Signal Detection in MIMO Systems With Hardware Imperfections: Message Passing on Neural Networks

Dawei Gao, *Member, IEEE*, Qinghua Guo, *Senior Member, IEEE*, Guisheng Liao, *Senior Member, IEEE*, Yonina C. Eldar, *Fellow, IEEE*, Yonghui Li, *Fellow, IEEE*, Yanguang Yu, *Senior Member, IEEE*, and Branka Vucetic, *Life Fellow, IEEE*

***Abstract*— We investigate signal detection in multiple-input-multiple-output (MIMO) communication systems with hardware impairments, such as power amplifier nonlinearity and in-phase/quadrature imbalance. To deal with the complex combined effects of hardware imperfections, neural network (NN) techniques, in particular deep neural networks (DNNs), have been studied to directly compensate for the impact of hardware impairments. However, it is difficult to train a DNN with limited pilot signals, hindering its practical application. In this work, we investigate how to achieve efficient Bayesian signal detection in MIMO systems with hardware imperfections. Characterizing combined hardware imperfections often leads to complicated signal models, making Bayesian signal detection challenging. To address this issue, we first train an NN to 'model' the MIMO system with hardware imperfections and then perform Bayesian inference based on the trained NN. Modelling the MIMO system with NN enables the design of NN architectures based on the signal flow of the MIMO system, minimizing the number of NN layers and parameters, which is crucial to achieving efficient training with limited pilot signals. We then represent the trained NN with a factor graph, and design an efficient message passing based Bayesian signal detector, leveraging the unitary approximate message passing (UAMP) algorithm. The implementation of a turbo receiver with the proposed Bayesian detector is also investigated. Extensive simulation results demonstrate that the proposed technique delivers remarkably better performance than state-of-the-art methods.**

***Index Terms*— Hardware imperfections, I/Q imbalance, power amplifier nonlinearity, multiple-input-multiple-output (MIMO), neural networks (NNs), factor graphs, approximate message passing (AMP), Bayesian inference.**

I. Introduction

We consider signal detection for multiple-input multi-output (MIMO) communications in the presence of hardware impairments, which arise, e.g., in millimeter wave (mm-wave) communications, where mm-wave front ends suffer from significant hardware imperfections, compromising signal transmission quality and degrading system performance [1], [2], [3], [4]. A pronounced impairment is in-phase/quadrature (I/Q) imbalance, i.e., the mismatch of amplitude, phase and frequency response between the I and Q branches, which impairs their orthogonality [5]. Power amplifier (PA) nonlinearity leads to nonlinear distortions to transmitted signals, which cannot be overlooked, especially in mm-wave communications [2]. The hardware imperfections need to be handled properly to avoid inducing significant system performance loss.

Many techniques have been considered to mitigate the impact of hardware imperfections. To handle PA nonlinearity, Volterra series based techniques were proposed for nonlinearity compensation at either transmitter or receiver [6], [7]. However, these methods often need to determine a large number of Volterra series coefficients, which is a difficult task. To address this, some simplified approaches such as those based on memory polynomials [8], Hammerstein model [9] and Wiener model [10] were proposed [8]. Addressing I/Q imbalance has also attracted much attention [11], [12], [13], [14]. In [13], a dual-input nonlinear model based on a real-valued Volterra series was proposed to model the I/Q imbalance, and its inverse model was employed at the transmitter to pre-compensate the I/Q imbalance. In [14], a single-user point-to-point mm-wave hybrid beamforming system with I/Q imbalance at the transmitter and its pre-compensation were considered. The pre-compensation technique [14] assumes availability of instantaneous channel state information at

Manuscript received 7 October 2022; revised 6 April 2023; accepted 22 May 2023. Date of publication 12 June 2023; date of current version 9 January 2024. The work of Dawei Gao was supported by the Proof-of-Concept Foundation of Xidian University Hangzhou Institute of Technology under Grant GNYZ2023QC0405 and in part by the Fundamental Research Funds for the Central Universities under Grant XJS222601. The work of Branka Vucetic was supported in part by the Australian Research Council Laureate Fellowship under Grant FL160100032 and in part by the Discovery Project under Grant DP210103410. The associate editor coordinating the review of this article and approving it for publication was A. Zappone. *(Corresponding author: Qinghua Guo.)*

Dawei Gao and Guisheng Liao are with the Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China, and also with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: gaodawei@xidian.edu.cn; liaogs@xidian.edu.cn).

Qinghua Guo and Yanguang Yu are with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: qguo@uow.edu.au; yanguang@uow.edu.au).

Yonina C. Eldar is with the Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Yonghui Li and Branka Vucetic are with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: yonghui.li@sydney.edu.au; branka.vucetic@sydney.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3283275.

Digital Object Identifier 10.1109/TWC.2023.3283275

the transmitter, which can be difficult to achieve in practical scenarios. With higher orders, polynomial-based methods have potential to handle severer nonlinear distortions, which, however, are more prone to numerical instability in determining their coefficients [15], [16], [17]. Most of the polynomial-based algorithms in the literature deal with a single type of hardware imperfection, i.e., either PA nonlinearity or I/Q imbalance. However, hardware imperfections may occur at the same time, leading to combined effects.

Recently, machine learning has been used in mmWave MIMO systems, e.g., the works in [18], [19], and [20]. Neural networks (NNs) have emerged as a promising technique to deal with nonlinear effects in communication systems [21], [22], [23]. In [24], a real-valued time-delay neural network (RVTDNN) was proposed to model PA behaviors. Various variants of RVTDNN were proposed [25], [26] to address the combined effects of hardware impairments. In [25], high-order signal components are applied to the RVTDNN to pre-compensate both the PA nonlinearity and I/Q imbalance. In [26], a deep NN (DNN) based technique was proposed to mitigate combined PA nonlinearity and I/Q imbalance at the transmitter of a MIMO system. In [27], a residual NN was proposed for digital predistortion, where shortcut connections are added between the input and output layer to improve the performance of PA nonlinearity mitigation. These predistortion based methods require feedback from the receiver, which can be inconvenient or difficult to implement, especially in the case of time-variant environments. Post-compensation techniques at the receiver have also been investigated [28], [29]. A recurrent NN (RNN) was proposed in [28] to compensate PA nonlinearity in a fiber-optic link. In [29], a deep-learning (DL) framework that integrates feedforward NN (FNN) and RNN was proposed to combat both the nonlinear distortion and linear interference. However, these works do not consider the impact of I/Q imbalance. Moreover, a significant problem with the DNN based techniques is that a large number of pilot symbols are required to train them properly, leading to unacceptable overhead and hindering their application especially in time-varying environments.

In this work, we investigate signal detection in an uplink multi-user mm-wave MIMO system, where transmitters (at users) suffer from combined distortions of PA nonlinearity and I/Q imbalance due to the use of low-cost mobile devices. To combat the combined effects of hardware imperfections and multi-user interference, the conventional approach is to design a DNN based detector with received signal as input and predicted symbols as output (shown in Fig. 1), which we call direct detection. However, it is difficult to train the DNN with limited pilot symbols. In this work, we investigate how to achieve efficient Bayesian detection in the presence of combined hardware imperfections. To this end, we require a signal model. However, characterizing combined hardware imperfections in a MIMO system leads to a complicated signal model (which may also be subject to modelling errors), making Bayesian signal detection challenging. We propose using an NN to 'model' the MIMO system (i.e., the NN serves as a substitute for the signal model), which captures combined effects of hardware imperfections and multi-user

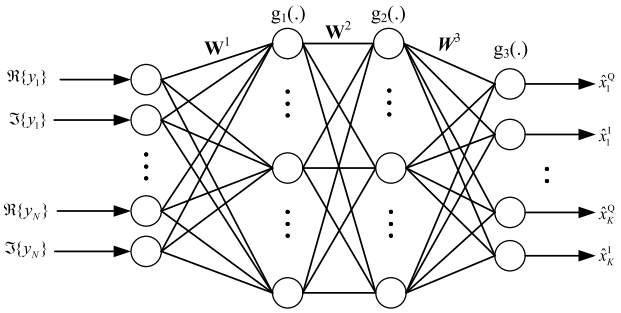


Fig. 1. Illustration of DNN based direct detector.

interference. Then we perform Bayesian inference based on the trained NN. This indirect detection strategy enables us to design the NN architecture based on the signal flow of the MIMO system and minimize the number of its layers and parameters, making it possible to achieve efficient training with limited pilot symbols.

To perform Bayesian inference with the trained NN, we represent it with a factor graph and develop message passing based Bayesian signal detection. The presence of densely connected factors due to the NN weight matrices makes the Bayesian inference difficult. Approximate message passing (AMP) algorithm is promising in handling densely connected factor graphs [30]. AMP works well for i.i.d (sub-) Gaussian matrices, but suffers severe performance degradation or easily diverges for a general matrix [30]. The work in [31] suggests that AMP can still work well in the case of a general matrix when a unitary transform of the original model is used. This leads to unitary AMP (UAMP), also known as UTAMP [31], [32], [33]. As NN weight matrices are normally not i.i.d. (sub-) Gaussian, we adopt UAMP and show that it plays a crucial role in achieving efficient message passing based Bayesian inference.

The contributions of this work are summarized as follows:

- A new strategy to achieve Bayesian signal detection for a communication system with complicated input-output relationship: We use an NN to model the MIMO system, followed by Bayesian inference based on the NN. This indirect detection strategy is more efficient than direct detection. Although this work focuses on MIMO systems with I/Q imbalance and PA nonlinearity, the developed method can be extended to deal with general systems with complicated input-output relationships.
- Signal-flow-based NN architecture design: The architecture of the NN is carefully designed based on the signal flow of the MIMO system, so that the number of layers and parameters of the NN is minimized, which is crucial to achieving efficient training.
- Message passing based Bayesian inference on NNs: To realize Bayesian signal detection based on an NN, we represent the NN as a factor graph and an efficient UAMP-based message passing inference algorithm (called MP-NN) is developed.
- Iterative detection and decoding in coded systems: Another advantage of the new strategy is that the

proposed MP-NN Bayesian detector is able to work with a soft-in-soft-out (SISO) decoder, leading to a powerful turbo receiver. In contrast, it is unknown how to develop a turbo receiver with existing DNN or polynomial based direct detection techniques.

- Comparisons with existing techniques: We carry out various comparisons with state-of-the-art methods and demonstrate that the proposed approach delivers remarkably better performance.

The remainder of the paper is organized as follows. In Section II, the signal model of MIMO communications with combined hardware imperfections is given and existing techniques are introduced. In Section III, with the new strategy, we investigate the NN architecture design and training, and develop a UAMP-based Bayesian detector by performing message passing on the trained NN. The extension to turbo receiver in a coded system is investigated in Section IV. Simulation results are provided in Section V, followed by conclusions in Section VI.

The notations used in this paper are as follows. Boldface lower-case and upper-case letters denote vectors and matrices, respectively. The superscript $(\cdot)^*$ represents the conjugate operation. The notations $(\cdot)^T$ and $(\cdot)^H$ represent the transpose and conjugate transpose operations, respectively. The notation ./ represents elementwise division of two vectors or matrices. We use $|x|$ and $||\mathbf{x}||$ to denote the amplitude of x and the norm of $\mathbf{x}$, and use $\Re\{\cdot\}$ and $\Im\{\cdot\}$ to represent the real and imaginary parts of a complex number, respectively. The notation $\langle f(x)\rangle_{p(x)}$ denotes the expectation of $f(x)$ with respect to distribution $p(x)$.

II. Signal Model and Existing Methods

A. Signal Model

We consider an uplink transmission of a multi-user mm-wave MIMO system with K users. Considering the cost of mobile devices, we assume that each user has a single antenna,[1] where low-cost modulators and PAs are used, resulting in I/Q imbalance and PA nonlinear distortions during transmission [34]. The base station (BS) is equipped with N antennas.

The mth symbol of user k is denoted by $x_k(m) \in \mathcal{A}$, where $\mathcal{A}$ denotes the symbol alphabet. The symbols of all users at time instant m form a vector $\mathbf{x}(m)$. At the transmitter side, the signal is up-converted to radio frequency through modulation, and the mismatch between I and Q branches is characterized as [26]

$$x_k^a(m) = \xi_k x_k(m) + \zeta_k x_k^*(m), \tag{1}$$

where

$$\xi_k = \cos(\frac{\theta_k}{2}) + j\lambda_k \sin(\frac{\theta_k}{2}), \tag{2}$$

$$\zeta_k = \lambda_k \cos(\frac{\theta_k}{2}) + j \sin(\frac{\theta_k}{2}) \tag{3}$$

[1]The extension of this work to the case of each user equipped with multiple antennas is straightforward.

with real valued amplitude imbalance parameter λ_k and phase imbalance parameter θ_k. The signal is then input to a PA.

The nonlinear distortion of the PA is characterized by the amplitude to amplitude conversion $A(|x_k^a(m)|)$ and amplitude to phase conversion $\phi(|x_k^a(m)|)$ [35]:

$$A(|x_k^a(m)|) = \frac{\alpha_a |x_k^a(m)|}{(1 + (\alpha_a \frac{|x_k^a(m)|}{x_{\text{sat}}})^{2\sigma_a})^{\frac{1}{2\sigma_a}}}, \tag{4}$$

$$\phi(|x_k^a(m)|) = \frac{\alpha_\phi |x_k^a(m)|^{q_1}}{1 + (\frac{|x_k^a(m)|}{\beta_\phi})^{q_2}}, \tag{5}$$

where α_a, α_ϕ, β_ϕ, σ_a, x_{sat}, q_1 and q_2 are model parameters. The distorted signal can then be expressed as

$$s_k(m) = f(x_k^a(m)) = A(|x_k^a(m)|) e^{j(\text{angle}(x_k^a(m)) + \phi(|x_k^a(m)|))}, \tag{6}$$

where $\text{angle}(x_k^a)$ denotes the phase of the complex signal x_k^a.

The received signal at time instant m is represented as

$$\mathbf{y}(m) = \mathbf{H}\mathbf{s}(m) + \boldsymbol{\omega}(m), \tag{7}$$

where $\mathbf{H} \in \mathbb{C}^{N \times K}$ is the MIMO channel matrix, $\mathbf{y}(m) = [y_1(m), y_2(m), \ldots, y_N(m)]^T$, $\mathbf{s}(m) = f(\mathbf{x}^a(m))$ with $\mathbf{x}^a(m) = [x_1^a(m), x_2^a(m), \ldots, x_K^a(m)]^T$ being the length-K vector, and $\boldsymbol{\omega}(m)$ denotes a white Gaussian noise vector. Note that the vectors and matrix in (7) are all complex-valued, which can be rewritten as the following real model:

$$\underbrace{\begin{bmatrix}\Re\{\mathbf{y}(m)\}\\ \Im\{\mathbf{y}(m)\}\end{bmatrix}}_{\mathbf{y}'(m)} = \underbrace{\begin{bmatrix}\Re\{\mathbf{H}\} & -\Im\{\mathbf{H}\}\\ \Im\{\mathbf{H}\} & \Re\{\mathbf{H}\}\end{bmatrix}}_{\mathbf{H}'} \underbrace{\begin{bmatrix}\Re\{\mathbf{s}(m)\}\\ \Im\{\mathbf{s}(m)\}\end{bmatrix}}_{\mathbf{s}'(m)} + \underbrace{\begin{bmatrix}\Re\{\boldsymbol{\omega}(m)\}\\ \Im\{\boldsymbol{\omega}(m)\}\end{bmatrix}}_{\boldsymbol{\omega}'(m)}. \tag{8}$$

Due to the combined effects of I/Q imbalance and PA nonlinearity, the input-output relationship of the MIMO system is complex, and is denoted as

$$\mathbf{y}'(m) = \mathcal{S}(\mathbf{x}(m)) + \boldsymbol{\omega}'(m), \tag{9}$$

where $\mathcal{S}(\cdot)$ is the system transfer function.

We assume that the channel matrix and the parameters of I/Q imbalance and PA nonlinearity models are unknown. Each user transmits a pilot signal followed by data. The aim of the receiver at the BS is to detect the transmitted data symbols of all users. To achieve this, there are two approaches.

- Direct detection: A symbol detector is trained directly using pilot symbols, where the input is the received signal and the output is the predicated symbols. As the system transfer function $\mathcal{S}(\cdot)$ is complicated, direct detection seems sensible. To deal with the nonlinearity, polynomial and DNN based techniques have been used in the literature. However, low order polynomials have limited capability to combat the nonlinearity. Although, high order polynomials have better capability, it is difficult to determine the polynomial coefficients due to numerical instability. DNN techniques are more effective in dealing with the nonlinearity, but are difficult to train with limited pilot symbols.
- Indirect detection: With the pilot symbols, the system function $\mathcal{S}(\cdot)$ is first identified, then a symbol detector

is developed based on the system function. This strategy allows the design of powerful Bayesian detectors, but the implementation of indirect detection is challenging. First, to identify $\mathcal{S}(\cdot)$ with pilot symbols, we need to estimate the parameters of the I/Q imbalance and PA nonlinearity models and the MIMO channel at the same time, which is a difficult task due to the nonlinearity. Second, even if we assume that $\mathcal{S}(\cdot)$ is known, it is still difficult to develop a detector, especially a Bayesian one, due to the nonlinearity of $\mathcal{S}(\cdot)$. The aim of this work is to develop a Bayesian detector by using NN and factor graph techniques, which is more powerful than direct detection in the literature.

B. Existing Detection Methods

1) Polynomial Based Direct Detection: A real-valued memory polynomial (RMP) model was developed in [13], where the I/Q branches after modulation are applied to the RMP model in order to compensate the I/Q imbalance. The work was extended to MIMO systems to address the joint effect of I/Q imbalance and PA nonlinearity in [34].

RMP can be used to directly compensate the hardware imperfections and deal with multi-user interference. The detector (for the kth user) is expressed as

$$\tilde{x}_k(m) = \text{argmin}_{\lambda_a \in \mathcal{A}} |\hat{x}_k(m) - \lambda_a| \tag{10}$$

with

$$\hat{x}_k(m) = \hat{x}_k^{\text{Q}}(m) + j\hat{x}_k^{\text{I}}(m). \tag{11}$$

Here,

$$\begin{aligned}\hat{x}_k^{\text{Q}}(m) = &\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{l=0}^{L} a_{n,p,l,k}^{\text{Q}} \Re\{y_n(m-l)\}^p \\ &+ b_{n,p,l,k}^{\text{Q}} \Im\{y_n(m-l)\}^p, \end{aligned} \tag{12}$$

$$\begin{aligned}\hat{x}_k^{\text{I}}(m) = &\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{l=0}^{L} a_{n,p,l,k}^{\text{I}} \Re\{y_n(m-l)\}^p \\ &+ b_{n,p,l,k}^{\text{I}} \Im\{y_n(m-l)\}^p, \end{aligned} \tag{13}$$

where P is the order of the polynomial, L is the memory length, and $\{a_{p,l,k}^{\text{Q}}, a_{p,l,k}^{\text{I}}\}$ and $\{b_{p,l,k}^{\text{Q}}, b_{p,l,k}^{\text{I}}\}$ are the coefficients of the polynomial with respect to the real and imaginary parts of the received signals, respectively. In the case of memoryless distortions (considered in this work), the memory of the RMP can be set to 0.

The RMP based detector is obtained by determining its polynomial coefficients $\{a_{p,l,k}^{\text{Q}}, a_{p,l,k}^{\text{I}}\}$ and $\{b_{p,l,k}^{\text{Q}}, b_{p,l,k}^{\text{I}}\}$ using pilot signals. The models (12) and (13) are linear with respect to the polynomial coefficients. With the squared error between $\{\hat{x}_k(m)\}$ and $\{x_k(m)\}$ as the cost function, the coefficients can be determined using least squares (LS). However, the determination of the coefficients suffers from numerical instability due to the involved matrix inversion, especially when the polynomial order is high [15].

2) DNN-Based Direct Detection: Another way to deal with the complex nonlinear relationship described in Section II-A is to use DNNs. As an example, a detector based on a real-valued DNN with two hidden layers is shown in Fig. 1. Here the received signal is input to the DNN and estimated symbols are output, i.e.,

$$\hat{\mathbf{x}}(m) = \mathcal{DNN}(\mathbf{y}'(m)), \tag{14}$$

where the DNN deals with the combined distortions and multi-user interference. A hard decision can be made based on $\hat{\mathbf{x}}(m)$, i.e., $\tilde{x}_k(m) = \text{argmin}_{\lambda_a \in \mathcal{A}} |\hat{x}_k(m) - \lambda_a|$.

Depending on the number of layers and hidden nodes, the number of parameters of the DNN can be large, leading to difficulties in training as a large number of pilot symbols are required. Furthermore, the training of DNN receivers is prone to overfitting.

III. Bayesian Signal Detection With Message Passing on Neural Networks

We adopt indirect detection and develop a Bayesian detector with the aid of NN and factor graph techniques. The development of the Bayesian detector relies on the signal model (9), in particular the system transfer function $\mathcal{S}(\cdot)$. However, it is difficult to estimate the unknown parameters and MIMO channel required in $\mathcal{S}(\cdot)$. To circumvent this, we train an NN (denoted by $\mathcal{NN}(\cdot)$) to substitute $\mathcal{S}(\cdot)$, i.e., we expect that

$$\mathcal{NN}(\mathbf{x}) \approx \mathcal{S}(\mathbf{x}), \tag{15}$$

for any symbol vector $\mathbf{x}$. The use of the substitute $\mathcal{NN}(\cdot)$ leads to the following benefits:

- Compared to estimating the parameters and MIMO channel involved in $\mathcal{S}(\cdot)$, training of the NN is much easier, i.e., $\mathcal{NN}(\cdot)$ is obtained using back-propagation. Moreover, the use of NNs is able to capture hardware imperfections that may not have explicit mathematical expressions.
- Very different from the use of DNNs in the literature (which are typically a black box), the NN in this work is used to model $\mathcal{S}(\cdot)$. Hence, the architecture of the NN can be carefully designed based on the signal flow of the MIMO system as detailed in Section III-A, so that the number of parameters of the NN can be minimized, which is crucial to achieving efficient training with limited number of pilot symbols.
- Bayesian inference based on $\mathcal{NN}(\cdot)$ is easier than that based on $\mathcal{S}(\cdot)$ as the building blocks of $\mathcal{NN}(\cdot)$ are matrix-vector products and activation functions. We will show in Section III-B that, leveraging UAMP, efficient Bayesian inference for symbol detection can be implemented with message passing.

A. Signal Flow Based NN Architecture Design and Training

As shown in Fig. 2, the NN consists of an input layer, two non-fully connected hidden layers and an output layer. We note that the NN is used to model the system characterized by (1), (6) and (8). The symbols of all users are input to the NN, and the outputs of the NN are the predicted received

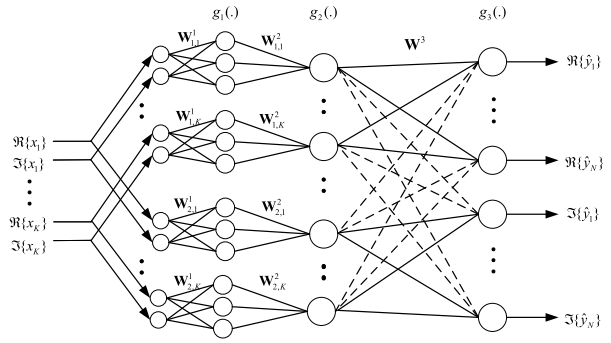


Fig. 2. Proposed NN to characterize hardware imperfections and multi-user interference.

signals, where the real and imaginary parts of the signals are separated to make the NN a real-valued one. The architecture of the NN is designed based on the signal flow expressed with (1), (6) and (8), i.e., the transmitted symbols are first distorted due to I/Q imbalance and PA nonlinearity and then undergo multi-user interference.

In Fig. 2, the input layer, the first hidden layer and the input to the second hidden layer are essentially $2K$ sub-NNs, which are used to model the I/Q imbalance and PA nonlinearity of the K users. Each sub-NN has two input nodes corresponding to the real and imaginary parts of a symbol, and a single hidden layer with N' hidden nodes, where the activation function *Tanh* is employed. As shown in Fig. 2, the real and imaginary parts of a symbol are shared by two sub-NNs, which are called a sub-NN pair. There are K sub-NN pairs in total, and they are indexed by (l,k), where $l=1,2$ and $k=1,\ldots,K$. The pair of sub-NN $(1,k)$ and sub-NN $(2,k)$ models the combined I/Q imbalance and PA nonlinearity of user k shown in (1) and (6), i.e., the output of one sub-NN is expected to be a good approximation to $\Re\{s_k(m)\}$ and the output of the other one is expected to be a good approximation to $\Im\{s_k(m)\}$.

According to Fig. 2, the input to the kth sub-NN pair is denoted as

$$\mathbf{c}_k(m)=[\Re\{x_k(m)\},\Im\{x_k(m)\}]^T. \tag{16}$$

Then, the output of the (l,k)th sub-NN is

$$\mathbf{d}_{l,k}(m)=g_1(\mathbf{W}^1_{l,k}\mathbf{c}_k(m)+\mathbf{b}^1_{l,k}), \tag{17}$$

where $\mathbf{W}^1_{l,k}$ and $\mathbf{b}^1_{l,k}$ are the corresponding weight matrix and bias vector of the sub-NN (l,k), and $\mathbf{W}^1_{l,k}=[\mathbf{w}^1_{l,k,1},\mathbf{w}^1_{l,k,2}]^T$ with $\mathbf{w}^1_{l,k,1}=[w^1_{l,k,11},w^1_{l,k,12},\ldots,w^1_{l,k,1N'}]^T$ and $\mathbf{w}^1_{l,k,2}=[w^1_{l,k,21},w^1_{l,k,22},\ldots,w^1_{l,k,2N'}]^T$. Each sub-NN has one output node, and the output of the (l,k)th sub-NN is expressed as

$$s_{l,k}(m)=(\mathbf{w}^2_{l,k})^T\mathbf{d}_{l,k}(m), \tag{18}$$

where $\mathbf{w}^2_{l,k}=[w^2_{l,k,1},w^2_{l,k,2},\ldots,w^2_{l,k,N'}]^T$ are the output weights of a sub-NN. It is known that an NN with a single hidden layer has the property of universal approximation [36]. We find that the sub-NNs with a single hidden layer in Fig. 2 are sufficient to model the combined PA nonlinear distortion and I/Q imbalance. When all transmitters have the same I/Q imbalance and PA nonlinearity, the sub-NN pairs share the weight and bias parameters, i.e., the parameters of the sub-NN pairs can be tied. This reduces the number of parameters of all sub-NNs from $6KN'$ to $6N'$.

Assume that the combined I/Q imbalance and PA nonlinearity are well modelled using the sub-NNs. The second hidden layer and the output layer are designed to model the multi-user interference. The activation functions of the two layers $g_2(.)$ and $g_3(.)$ are linear as the interference shown in (8) is in a linear form. The second hidden layer is fully connected to the output layer, yielding the predicted in-phase and quadrature components of the received signal. Considering the structure of $\mathbf{H}'$ in (8), the weight matrix $\mathbf{W}^3$ between the second hidden layer and output layer should have the same structure. To impose such a structure on the weight matrix, we tie the elements of the weight matrix, leading to

$$\mathbf{W}^3=\begin{bmatrix}\mathbf{W}^{31} & \mathbf{W}^{32}\\ -\mathbf{W}^{32} & \mathbf{W}^{31}\end{bmatrix}, \tag{19}$$

where $\mathbf{W}^{31}$ and $\mathbf{W}^{32}$ are sub-weight matrices with dimension $N\times K$. It can be seen that the weight matrix has $2KN$ parameters, which is in contrast to the unstructured weight matrix that has $4KN$ parameters. Then the output of the NN is

$$\hat{\mathbf{y}}'(m)=\mathbf{W}^3\mathbf{s}'(m), \tag{20}$$

where $\mathbf{s}'(m)=[s_{1,1}(m),\ldots,s_{1,K}(m),\ldots,s_{2,K}(m)]^T$ is the output vector from the $2K$ sub-NNs, and $\hat{\mathbf{y}}'(m)=[v_{1,1}(m),\ldots,v_{1,N}(m),\ldots,v_{2,N}(m)]^T$ is a length-$2N$ output vector with $v_{1,n}(m)=\Re\{\hat{y}_n(m)\}$ and $v_{2,n}(m)=\Im\{\hat{y}_n(m)\}$. Hence, the predicted signal of the nth receive antenna is represented as $\hat{y}_n(m)=v_{1,n}(m)+jv_{2,n}(m)$.

Training of the NN is straightforward. Suppose the length of the pilot signal is M_0, i.e., we have M_0 training samples $\{(\mathbf{p}(m),\mathbf{t}(m)),m=1,\ldots,M_0\}$, where $\mathbf{t}(m)=[t_1(m),t_2(m),\ldots,t_K(m)]^T$ and $\mathbf{p}(m)=[p_1(m),p_2(m),\ldots,p_N(m)]^T$ denote the pilot symbols and corresponding received signal. The NN is trained with input $\mathbf{t}'(m)=[\Re\{\mathbf{t}(m)\}^T,\Im\{\mathbf{t}(m)\}^T]^T$, expected output $\mathbf{p}'(m)=[\Re\{\mathbf{p}(m)\}^T,\Im\{\mathbf{p}(m)\}^T]^T$ and loss function

$$\text{Loss}=\frac{1}{2N}\frac{1}{M_0}\sum_{m=1}^{M_0}\sum_{n=1}^{2N}(v_n(m)-p'_n(m))^2, \tag{21}$$

i.e., the weights $\{\mathbf{W}^1_{l,k},\mathbf{w}^2_{l,k},\mathbf{W}^3\}$ and biases $\{\mathbf{b}^1_{l,k}\}$ are determined with back-propagation [37].

After training, we obtain the following model:

$$\begin{aligned}\mathbf{y}'(m)&=\hat{\mathbf{y}}'(m)+\boldsymbol{\omega}'(m)\\&=\mathcal{NN}(\mathbf{x}(m))+\boldsymbol{\omega}'(m),\end{aligned} \tag{22}$$

where $\mathcal{NN}(\cdot)$ denotes the trained NN and the term $\boldsymbol{\omega}'(m)$ denotes a noise vector that accounts for training and modelling errors. We then detect the transmitted symbols based on the trained NN, as elaborated in the next section.

B. Bayesian Signal Detection Based on the Trained NN

During the phase of data transmission, we perform Bayesian inference for the transmitted symbols based on the trained NN, i.e., model (22). We assume the error $\boldsymbol{\omega}'(m)$ is white

Gaussian with mean zero and unknown variance ϵ^{-1} (ϵ is the precision). Our aim is to determine the transmitted symbol vector $\mathbf{x}(m)$ based on the received signal $\mathbf{y}(m)$. We use a Bayesian approach, in particular, message passing, where we represent the trained NN (22) as a factor graph. The weight matrix $\mathbf{W}^3$ in the NN leads to a densely connected factor graph, resulting in difficulties in message passing in terms of complexity and convergence. The AMP algorithm is efficient in handling short loops induced by i.i.d. (sub-) Gaussian matrices, but the weight matrix $\mathbf{W}^3$ here is not i.i.d. (sub-)Gaussian. Instead, we use the UAMP algorithm.

According to (20) and (22), we have

$$\mathbf{y}' = \mathbf{W}^3\mathbf{s}' + \boldsymbol{\omega}', \tag{23}$$

where the time index m is dropped for simplicity. As UAMP works with a unitary transformed model, we perform a unitary transformation to (23), i.e.,

$$\mathbf{r} = \mathbf{U}^H\mathbf{y}' = \boldsymbol{\Phi}\mathbf{s}' + \tilde{\boldsymbol{\omega}}, \tag{24}$$

where $\boldsymbol{\Phi} = \mathbf{U}^H\mathbf{W}^3 = \boldsymbol{\Lambda}\mathbf{V}$, $\mathbf{U}$ is obtained from the SVD $\mathbf{W}^3 = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}$, and the noise $\tilde{\boldsymbol{\omega}} = \mathbf{U}^H\boldsymbol{\omega}'$ has the same distribution as $\boldsymbol{\omega}'$ since $\mathbf{U}$ is a unitary matrix. As the noise precision ϵ is unknown, its estimation is included in the detector. Define an auxiliary vector

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{s}', \tag{25}$$

which is treated as a latent variable. Then the joint distribution of $\mathbf{x}$, $\mathbf{s}'$, $\mathbf{z}$ and ϵ given $\mathbf{r}$ can be expressed as

$$p(\mathbf{x}, \mathbf{z}, \mathbf{s}', \epsilon|\mathbf{r}) \propto p(\epsilon)p(\mathbf{r}|\mathbf{z}, \epsilon)p(\mathbf{z}|\mathbf{s}')p(\mathbf{s}'|\mathbf{x})p(\mathbf{x}), \tag{26}$$

where we assume an improper prior for the noise precision, i.e., $p(\epsilon) \propto 1/\epsilon$ [38],

$$p(\mathbf{r}|\mathbf{z}, \epsilon) = \prod_n p(r_n|z_n, \epsilon) \tag{27}$$

with $p(r_n|z_n, \epsilon) = \mathcal{N}(r_n; z_n, \epsilon^{-1})$,

$$p(\mathbf{z}|\mathbf{s}') = \delta(\mathbf{z} - \boldsymbol{\Phi}\mathbf{s}') = \prod_n \delta(z_n - \boldsymbol{\Phi}_n^T\mathbf{s}'), \tag{28}$$

with $\boldsymbol{\Phi}_n^T$ being the nth row of $\boldsymbol{\Phi}$,

$$p(\mathbf{s}'|\mathbf{x}) = \prod_{l,k} p(s_{l,k}|x_{1,k}, x_{2,k}) = \prod_{l,k} \delta(s_{l,k} - f_l(\mathbf{x}'_k)), \tag{29}$$

with $f_l(\mathbf{x}'_k)$ given later in (44) and $\mathbf{x}'_k = [x_{1,k}, x_{2,k}]^T$, and

$$p(\mathbf{x}) = \prod_k p(x_k) = \prod_k (1/|\mathcal{A}|) \sum_{a=1}^{|\mathcal{A}|} \delta(x_k - \lambda_a). \tag{30}$$

Our aim is to obtain the (approximate) marginal (a posteriori distribution) of each transmitted symbol $p(x|\mathbf{r})$, based on which a hard decision can be made with the maximum posterior probability (MAP) criterion.

The factor graph representation for the factorization in (26)-(30) is depicted in Fig. 3, where squares and circles represent function nodes and variable nodes, respectively. To facilitate the factor graph representation, we introduce the notations in Table I, which shows the correspondence between the factor labels and the corresponding distributions they represent.

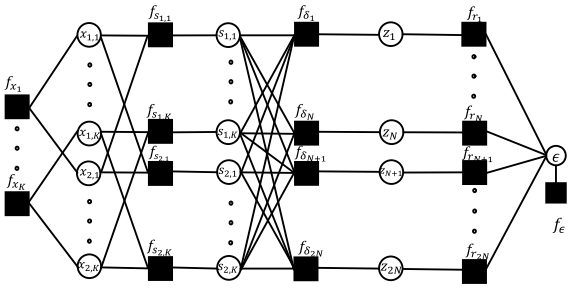


Fig. 3. Factor graph representation of the NN-modeled system.

TABLE I

FACTORS, UNDERLYING DISTRIBUTIONS AND FUNCTIONAL FORMS ASSOCIATED WITH (26)

Factor	Distribution	Functional Form
f_{r_n}	$p(r_n\|z_n, \epsilon)$	$\mathcal{N}(r_n; z_n, \epsilon^{-1})$
f_{δ_n}	$p(z_n\|\mathbf{s}')$	$\delta(z_n - \boldsymbol{\Phi}_n^T\mathbf{s}')$
$f_{s_{l,k}}$	$p(s_{l,k}\|x_{1,k}, x_{2,k})$	$\delta(s_{l,k} - f_l(\mathbf{x}'_\mathbf{k}))$
f_{x_k}	$p(x_k)$	$(1/\|\mathcal{A}\|)\sum_{a=1}^{\|\mathcal{A}\|}\delta(x_k - \lambda_a)$
f_ϵ	$p(\epsilon)$	$\propto \epsilon^{-1}$

We develop a message passing algorithm based on the factor graph in Fig. 3. Due to the presence of loops in the graph, an iterative process is required, which involves several rounds of forward and backward recursions. In particular, we use UAMP to handle the densely connected part of the graph, which is crucial to achieving high performance with low complexity. To deal with various factor nodes, both belief propagation (BP) [39] and variational message passing (VMP) [40] are used. In the following we derive the message updates, where the message passed from node A to node B is denoted by $m_{A\to B}(c)$, which is a function of c. It is noted that the message passing algorithm is an iterative one, and some message computations in the current iteration require messages computed in the last iteration. The message passing algorithm is summarized in Algorithm 1, and the derivations of the algorithm line by line are elaborated in the following.

According to the derivation of (U)AMP using loopy BP, (U)AMP provides the message from variable node z_m to function node f_{r_m}. Due to the Gaussian approximation in the derivation of (U)AMP, the message is Gaussian, i.e.,

$$m_{z_n\to f_{r_n}}(z_n) = m_{f_{\delta_n}\to z_n}(z_n) \propto \mathcal{N}(z_n; p_n, \tau_{p_n}), \tag{31}$$

where the mean p_n and the variance τ_{p_n} are the nth elements of $\mathbf{p}$ and $\boldsymbol{\tau}_p$ given in Lines 1 and 2 of Algorithm 1.

Following VMP, the message $m_{f_{r_n}\to\epsilon}(\epsilon)$ from factor node f_{r_n} to variable node ϵ can be expressed as

$$m_{f_{r_n}\to\epsilon}(\epsilon) \propto \exp\left\{\langle \log f_{r_n}(z_n, \epsilon)\rangle_{b(z_n)}\right\}, \tag{32}$$

where the belief of z_n is given as

$$b(z_n) \propto m_{z_n\to f_{r_n}}(z_n) m_{f_{r_n}\to z_n}(z_n). \tag{33}$$

Later in (39), we will show that $m_{f_{r_n}\to z_n}(z_n) \propto \mathcal{N}(z_n; r_n, \hat{\epsilon}^{-1})$ with $\hat{\epsilon}^{-1}$ being the estimate of ϵ^{-1} in last iteration, and its computation is given in (42). Hence $b(z_n)$ is

Algorithm 1 MP-NN Message Passing Detector

Define vector $\boldsymbol{\lambda} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^H\mathbf{1}$. Initialization: $\tau_s^{(0)} = 1$, $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{c} = \mathbf{0}$, $\hat{\mathbf{x}}^{(0)} = \mathbf{0}$, $\hat{\epsilon} = 1$ and $i = 0$.

Repeat

1: $\boldsymbol{\tau}_p = \tau_s^i\boldsymbol{\lambda}$
2: $\mathbf{p} = \boldsymbol{\Phi}\hat{\mathbf{s}}^i - \boldsymbol{\tau}_p \cdot \mathbf{c}$
3: $\tau_z = \boldsymbol{\tau}_p./(\mathbf{1} + \hat{\epsilon}\boldsymbol{\tau}_p)$
4: $\hat{\mathbf{z}} = (\hat{\epsilon}\boldsymbol{\tau}_p \cdot \mathbf{r} + \mathbf{p})./(\mathbf{1} + \hat{\epsilon}\boldsymbol{\tau}_p)$
5: $\hat{\epsilon} = 2N/(||\mathbf{r} - \hat{\mathbf{z}}||^2 + \mathbf{1}^H\mathbf{v}_z)$
6: $\boldsymbol{\tau}_c = \mathbf{1}./(\boldsymbol{\tau}_p + \hat{\epsilon}^{-1}\mathbf{1})$
7: $\mathbf{c} = \boldsymbol{\tau}_c \cdot (\mathbf{r} - \mathbf{p})$
8: $1/\tau_q = (1/2K)\boldsymbol{\lambda}^H\boldsymbol{\tau}_c$
9: $\mathbf{q} = \hat{\mathbf{s}}^i + \tau_q(\boldsymbol{\Phi}^H\mathbf{c})$
10: $\forall l,k,\ \tilde{q}_{l,k}^i = (\mathbf{w}_{l,k}^2)^T g(\mathbf{W}_{l,k}^1\hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1)$
11: $\forall l,k,\ \eta_{l,k}^i = (\mathbf{w}_{l,k}^2 \cdot \mathbf{w}_{l,k,1}^1)^T g'(\mathbf{W}_{l,k}^1\hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1)$,
12: $\forall l,k,\ \gamma_{l,k}^i = (\mathbf{w}_{l,k}^2 \cdot \mathbf{w}_{l,k,2}^1)^T g'(\mathbf{W}_{l,k}^1\hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1)$
13: $\forall l,k,\ \tau_{\psi_{l,k}}^{l,1} = (\tau_q + \gamma_{l,k}^2\tau_{x_{l'(l'\neq l),k}})/\eta_{l,k}^2$
14: $\forall l,k,\ \tau_{\psi_{l,k}}^{l,2} = (\tau_q + \eta_{l,k}^2\tau_{x_{l'(l'\neq l),k}})/\gamma_{l,k}^2$
15: $\forall l,k,\ \psi_{l,k}^{l,1} = (q_{l,k} - \tilde{q}_{l,k})/\eta_{l,k} + \hat{x}_{l,k}$
16: $\forall l,k,\ \psi_{l,k}^{l,2} = (q_{l,k} - \tilde{q}_{l,k})/\gamma_{l,k} + \hat{x}_{l,k}$
17: $\forall l,k,\ \tau_{\psi_{l,k}} = (1/\tau_{\psi_{l,k}}^{l,1} + 1/\tau_{\psi_{l,k}}^{l,2})^{-1}$
18: $\forall l,k,\ \psi_{l,k} = (\psi_{l,k}^{l,1}/\tau_{\psi_{l,k}}^{l,1} + \psi_{l,k}^{l,2}/\tau_{\psi_{l,k}}^{l,2})\tau_{\psi_{l,k}}$
19: $\forall k,\ \tau_{\tilde{\psi}_k} = \tau_{\psi_{1,k}} + \tau_{\psi_{2,k}}$
20: $\forall k,\ \tilde{\psi}_k = \psi_{1,k} + j\psi_{2,k}$
21: $\forall k,a,\ \xi_{k,a} = \exp(-\tau_{\tilde{\psi}_k}^{-1}|\lambda_a - \tilde{\psi}_k|^2)$
22: $\forall k,a,\ \mu_{k,a} = \xi_{k,a}/\sum_{a=1}^{|A|}\xi_{k,a}$
23: $\forall k,\ \hat{x}_k^{i+1} = \sum_{a=1}^{|A|}\lambda_a\mu_{k,a}$
24: $\forall k,\ \tau_{x_k^{i+1}} = \sum_{a=1}^{|A|}\mu_{k,a}|\lambda_a - \hat{x}_k^{i+1}|^2$
25: Calculate $\eta_{l,k}^{i+1}, \gamma_{l,k}^{i+1}, \tilde{q}_{l,k}^{i+1}$ using Lines 10-12 with $\hat{x}_k^{i+1}$
26: $\forall k,\ \tau_{x_{1,k}}^{i+1} = \tau_{x_{2,k}}^{i+1} = 1/2\tau_{x_k}^{i+1}$
27: $\forall k,\ \hat{x}_{1,k}^{i+1} = \Re\{\hat{x}_k^{i+1}\}, \hat{x}_{2,k}^{i+1} = \Im\{\hat{x}_k^{i+1}\}$
28: $\forall l,k,\ \tau_{s_{l,k}}^{i+1} = (\eta_{l,k}^{i+1})^2\tau_{x_{1,k}}^{i+1} + (\gamma_{l,k}^{i+1})^2\tau_{x_{2,k}}^{i+1}$
29: $\forall l,k,\ \hat{s}_{l,k}^{i+1} = \tilde{q}_{l,k}^{i+1}$
30: $\tau_s^{i+1} = \frac{1}{4K}\sum_{l=1}^{2}\sum_{k=1}^{2K}\tau_{s_{l,k}}^{i+1}$
31: $i = i + 1$

Until terminated

Gaussian according to the property of the product of Gaussian functions, i.e., $b(z_n) = \mathcal{N}(z_n; \hat{z}_n, v_{z_n})$ with

$$v_{z_n} = (1/\tau_{p_n} + \hat{\epsilon})^{-1} \tag{34}$$

$$\hat{z}_n = v_{z_n}(\hat{\epsilon}r_n + p_n/\tau_{p_n}). \tag{35}$$

Note that $\boldsymbol{\tau}_p$ may contain zero elements. To avoid numerical problems in (34) and (35), they are rewritten (in vector form) as

$$\boldsymbol{\tau}_z = \boldsymbol{\tau}_p./(\mathbf{1} + \hat{\epsilon}\boldsymbol{\tau}_p), \tag{36}$$

$$\hat{\mathbf{z}} = (\hat{\epsilon}\boldsymbol{\tau}_p \cdot \mathbf{r} + \mathbf{p})./(\mathbf{1} + \hat{\epsilon}\boldsymbol{\tau}_p), \tag{37}$$

which are Lines 3 and 4 of Algorithm 1.

From (32) and the Gaussianity of $b(z_n)$, the message $m_{f_{r_n}\to\epsilon}(\epsilon)$ is expressed as

$$m_{f_{r_n}\to\epsilon}(\epsilon) \propto \sqrt{\epsilon}\frac{1}{2}\exp(-\epsilon(|r_n - \hat{z}_n|^2 + v_{z_n}))). \tag{38}$$

According to VMP, the message from function node f_{r_n} to variable node z_n is

$$\begin{aligned} m_{f_{r_n}\to z_n}(z_n) &\propto \exp\left\{\langle \log f_{r_n}(z_n, \epsilon)\rangle_{b(\epsilon)}\right\} \\ &\propto \mathcal{N}\left(z_n; r_n, \hat{\epsilon}^{-1}\right), \end{aligned} \tag{39}$$

where $\hat{\epsilon} = \langle\epsilon\rangle_{b(\epsilon)}$ with

$$\begin{aligned} b(\epsilon) &\propto m_{\epsilon\to f_{r_n}}(\epsilon)m_{f_{r_n}\to\epsilon}(\epsilon) \\ &= f_\epsilon(\epsilon)\prod_n^{2N} m_{f_{r_n}\to\epsilon}(\epsilon) \\ &\propto \epsilon^{N-1}\exp\left\{-\frac{\epsilon}{2}\sum_n\left(|r_n - \hat{z}_n|^2 + v_{z_n}\right)/2\right\} \end{aligned} \tag{40}$$

and

$$m_{\epsilon\to f_{r_n}}(\epsilon) = f_\epsilon(\epsilon)\prod_{n'\neq n} m_{f_{r_{n'}}\to\epsilon}(\epsilon). \tag{41}$$

It is noted that $b(\epsilon)$ follows a Gamma distribution with rate parameter $-\frac{1}{2}\sum_n\left(|r_n - \hat{z}_n|^2 + v_{z_n}\right)$ and shape parameter N, so $\hat{\epsilon} = \langle\epsilon\rangle_{b(\epsilon)}$ can be computed as

$$\hat{\epsilon} = \frac{2N}{\sum_{n=1}^{2N}(|r_n - \hat{z}_n|^2 + v_{z_n})}, \tag{42}$$

which can be rewritten in vector form shown in Line 5 of Algorithm 1. From (39), the Gaussian form of the message $m_{f_{r_n}\to z_n}(z_n)$ suggests the following model

$$r_n = z_n + \omega_n, n = 1, \ldots, 2N, \tag{43}$$

where w_n is a Gaussian noise with mean 0 and variance $\hat{\epsilon}^{-1}$. This fits the forward recursion of the UAMP algorithm with a known noise variance, corresponding to Lines 6 - 9 of Algorithm 1.

According its derivation, UAMP produces the message $m_{s_{l,k}\to f_{s_{l,k}}}(s_{l,k}) \propto \mathcal{N}(s_{l,k}; q_{l,k}, \tau_q)$ with mean $q_{l,k}$ and variance τ_q, which are given in Lines 8 and 9 of Algorithm 1. Next, we need to compute the outgoing message of the function node $f_{s_{l,k}} = \delta(s_{l,k} - f_l(\mathbf{x}_\mathbf{k}'))$. It is noted that the local function is nonlinear with the following expression:

$$f_l(\mathbf{x}_k') = (\mathbf{w}_{l,k}^2)^T g_1(\mathbf{w}_{l,k,1}^1 x_{1,k} + \mathbf{w}_{l,k,2}^1 x_{2,k} + \mathbf{b}_{l,k}^1), \tag{44}$$

where $g_1(\cdot) = Tanh(\cdot)$. The nonlinear function makes the computation of the message $m_{f_{s_{l,k}}\to x_{l,k}}(x_{l,k})$ intractable. To solve this problem, $f_l(\mathbf{x}_k')$ is linearized by using the first order Taylor expansion at the estimate of $\mathbf{x}_k'$ in the last iteration, i.e.,

$$f_l(\mathbf{x}_k') \approx f_l(\hat{\mathbf{x}}_k') + f_l'(\hat{\mathbf{x}}_k')(\mathbf{x}_k' - \hat{\mathbf{x}}_k') \tag{45}$$

with

$$f_l(\hat{\mathbf{x}}_k') = \tilde{q}_{l,k} = (\mathbf{w}_{l,k}^2)^T g_1(\mathbf{W}_{l,k}^1\hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1), \tag{46}$$

which is Line 10 of Algorithm 1, and

$$f_l'(\hat{\mathbf{x}}_k') = \left[\frac{\partial f_l(\hat{\mathbf{x}}_k')}{\partial x_{1,k}}, \frac{\partial f_l(\hat{\mathbf{x}}_k')}{\partial x_{2,k}}\right]^T = [\eta_{l,k}, \gamma_{l,k}]^T, \tag{47}$$

where

$$\eta_{l,k} = \left(\mathbf{w}_{l,k}^2 \cdot \mathbf{w}_{l,k,1}^1\right)^T g_1'(\mathbf{W}_{l,k}^1 \hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1), \tag{48}$$

$$\gamma_{l,k} = (\mathbf{w}_{l,k}^2 \cdot \mathbf{w}_{l,k,2}^1)^T g_1'(\mathbf{W}_{l,k}^1 \hat{\mathbf{x}}_k' + \mathbf{b}_{l,k}^1), \tag{49}$$

which are Lines 11 - 12 of Algorithm 1. In the derivations, we use the property $g_1'(\cdot) = 1 - g_1(\cdot)^2$.

With indexes $l, l' \in \{1, 2\}$, the message $m_{f_{s_{l,k}} \to x_{l',k}}(x_{l',k})$ is computed by the BP rule with messages $m_{s_{l,k} \to f_{s_{l,k}}}(s_{l,k})$ and $\forall l'' \neq l', m_{x_{l'',k} \to f_{s_{l,k}}}(x_{l'',k})$ later computed in (65), yielding

$$\begin{aligned} &m_{f_{s_{l,k}} \to x_{l',k}}(x_{l',k}) \\ &\quad = \langle f_{s_{l,k}}(s_{l,k}, \mathbf{x}_k') \rangle_{m_{s_{l,k} \to f_{s_{l,k}}}(s_{l,k}) m_{x_{l'',k} \to f_{s_{l,k}}}(x_{l'',k})} \\ &\quad \propto \mathcal{N}(x_{l,k}; \psi_{l,k}^{l,l'}, \tau_{\psi_{l,k}}^{l,l'}), \end{aligned} \tag{50}$$

where for $l' = 1$

$$\tau_{\psi_{l,k}}^{l,1} = (\tau_q + \gamma_{l,k}^2 \tau_{x_{2,k}})/\eta_{l,k}^2 \tag{51}$$

$$\psi_{l,k}^{l,1} = (q_{l,k} - \tilde{q}_{l,k})/\eta_{l,k} + \hat{x}_{1,k} \tag{52}$$

and for $l' = 2$

$$\tau_{\psi_{l,k}}^{l,2} = (\tau_q + \eta_{l,k}^2 \tau_{x_{1,k}})/\gamma_{l,k}^2 \tag{53}$$

$$\psi_{l,k}^{l,2} = (q_{l,k} - \tilde{q}_{l,k})/\gamma_{l,k} + \hat{x}_{2,k} \tag{54}$$

given in Lines 13 - 16 of Algorithm 1. The message $m_{x_{l,k} \to f_{x_{l,k}}}(x_{l,k})$ is calculated as

$$\begin{aligned} n_{x_{l,k} \to f_{x_{l,k}}}(x_{l,k}) &= m_{f_{s_{1,k}} \to x_{l,k}}(x_{l,k}) m_{f_{s_{2,k}} \to x_{l,k}}(x_{l,k}) \\ &\propto \mathcal{N}(x_{l,k}; \psi_{l,k}, \tau_{\psi_{l,k}}), \end{aligned} \tag{55}$$

with

$$\tau_{\psi_{l,k}} = (1/\tau_{\psi_{l,k}}^{l,1} + 1/\tau_{\psi_{l,k}}^{l,2})^{-1}, \tag{56}$$

$$\psi_{l,k} = (\psi_{l,k}^{l,1}/\tau_{\psi_{l,k}}^{l,1} + \psi_{l,k}^{l,2}/\tau_{\psi_{l,k}}^{l,2}) \tau_{\psi_{l,k}}, \tag{57}$$

given in Lines 17 and 18 of Algorithm 1.

Note that all the values in the above computations are real as the real parts and imaginary parts of the variables are separated. To facilitate the estimation of the complex-valued symbols, we merge the real and imaginary components:

$$\tau_{\tilde{\psi}_k} = \tau_{\psi_{1,k}} + \tau_{\psi_{2,k}} \tag{58}$$

$$\tilde{\psi}_k = \psi_{1,k} + j\psi_{2,k}, \tag{59}$$

which are shown in Lines 19 and 20 of Algorithm 1.

The prior of x_k is a uniform discrete distribution, i.e.,

$$p(x_k = \lambda_a) = 1/|\mathcal{A}|. \tag{60}$$

It is not hard to show that the *a posteriori* mean $\hat{x}_k$ and variance τ_{x_k} of x_k are given by (also shown in Lines 21 - 24 of Algorithm 1)

$$\hat{x}_k = \sum_{a=1}^{|A|} \lambda_a \mu_{k,a} \tag{61}$$

$$\tau_{x_k} = \sum_{a=1}^{|A|} \mu_{k,a} |\lambda_a - \hat{x}_k|^2, \tag{62}$$

where

$$\mu_{k,a} = \xi_{k,a} / \sum_{a=1}^{|A|} \xi_{k,a}, \tag{63}$$

with

$$\xi_{k,a} = \exp(-\tau_{\tilde{\psi}_k}^{-1} |\lambda_a - \tilde{\psi}_k|^2). \tag{64}$$

To simplify the message computations, we use the following approximation:

$$m_{x_{l,k} \to f_{s_{1,k}}} = m_{x_{l,k} \to f_{s_{2,k}}} = m_{f_{x_k} \to x_{l,k}}. \tag{65}$$

Since the *a posteriori* mean $\hat{x}_k$ of x_k are updated in (61), we update $f_l(\mathbf{x}_k')$ (including $\tilde{q}_{l,k}$, $\eta_{l,k}$ and $\gamma_{l,k}$) in (45) with the updated $\hat{x}_k$. This is Line 25 of Algorithm 1.

To compute the message $m_{f_{s_{l,k}} \to s_{l,k}}$, we separate the real part and imaginary part of x_k and assume that they have the same variance, so

$$\tau_{x_{1,k}} = \tau_{x_{2,k}} = \tau_{x_k}/2 \tag{66}$$

$$\hat{x}_{1,k} = \Re\{\hat{x}_k\}, \hat{x}_{2,k} = \Im\{\hat{x}_k\}, \tag{67}$$

which are Lines 26 and 27 of Algorithm 1. Then, we are ready to compute the message from $f_{s_{l,k}}$ to $s_{l,k}$, i.e.,

$$\begin{aligned} m_{f_{s_{l,k}} \to s_{l,k}}(s_{l,k}) &= \langle f_{s_{l,k}}(s_{l,k}, \mathbf{x}_k') \rangle_{\prod_{l'} m_{x_{l',k} \to f_{s_{l,k}}}(x_{l',k})} \\ &\propto \mathcal{N}(s_{l,k}; \overleftarrow{s}_{l,k}, \overleftarrow{\tau}_{s_{l,k}}), \end{aligned} \tag{68}$$

with

$$\overleftarrow{\tau}_{s_{l,k}} = \eta_{l,k}^2 \tau_{x_{1,k}} + \gamma_{l,k}^2 \tau_{x_{2,k}} \tag{69}$$

$$\overleftarrow{s}_{l,k} = \tilde{q}_{l,k}, \tag{70}$$

which are Lines 28 and 29 of Algorithm 1. According to UAMP version 2 [33], an averaged variance is required, i.e.,

$$\tau_s = \frac{1}{2K} \sum_{l=1}^{2} \sum_{k=1}^{2K} \tau_{s_{l,k}}, \tag{71}$$

which is Line 30 of Algorithm 1. This is the end of a single round iteration of the iterative process. A number of iterations are performed until the algorithm converges, or the algorithm is terminated when a pre-set number of iterations is reached.

C. Complexity Analysis

We analyze the complexity of the proposed detector and the benchmark detectors, including the DNN-based direct detector and RMP-based direct detector. Only multiplications/divisions are counted in the analysis.

For training of the proposed detector, the back-propagation consists of a forward phase for signal flow calculation and a backward phase for gradient calculation. The forward phase requires $8KN' + 2KN$ operations for a single training sample,

and the backward phase requires $4K^3N' + (N')^3N$ operations with tied weights. Hence, for $\mathcal{L}$ iterations of gradient descent with M_0 training samples, the proposed detector has the complexity $\mathcal{O}(M_0\mathcal{L}(KN' + KN + K^3N' + (N')^3N))$. Similarly, for the DNN based direct detector, the complexity for training is $\mathcal{O}(M_0\mathcal{L}KNN_1N_2)$, where N_1 and N_2 are the number of nodes of the first and second hidden layers. As N_1 and N_2 are normally much larger than K, N and N', the proposed detector has less training complexity than the DNN-based direct detector. The training of the RMP-based direct detector requires a complexity of $\mathcal{O}((NP(L+1))^3 + (NP(L+1))^2M_0)$, where matrix inversion is required to determine the polynomial coefficients (the number of coefficients can be very large). For the proposed detector, the training length M_0 can be greatly reduced and it converges fast with less hyperparameters. Compared to the RMP-based direct detector (with zero memory length), the proposed detector may require higher complexity in training, but the proposed detector delivers remarkably better performance as shown in Section V.

For signal detection, the proposed detector performs UAMP on the trained NN. The complexity of UAMP is dominated by two matrix-vector product operations in Line 2 and Line 9 of Algorithm 1, resulting in $\mathcal{O}(KN)$ per iteration. Besides, the determination of $\tilde{q}^i_{l,k}$, $\eta^i_{l,k}$ and $\gamma^i_{l,k}$ in Lines 11-13 of Algorithm 1 also involves matrix-vector product operations with $\mathcal{O}(KN')$ per iteration. An SVD with complexity $\mathcal{O}(\min(K^2N, N^2K))$ is also needed, but it only needs to be computed once, and can be shared by the whole symbol sequence with length L_s. Therefore, with I_c iterations, the complexity of the proposed detector is $\mathcal{O}(I_cKN + I_cKN' + \min(K^2N, N^2K)/L_s)$. The proposed detector has a fast convergence speed (it requires about $I_c = 10$ iterations to converge according to the simulations in Section V). The DNN-based detector requires a complexity of $\mathcal{O}(NN_1 + N_1N_2 + N_2K)$, and the RMP-based detector requires a complexity of $\mathcal{O}(KNP(L+1))$. So, the complexity of proposed detector can be slightly larger, but it performs remarkably better than the benchmark detectors as shown in Section V.

IV. Extension To Coded System With Turbo Receiver

In a turbo receiver, the detector and decoder work in an iterative manner to achieve joint detection and decoding. It is well known that a turbo receiver is much more powerful than a conventional non-iterative receiver [41], [42]. Compared to the direct detectors, the proposed Bayesian detector is readily extended to a SISO detector so that a turbo receiver can be implemented. In a turbo system, the information bits are firstly encoded and then interleaved before mapping. Each symbol $x_k \in \mathcal{A} = [\lambda_1, \ldots, \lambda_{|\mathcal{A}|}]$ is mapped from a sub-sequence of the coded bit sequence, which is denoted by $\mathbf{u}_k = [u_k^1, \ldots, u_k^{\log|\mathcal{A}|}]$. Each λ_a corresponds to a length-$\log|\mathcal{A}|$ binary sequence denoted by $\{\lambda_a^1, \ldots, \lambda_a^{\log|\mathcal{A}|}\}$.

The turbo receiver is shown in Fig. 4, which consists of the UAMP-based Bayesian detector and a SISO decoder, working in an iterative manner to exchange extrinsic log-likelihood ratios (LLRs) of the coded bits. For simplicity, a single user is assumed in Fig. 4. The detector calculates the extrinsic LLRs for each coded bit with the extrinsic LLRs from the decoder as the *a priori* information. Then, with the extrinsic LLRs from the detector, the decoder refines the LLRs with the code constraints. In this work, we assume a standard SISO decoder (e.g., the Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm for convolutional codes) is employed, and we adapt the detector proposed in Section III to a SISO one.

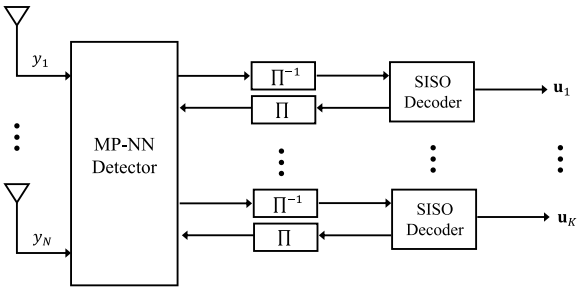


Fig. 4. Block diagram of turbo receiver, where Π and Π^{-1} denote an interleaver and the corresponding deinterleaver, respectively.

The task of the detector is to calculate the extrinsic LLR for each code bit $u_k^q(m)$, which is represented as

$$L^e(u_k^q) = \ln\left(\frac{p(u_k^q = 0|\mathbf{r})}{p(u_k^q = 1|\mathbf{r})}\right) - L^a(u_k^q), \tag{72}$$

where $L^a(u_k^q)$ is the output extrinsic LLR of the decoder in the previous iteration. The extrinsic LLR $L^e(u_k^q)$ is passed to the decoder. The derivation for $L^e(u_k^q)$ in terms of extrinsic mean and variance can be found in [43], and $L^e(u_k^q)$ can be expressed as

$$L^e(u_k^q) = \ln \frac{\sum\limits_{\lambda_a \in \mathcal{A}_q^0} \exp\left(-\frac{|\lambda_a - m^e_{x_k}|^2}{v^e_k}\right) \prod\limits_{q' \neq q} p(u_k^{q'} = \lambda_a^{q'})}{\sum\limits_{\lambda_a \in \mathcal{A}_q^1} \exp\left(-\frac{|\lambda_a - m^e_{x_k}|^2}{v^e_k}\right) \prod\limits_{q' \neq q} p(u_k^{q'} = \lambda_a^{q'})}, \tag{73}$$

where $\mathcal{A}_q^0$ and $\mathcal{A}_q^1$ denote subsets of all $\lambda_a \in D$ whose label in position q has the value of 0 and 1, respectively, and m_k^e and v_k^e are the extrinsic mean and variance of x_k. According to [43], the extrinsic variance and mean are defined as

$$v_k^e = \left(\frac{1}{v_k^p} - \frac{1}{v_k}\right)^{-1} \tag{74}$$

$$m^e_{x_k} = v^e_{x_k}\left(\frac{m^p_{x_k}}{v^p_{x_k}} - \frac{m_{x_k}}{v_{x_k}}\right), \tag{75}$$

where m_{x_k} and v_{x_k} are the *a priori* mean and variance of x_k calculated based on the output LLRs of the SISO decoder [41], [42], [44], and $m^p_{x_k}$ and $v^p_{x_k}$ are the *a posteriori* mean and variance of x_k. By examining the derivation of the Bayesian detector in Algorithm 1, we can find that $\tilde{\psi}_k$ and $\tau_{\tilde{\psi}_k}$ consist of the extrinsic mean and variance of x_k as they are the messages passed from observation and do not contain the immediate *a priori* information about x_k. Therefore, we have

$$m^e_{x_k} = \tilde{\psi}_k, \quad v^e_{x_k} = \tau_{\tilde{\psi}_k}. \tag{76}$$

Then, (73) can be readily used to calculate the extrinsic LLRs of the coded bits. Note that with the LLRs output from the

SISO decoder, we can compute the probability $p(x_k = \lambda_a)$ for each x_k, which is no longer $1/|\mathcal{A}|$ in Algorithm 1. Therefore, $\xi_{k,a}$ in Line 21 of Algorithm 1 needs to be changed to

$$\xi_{k,a} = p(x_k = \lambda_a) \exp(-v_{\psi_k}^{-1} |\lambda_a - \tilde{\psi}_k|^2). \tag{77}$$

In addition, we note that the iteration of the detector can be incorporated into the iteration between the SISO decoder and detector, i.e., only a single loop iteration (without inner iteration) is needed.

V. Simulation Results

Assume that the BS is equipped with a uniform linear antenna array, and in the simulations we set $N = 10$ and $K = 5$ (unless specified). The modulation scheme used is 16-QAM. The Saleh-Valenzuela channel model [45] is employed. The channel vector $\mathbf{h}_k$ between the kth user and the N receive antennas is represented as

$$\mathbf{h}_k = \sqrt{\frac{N}{Q_k}} \sum_{q=1}^{Q_k} \beta_{kq} \mathbf{a}(\theta_{kq}), \tag{78}$$

where θ_{kq} is the incident angle of the qth path, $\mathbf{a}(\theta_{kq}) = \frac{1}{\sqrt{N}}[1, e^{-j2\pi d sin(\theta_{kq})/\lambda}, \ldots, e^{-j2\pi d sin(\theta_{kq})(N-1)/\lambda}]^T$ is a length-N steering vector with antenna spacing d, λ is the wavelength of carrier, Q_k is the number of paths for user k, and β_{kq} is the complex gain of the qth path. We use the same parameter settings as in [46], where $d = \lambda/2$, $Q = 3$, β_{kq} follows Gaussian distribution with zero mean and unity variance, and θ_{kq} is uniformly drawn from $(-0.5\pi, 0.5\pi]$. As in [35] and [47], the parameters used for the PA nonlinearity are $\alpha_a = 4.65$, $\alpha_\phi = 2560$, $\beta_\phi = 0.114$, $\sigma_a = 0.81$, $x_{\text{sat}} = 0.58$, $q_1 = 2.4$ and $q_2 = 2.3$. For I/Q imbalance, the parameters are $\theta_k = 4^\circ$ and $\lambda_k = 0.05$. The SNR is defined as P_x/σ_n^2, where P_x is the power of the transmitted signal of a user (assuming all users have the same transmit power), and σ_n^2 is the power of the noise (per receive antenna) at the receiver. We compare the proposed detector called MP-NN, where the parameters of the sub-NN pairs are tied, with existing detectors, including DNN based direct detector [26] and RMP-based direct detector [34], which are called D-DNN and D-RMP, respectively.

The deep learning framework *Tensorflow* is used for (D)NN training and validation. Batch gradient descent is adopted, and cross-validation is used to avoid overfitting and ensure the generality of the trained model. We use 80% and 20% of the dataset for training (including 3-fold validation) and testing. Through the validation data set, we determine that the batch size is 100 and the number of epochs is 300. The Adam optimizer with a learning rate 0.01 is employed to update the (D)NN parameters. For the proposed NN, the number of hidden nodes N' in the sub-NNs is 20. For D-DNN, the activation function $Tanh$ is employed for hidden layers. The number of hidden layers is 2, and the numbers of nodes of the hidden layers are 30 and 40, unless these parameters are specified. For D-RMP, a fifth order polynomial is employed and the memory length L is set to 0.

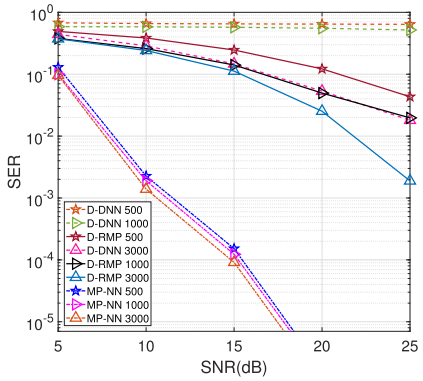


Fig. 5. SER performance comparisons of MP-NN, D-DNN and D-RMP based detectors with different training lengths.

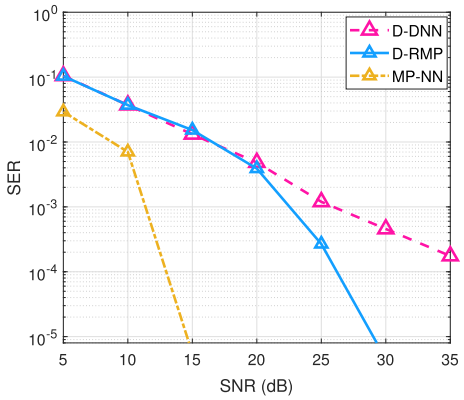


Fig. 6. SER performance comparisons of MP-NN, D-DNN and D-RMP based detectors with K=16 and N=32.

A. Uncoded System

We first consider a uncoded system. Fig. 5 shows the symbol error rate (SER) of the detectors, where the training lengths 500, 1000 and 3000 are used to examine the impact of training length on the performance of the detectors. From the results, we can see that in all the cases, the proposed MP-NN detector always performs remarkably better than other detectors. We can also see that D-RMP performs better than D-DNN. Moreover, when the training length is decreased from 3000 to 500, there are only minor changes in the performance of MP-NN, which indicates that the training length 500 is sufficient for MP-NN. In contrast, the impact of the training length on the performance of D-RMP and D-DNN is significant, and their performance degrades rapidly with the reduce of the training length. These results demonstrate the effectiveness of the proposed detector, i.e., it can be trained more effectively and the Bayesian detector is much more powerful. Considering that neither D-RMP nor D-DNN works well with training lengths 500 and 1000, we use training length 3000 in the subsequent simulations. We also examine the SER performance of the proposed detector with a larger number of users and receive antennas. The results are shown in Fig. 6, where the number of users K is increased to 16, the number of receive antennas N is increased to 32, and the other setups remain the same as those in Fig. 5 (with training length 3000). It can be seen that the proposed MP-NN detector still achieves significant performance gain compared to the D-DNN and D-RMP detectors.

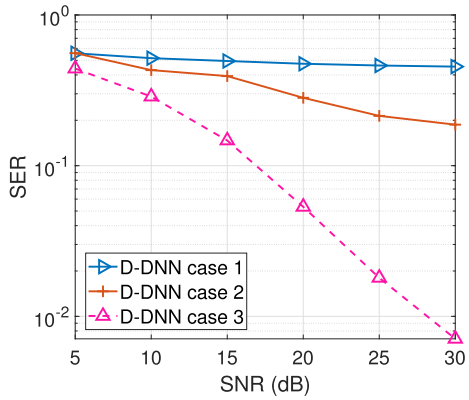


Fig. 7. SER performance comparisons of D-DNN with different hyper-parameters.

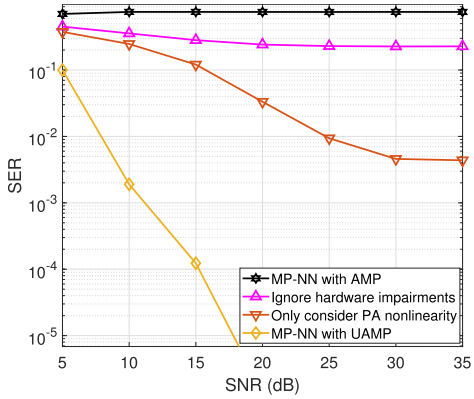


Fig. 8. SER performance of the MP-NN detector, the receiver without handling I/Q imbalance, and the receiver without handling nonlinearity and I/Q imbalance.

With the training length fixed to 3000, we examine the performance of D-DNN by changing its hyper-parameters including the number of layers and hidden nodes, which are indicated by cases 1, 2 and 3. In case 1, the number of hidden layers is 2, we increase the number of hidden nodes in the two hidden layers to 300 and 400, respectively. In case 2, we increase the number of hidden layers to 3 with hidden nodes 30, 40 and 50, respectively. In case 3, we use the default setup as before. The results are shown in Fig. 7. It can be seen that, compared to the default hyper-parameter setting (case 3), the SER performance of D-DNN deteriorates significantly with other settings. This is because the number of parameters for the DNN is increased significantly in cases 1 and 2, and the training samples are insufficient. Hence in the subsequent examples, we will use the default setting for D-DNN.

It is mentioned in the previous section that, UAMP plays a crucial role in the message passing based Bayesian detector MP-NN. To demonstrate this, we also use AMP to deal with the densely connected part of the factor graph (i.e., AMP is integrated into the message passing algorithm). We compare the SER performance of the detector with AMP and UAMP in Fig. 8. We can see that the AMP based detector simply does not work as the AMP algorithm does not converge. To demonstrate that it is necessary to handle the I/Q imbalance and PA nonlinearity at the receiver side, we compare

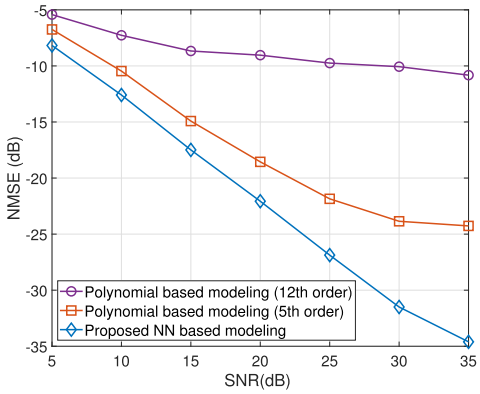


Fig. 9. Modeling performance comparison of the proposed NN and polynomial methods with 5th and 12th orders, respectively.

the MP-NN receiver with the receiver without considering I/Q imbalance and nonlinearity, where the zero-forcing (ZF) detector with known MIMO channel matrix is employed. We also compared the proposed receiver with the receiver without considering I/Q imbalance, where polynomial based detector is employed to handle PA nonlinearity. The results are also shown in Fig. 8. It can be seen that, without considering both I/Q imbalance and PA nonlinearity, the receiver simply does not work properly. If only PA nonlinearity is considered, the receiver performs poorly and a very high SER floor is observed. The results indicate that both I/Q imbalance and PA nonlinearity need to be properly handled by the receiver to achieve good performance.

As discussed in the previous section, the architecture of the NN proposed in this work is designed based on signal flow to model the joint effects of hardware impairments and co-channel interference. We note that the polynomial techniques [34] can also be used to model the joint effects. It is interesting to compare the performance of the two methods. We use the normalized mean square error (NMSE) to evaluate the modelling performance and the results are shown in Fig. 9, where polynomials with the 5th and 12th order are used. We note that, although the use of higher order polynomial may improve the modelling capability of the polynomial technique, it causes difficulties in determining the polynomial parameters due to numerical instability. Moreover, it is noted that when the order of polynomial increases by one, the number of parameters to be determined is increased by $4KN(L+1)$, which is a significant increase, making it prone to overfitting due to the limited number of training samples. As shown in Fig. 9, the proposed NN significantly outperforms the 5th-order polynomial, indicating that the proposed NN has much better modeling capability. When increasing the polynomial order to 12, the performance of the polynomial method becomes extremely poor due to numerical instability and overfitting. The results demonstrate the advantage of the proposed NN in modelling.

So far, we have compared the performance of the receivers with moderate hardware imperfections. It is also interesting to test the capabilities of the receivers in handling severer hardware imperfections. According to [48], we increase the gain α_a of amplitude to amplitude conversion to 6.5 to simulate

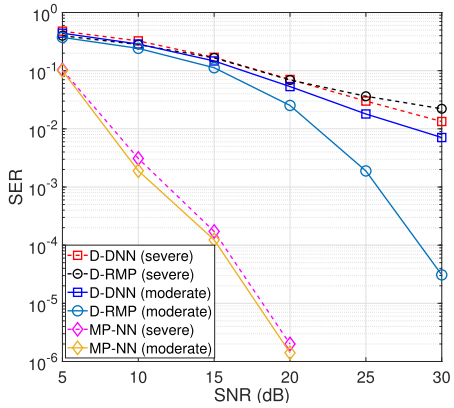


Fig. 10. SER performance comparisons of the receivers with moderate and severe hardware imperfections.

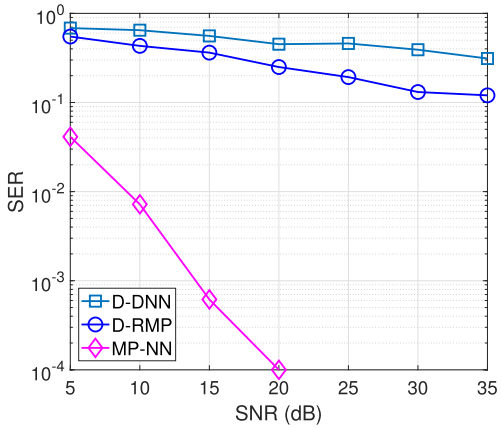


Fig. 11. SER performance comparison with extreme I/Q imbalance and PA nonlinear distortion.

severer PA nonlinearity. Fig. 10 shows the SER performance of the receivers. It can be seen that the performance of D-RMP and D-DDN deteriorate significantly with severer hardware imperfections. In contrast, the proposed MP-NN receiver only incurs marginal performance loss, and it still delivers outstanding performance. We also adjust the I/Q imbalance and PA nonlinearity to an extreme condition. The PA nonlinearity is simulated using a fifth-order polynomial in [16]. The I/Q imbalance parameter θ_k is increased to 10°. The results are shown in Fig. 11, where we can see that D-RMP and D-DNN simply do not work under the extreme hardware imperfections. In contrast, the proposed MP-NN detector still performs very well. These results demonstrate the high capability of MP-NN to deal with hardware distortions.

B. Coded System

We then evaluate the performance of the detectors in a coded system, and compare the performance of the systems with and without turbo receiver. We use a rate-2/3 convolutional code with generators [23], [35], followed by a random interleaver and 16-QAM modulation, where Gray mapping is used in symbol mapping. The BCJR algorithm is used to implement the SISO decoder. As it is unknown how to implement a turbo receiver based on the direct detectors D-RMP and D-DNN, so non-iterative receivers are implemented for them, where the outputs of detectors after hard decision are fed to a Viterbi decoder. The other settings are the same as those in the previous section, and the bit error rate (BER) is used to evaluate the performance of the receivers. We compare the performance of the MP-NN turbo receiver, D-RMP receiver and D-DNN receiver in the coded system. Fig. 12 shows the BER performance of the receivers. We can see that the proposed MP-NN detector performs significantly better than other receivers. Similar to the previous results, the D-RMP receiver performs slightly better than the D-DNN receiver.

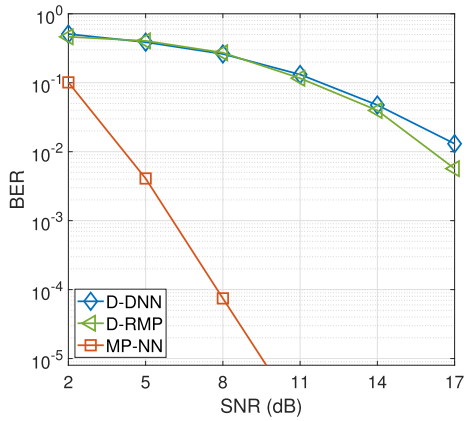


Fig. 12. BER performance of MP-NN, D-RMP and D-DNN receivers in a coded system.

VI. Conclusion

In this work, we developed a Bayesian detector for MIMO communications with combined hardware imperfections. Based on the signal flow, we first design the architecture of an NN to model the hardware imperfections and multi-user interference, so that the NN can be trained much more efficiently, compared to conventional DNN-based methods. Then, representing the trained NN as a factor graph and leveraging UAMP, we develop an efficient message passing based Bayesian detector MP-NN. Both non-iterative receiver and turbo receiver are investigated. Extensive simulation results demonstrate that the proposed method significantly outperforms state-of-the-art algorithms.

By combining NN and factor graph techniques, this work provides a general way to achieve Bayesian signal detection for a communication system with complicated input-output relationship. Interestingly, a recent work in [49] also combines NNs and factor graphs for stationary time sequence inference. However, the ways of combining NNs and factor graphs in this work and [49] are very different. Here, NNs are represented as factor graphs to develop efficient message passing algorithms for Bayesian inference, where message passing is carried out on NNs. In [49], NNs are used to learn specific components of a factor graph describing the distribution of the time sequence, where NNs are involved in the computation of local messages. Combining NNs and factor graphs is promising to tackle challenging signal processing tasks, which is worth further exploration.

REFERENCES

[1] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[2] Y. Wu, Y. Gu, and Z. Wang, "Channel estimation for mmWave MIMO with transmitter hardware impairments," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 320–323, Feb. 2018.

[3] A. Chung, M. Ben Rejeb, Y. Beltagy, A. M. Darwish, H. A. Hung, and S. Boumaiza, "IQ imbalance compensation and digital predistortion for millimeter-wave transmitters using reduced sampling rate observations," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 7, pp. 3433–3442, Jul. 2018.

[4] C. Qi, K. Chen, O. A. Dobre, and G. Y. Li, "Hierarchical codebook-based multiuser beam training for millimeter wave massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8142–8152, Dec. 2020.

[5] X. Cheng, Y. Yang, and S. Li, "Joint compensation of transmitter and receiver I/Q imbalances for SC-FDE systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8483–8498, Aug. 2020.

[6] C. Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Trans. Signal Process.*, vol. 45, no. 1, pp. 223–227, Jan. 1997.

[7] C. Yu, L. Guan, E. Zhu, and A. Zhu, "Band-limited Volterra series-based digital predistortion for wideband RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 12, pp. 4198–4208, Dec. 2012.

[8] L. Ding et al., "A robust digital baseband predistorter constructed using memory polynomials," *IEEE Trans. Commun.*, vol. 52, no. 1, pp. 159–165, Jan. 2004.

[9] L. Ding, R. Raich, and G. T. Zhou, "A Hammerstein predistortion linearization design based on the indirect learning architecture," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 3, 2002.

[10] F. M. Ghannouchi and O. Hammi, "Behavioral modeling and predistortion," *IEEE Microw. Mag.*, vol. 10, no. 7, pp. 52–64, Dec. 2009.

[11] J. Zheng, J. Zhang, L. Zhang, X. Zhang, and B. Ai, "Efficient receiver design for uplink cell-free massive MIMO with hardware impairments," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4537–4541, Apr. 2020.

[12] L. Ding, Z. Ma, D. R. Morgan, M. Zierdt, and G. Tong Zhou, "Compensation of frequency-dependent gain/phase imbalance in predistortion linearization systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 1, pp. 390–397, Feb. 2008.

[13] H. Cao, A. Soltani Tehrani, C. Fager, T. Eriksson, and H. Zirath, "I/Q imbalance compensation using a nonlinear modeling approach," *IEEE Trans. Microw. Theory Techn.*, vol. 57, no. 3, pp. 513–518, Mar. 2009.

[14] R. Mahendra, S. K. Mohammed, and R. K. Mallik, "Transmitter IQ imbalance pre-compensation for mm-Wave hybrid beamforming systems," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–7.

[15] R. Raich, H. Qian, and G. T. Zhou, "Orthogonal polynomials for power amplifier modeling and predistorter design," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1468–1479, Sep. 2004.

[16] W. Zhao et al., "Orthogonal polynomial-based nonlinearity modeling and mitigation for LED communications," *IEEE Photon. J.*, vol. 8, no. 4, pp. 1–12, Aug. 2016.

[17] W. Zhao, Q. Guo, J. Tong, J. Xi, Y. Yu, and P. Niu, "Frequency domain equalization and post distortion for LED communications with orthogonal polynomial based joint LED nonlinearity and channel estimation," *IEEE Photon. J.*, vol. 10, no. 4, pp. 1–11, Aug. 2018.

[18] J. Zhang, Y. He, Y. Li, C. Wen, and S. Jin, "Meta learning-based MIMO detectors: Design, simulation, and experimental test," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1122–1137, Feb. 2021.

[19] C. Qi, P. Dong, W. Ma, H. Zhang, Z. Zhang, and G. Y. Li, "Acquisition of channel state information for mmWave massive MIMO: Traditional and machine learning-based approaches," *Sci. China Inf. Sci.*, vol. 64, no. 8, pp. 1–16, Aug. 2021.

[20] W. Ma, C. Qi, Z. Zhang, and J. Cheng, "Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2838–2849, May 2020.

[21] D. Gao and Q. Guo, "Extreme learning machine-based receiver for MIMO LED communications," *Digit. Signal Process.*, vol. 95, Dec. 2019, Art. no. 102594. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200419301484

[22] D. Gao, Q. Guo, and Y. C. Eldar, "Massive MIMO as an extreme learning machine," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 1046–1050, Jan. 2021.

[23] D. Gao, Q. Guo, M. Jin, Y. Yu, and J. Xi, "Adaptive extreme learning machine-based nonlinearity mitigation for LED communications," *IEEE J. Sel. Topics Quantum Electron.*, vol. 27, no. 2, pp. 1–9, Mar. 2021.

[24] T. Liu, S. Boumaiza, and F. M. Ghannouchi, "Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.

[25] D. Wang, M. Aziz, M. Helaoui, and F. M. Ghannouchi, "Augmented real-valued time-delay neural network for compensation of distortions and impairments in wireless transmitters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 242–254, Jan. 2019.

[26] P. Jaraut, M. Rawat, and F. M. Ghannouchi, "Composite neural network digital predistortion model for joint mitigation of crosstalk, I/Q imbalance, nonlinearity in MIMO transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 11, pp. 5011–5020, Nov. 2018.

[27] Y. Wu, U. Gustavsson, A. G. I. Amat, and H. Wymeersch, "Residual neural networks for digital predistortion," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[28] R. J. Thompson and X. Li, "Integrating Volterra series model and deep neural networks to equalize nonlinear power amplifiers," in *Proc. 53rd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2019, pp. 1–6.

[29] H. Liu, X. Yang, P. Chen, M. Sun, B. Li, and C. Zhao, "Deep learning based nonlinear signal detection in millimeter-wave communications," *IEEE Access*, vol. 8, pp. 158883–158892, 2020.

[30] F. Caltagirone, L. Zdeborová, and F. Krzakala, "On convergence of approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 1812–1816.

[31] Q. Guo and J. Xi, "Approximate message passing with unitary transformation," 2015, *arXiv:1504.04799*.

[32] M. Luo, Q. Guo, M. Jin, Y. C. Eldar, D. Huang, and X. Meng, "Unitary approximate message passing for sparse Bayesian learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 6023–6039, 2021.

[33] Z. Yuan, Q. Guo, and M. Luo, "Approximate message passing with unitary transformation for robust bilinear recovery," *IEEE Trans. Signal Process.*, vol. 69, pp. 617–630, 2021.

[34] Z. A. Khan, E. Zenteno, P. Händel, and M. Isaksson, "Digital predistortion for joint mitigation of I/Q imbalance and MIMO power amplifier distortion," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 1, pp. 322–333, Jan. 2017.

[35] E. Perahia, "IEEE p802.11 wireless LANs TGad evaluation methodology," *IEEE Standards Assoc.*, vol. 29, pp. 9–15, 2010.

[36] G.-B. Huang and H. A. Babri, "Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions," *IEEE Trans. Neural Netw.*, vol. 9, no. 1, pp. 224–229, Dec. 1998.

[37] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: Perceptron, madaline, and backpropagation," *Proc. IEEE*, vol. 78, no. 9, pp. 1415–1442, Jun. 1990.

[38] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[39] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[40] J. Winn, C. M. Bishop, and T. Jaakkola, "Variational message passing," *J. Mach. Lear. Res.*, vol. 6, no. 4, pp. 1–34, 2005.

[41] M. Tuchler, A. C. Singer, and R. Koetter, "Minimum mean squared error equalization using a priori information," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 673–683, Mar. 2002.

[42] Q. Guo and L. Ping, "LMMSE turbo equalization based on factor graphs," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 311–319, Mar. 2008.

[43] Q. Guo and D. D. Huang, "A concise representation for the soft-in soft-out LMMSE detector," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 566–568, May 2011.

[44] B. Vucetic and J. Yuan, *Turbo Codes: Principles and Applications*, vol. 559. Cham, Switzerland: Springer, 2012.

[45] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[46] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 3679–3684.

[47] B. Li, C. Zhao, M. Sun, H. Zhang, Z. Zhou, and A. Nallanathan, "A Bayesian approach for nonlinear equalization and signal detection in millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3794–3809, Jul. 2015.

[48] L. Cho, X. Yu, C. Hsu, and P. Ho, "Mitigation of PA nonlinearity for IEEE 802.11ah power-efficient uplink via iterative subcarrier regularization," *IEEE Access*, vol. 9, pp. 15659–15669, 2021.

[49] N. Shlezinger, N. Farsad, Y. C. Eldar, and A. J. Goldsmith, "Learned factor graphs for inference from stationary time sequences," *IEEE Trans. Signal Process.*, vol. 70, pp. 366–380, 2022.

Dawei Gao (Member, IEEE) received the B.E. (Hons.) and Ph.D. degrees in telecommunications engineering from the University of Wollongong, Wollongong, NSW, Australia, in 2016 and 2020, respectively. He is currently a Lecturer with the Hangzhou Institute of Technology, Xidian University, Hangzhou, Zhejiang, China. His research interests include machine learning, array signal processing, and joint sensing and communications.

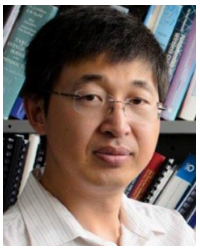
Qinghua Guo (Senior Member, IEEE) received the B.E. degree in electronic engineering and the M.E. degree in signal and information processing from Xidian University, Xi'an, China, in 2001 and 2004, respectively, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, SAR, China, in 2008. He is currently an Associate Professor with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia, and an Adjunct Associate Professor with the School of Engineering, The University of Western Australia, Perth, WA, Australia. His research interests include signal processing, telecommunications, radar, and optical sensing. He was a recipient of the Australian Research Council's Inaugural Discovery Early Career Researcher Award.

Guisheng Liao (Senior Member, IEEE) was born in Guilin, Guangxi, China, in 1963. He received the B.S. degree in mathematics from Guangxi University, Guangxi, in 1985, and the M.S. degree in computer science and the Ph.D. degree in electrical engineering from Xidian University, Xi'an, China, in 1990 and 1992, respectively. From 1999 to 2000, he was a Senior Visiting Scholar with The Chinese University of Hong Kong, Hong Kong. Since 2006, he has been the panelists for the medium and long term development plan in high-resolution and remote sensing systems. Since 2007, he has been the Lead of the Yangtze River Scholars Innovative Team and devoted in advanced techniques in signal and information processing. Since 2009, he has been the Evaluation Expert for the International Cooperation Project of the Ministry of Science and Technology in China. He is currently a Yangtze River Scholars Distinguished Professor with the National Laboratory of Radar Signal Processing and the Dean of the Hangzhou Institute of Technology, Xidian University. His research interests include signal processing, space-time adaptive processing, radar waveform design, and airborne/space surveillance and warning radar systems.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002.

She was a Professor with the Department of Electrical Engineering, Technion. She was a Visiting Professor with Stanford University. She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, and an Adjunct Professor with Duke University. She is the author of the book *Sampling Theory: Beyond Bandlimited Systems* and the coauthor of five other books published by Cambridge University Press. Her research interests include statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging, and optics.

Dr. Eldar was elected as a member of the Israel Academy of Sciences and Humanities, in 2017. She is an EURASIP Fellow. She was a Horev Fellow of the Leaders in Science and Technology Program, Technion, and an Alon Fellow. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She was a member of the IEEE Signal Processing Theory and Methods Technical Committee and the Bio Imaging and Signal Processing Technical Committee. She is a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. She has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award in 2013, the IEEE/AESS Fred Nathanson Memorial Radar Award in 2014, and the IEEE Kiyo Tomiyasu Award in 2016. She received the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel and David Jacknow Award for Excellence in Teaching, and the Technion Award for Excellence in Teaching (two times). She received several best paper awards and best demo awards together with her research students and colleagues, including the SIAM Outstanding Paper Prize, the UFFC Outstanding Paper Award, the Signal Processing Society Best Paper Award, and the IET Circuits, Devices and Systems Premium Award. She was selected as one of the 50 most influential women in Israel and Asia. She was the co-chair and the technical co-chair of several international conferences and workshops. She served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*. She is the Editor-in-Chief of *Foundations and Trends in Signal Processing*. In the past, she was a Signal Processing Society Distinguished Lecturer. She is a Highly Cited Researcher.



Yonghui Li (Fellow, IEEE) received the Ph.D. degree from the Beijing University of Aeronautics and Astronautics in November 2002.

Since 2003, he has been with the Centre of Excellence in Telecommunications, The University of Sydney, Australia. He is currently a Professor and the Director of the Wireless Engineering Laboratory, School of Electrical and Information Engineering, The University of Sydney. His current research interests include wireless communications, with a particular focus on MIMO, millimeter wave communications, machine to machine communications, coding techniques, and cooperative communications. He holds a number of patents granted and pending in these fields. He was a recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012. He received the Best Paper Award from the IEEE International Conference on Communications (ICC) 2014, IEEE PIRMC 2017, and the IEEE Wireless Days Conferences (WD) 2014. He was an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He also served as the Guest Editor for several IEEE journals, such as IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *IEEE Communications Magazine*, IEEE INTERNET OF THINGS JOURNAL, and IEEE ACCESS.

Yanguang (Sunny) Yu (Senior Member, IEEE) received the B.E. degree from the Huazhong University of Science and Technology, China, in 1986, and the Ph.D. degree from the Harbin Institute of Technology, China, in 2000. She was with the College of Information Engineering, Zhengzhou University, China, on various appointments, including a Lecturer from 1986 to 1999, an Associate Professor from 2000 to 2004, and a Professor from 2005 to 2007. From 2001 to 2002, she was a Post-Doctoral Fellow with the Opto-Electronics Information Science and Technology Laboratory, Tianjin University, China. She also had a number of visiting appointments, including a Visiting Fellow with the Optoelectronics Group, Department of Electronics, University of Pavia, Italy, from 2002 to 2003, a Principal Visiting Fellow with the University of Wollongong, Australia, from 2004 to 2005, and a Visiting Associate Professor and a Professor with the Engineering School ENSEEIHT, Toulouse, France, in 2004 and 2006, respectively. She joined the University of Wollongong in 2007, where she is currently an Associate Professor with the School of Electrical, Computer and Telecommunications Engineering. Her research interests include semiconductor lasers with optical feedback and their applications in sensing and instrumentations, secure chaotic communications, and signal processing and its applications to 3-D profile measurement and telecommunication systems.

Branka Vucetic (Life Fellow, IEEE) is currently an Australian Laureate Fellow, a Professor of telecommunications, and the Director of the Centre for IoT and Telecommunications, The University of Sydney. Her current research work is in wireless networks and Industry 5.0. In the area of wireless networks, she works on communication system design for 6G and wireless AI. In the area of Industry 5.0, her research is focused on the design of cyber-physical-human systems and wireless networks for applications in healthcare, energy grids, and advanced manufacturing. She is a fellow of the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science.