for radar systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 5, pp. 2501–2515, 2019.
- [17] U. Singh *et al.*, "Range and Velocity Estimation using Kernel Maximum Correntropy based Non-linear Estimators in Non-Gaussian Clutter," *IEEE Trans. Aerosp. Electron. Syst.*, 2019.
- [18] V. Bhatia, S. Jain, K. K. Garg, and R. Mitra, "Performance analysis of RKHS based detectors for nonlinear NLOS ultraviolet communications," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2021.
- [19] R. Mitra and G. Kaddoum, "Random Fourier feature based deep learning for wireless communications," *arXiv preprint arXiv:2101.05254*, 2021.
- [20] R. Mitra, G. Kaddoum, and G. Poitou, "Hyperparameter free RFF based deep learning in the low-data regime," *IEEE Transactions on Circuits and Systems II: Express Briefs (Under Review)*, 2021.
- [21] K. R. Sekhar and R. Mitra, "MBER combining for MIMO VLC with user mobility and imperfect CSI," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 376–380, 2019.
- [22] K. R. Sekhar, F. Miramirkhani, R. Mitra, and A. C. Turlapaty, "Generic BER analysis of VLC channels impaired by 3d user-mobility and imperfect CSI," *IEEE Commun. Lett.*, 2021.

- [23] V. Bhatia and B. Mulgrew, "Non-parametric likelihood based channel estimator for Gaussian mixture noise," *Signal Processing*, vol. 87, no. 11, pp. 2569–2586, 2007.
- [24] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.
- [25] B. Chen, Y. Zhu, J. Hu, and J. C. Principe, *System parameter identification: information criteria and algorithms*. Newnes, 2013.
- [26] R. Mitra, G. Kaddoum, G. Dahman, and G. Poitou, "Error analysis of localization based on minimum-error entropy with fiducial points," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1187–1191, 2021.
- [27] —, "Error analysis of localization based on minimum-error entropy with fiducial points," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1187–1191, 2021.
- [28] E. Sfeir, R. Mitra, G. Kaddoum, and V. Bhatia, "Performance analysis of maximum-correntropy based detection for SCMA," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1114–1118, 2021.
- [29] C. Xu, M. Ji, Y. Qi, and X. Zhou, "MCC-CKF: A distance constrained Kalman filter method for indoor TOA localization applications," *Electronics*, vol. 8, no. 5, p. 478, 2019.
- [30] B. Xu, S. Li, A. A. Razzaqi, and J. Zhang, "Cooperative localization in harsh underwater environment based on the MC-ANFIS," *IEEE Access*, vol. 7, pp. 55 407–55 421, 2019.
- [31] Y. Xie, Y. Li, Y. Gu, J. Cao, and B. Chen, "Fixed-point minimum error entropy with fiducial points," *IEEE Trans. Signal Process.*, vol. 68, pp. 3824–3833, 2020.
- [32] K. Li and J. C. Principe, "Fast estimation of information theoretic learning descriptors using explicit inner product spaces," *arXiv preprint arXiv:2001.00265*, 2020.
- [33] A. H. Sayed, *Adaptive filters*. John Wiley & Sons, 2011.
- [34] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IRE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, 1958.
- [35] A. Gupta and P. Garg, "Statistics of SNR for an indoor VLC system and its applications in system performance," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1898–1901, 2018.
- [36] K. Reddy Sekhar and R. Mitra, "MBER Combining for MIMO VLC with User Mobility and Imperfect CSI," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 376–380, 2020.
- [37] H. Marshoud, P. C. Sofotasios, S. Muhaidat, G. K. Karagiannidis, and B. S. Sharif, "On the performance of visible light communication systems with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6350–6364, 2017.
- [38] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, 2016.
- [39] K. F. McDonald and R. S. Blum, "A statistical and physical mechanisms-based interference and noise model for array observations," *IEEE Trans. Signal Proc.*, vol. 48, no. 7, pp. 2044–2056, 2000.
- [40] K. M. Gharaibeh, *Nonlinear distortion in wireless systems: Modeling and simulation with MATLAB*. John Wiley & Sons, 2011.
- [41] S. M. Zabin and H. V. Poor, "Parameter estimation for middleton class A interference processes," *IEEE Trans. Commun.*, vol. 37, no. 10, pp. 1042–1051, 1989.