EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Low-complexity deep unfolded neural network receiver for MIMO systems based on the probability data association detector

Pedro H. C. de Souza* and Luciano L. Mendes*

*Correspondence:
pedro.carneiro@dtel.inatel.br;
luciano@inatel.br

National Institute of Telecommunications -INATEL, Av. João de Camargo, 510 - Centro, Santa Rita do Sapucaí 37540-000, Brazil

Abstract

The interest on applications where machine learning algorithms and communications are combined has been on a rising in recent years. Machine learning and neural networks are being advocated as a way of improving the performance of several functions across all layers of future communication systems. Furthermore, in applications where complexity reduction is essential for the system feasibility at the cost of an affordable performance loss, more efficient systems might be achieved with the aid of machine learning algorithms. Signal detection for multiple-input multiple-output (MIMO) systems has become a hot topic in recent years given its prominent role in fourth and fifth generations of mobile networks. However, the computational complexity in MIMO systems can become prohibitive when the number of antennas increases. Therefore, by leveraging neural networks architectures we propose a deep unfolded detector, whereby the algorithm of the probability data association (PDA) detector is adapted and enhanced by means of neural network learning capabilities. We unveil that the proposed detector is orders-of-magnitude less complex than the PDA detector, yet presenting no severe penalties in performance in terms of bit error rate (BER).

Keywords: MIMO, Signal detection, Machine learning, Neural networks, Deep unfolding, Low-complexity

1 Introduction

The purported success of multiple-input multiple-output (MIMO) systems is being confirmed since the fourth generation of mobile networks (4G) and continued to show its importance in recent deployments of the fifth generation of mobile networks (5G) technology. Early studies on the sixth generation of mobile networks (6G) also show MIMO systems as a key enabler for future wireless systems [1]. Its advantages over classical single-input single-output (SISO) systems are extremely attractive and relatively simple to understand from a theoretical standpoint [2, 3]: By increasing the number of service antennas, an overall increase in data throughput is obtained.

It was shown in [4] that detectors based on neural networks (NNs) have a competitive performance when compared to the optimum maximum likelihood detector (MLD), while the former is more robust to imperfect channel estimations and

less complex than the latter. However, the system model in [4] considers a system. Recently, several works [5, 6] proposed solutions that attempt to integrate machine learning (ML) and NN to MIMO systems. One emerging solution involves adapting NN architectures according to model-driven detection algorithms, such that its iterations are unfolded on NN layers. This solution is called deep unfolding [5, 7].

Therefore, in this work we propose a deep unfolded detector [8] based on the probability data association (PDA) detector [9] for MIMO systems. The main aim is to achieve the aforementioned advantages of data-driven detectors for SISO systems in MIMO systems, while advantageous features of the PDA detector [3] are maintained. To the best of authors knowledge, this is the first attempt at combining the deep unfolded architecture with the algorithm of the PDA detector for MIMO systems.

MIMO systems are also largely used for beamforming and beamsteering in the most recent mobile networks, where precoding can provide spatial multiplexing and improve the system performance without increasing the complexity on the receiver side [10]. It is clear that precoding will play an important role in future MIMO systems for mobile communications. Nevertheless, this work focuses on MIMO systems where multiple antennas transmit data over a rich scattering environment without considering precoding, relying on detection techniques that can resolve the inter-antenna interference (IAI) with affordable complexity, a scenario where the PDA detector is an interesting solution [3].

1.1 Contributions and paper organization

In this paper, we make the following contributions:

- We propose a novel combination of the data-driven deep unfolded detector and the PDA algorithm for signal detection in MIMO systems;
- Differently from other similar proposals [5, 6, 8], we employ the categorical cross-entropy loss function and dispense with the use of optimal Gaussian denoisers;
- The computational complexity of the proposed detector is evaluated and compared with the complexity presented by detectors of interest;
- A low-complexity variation of the deep unfolded PDA (DU-PDA) is also presented, its computational complexity being lower than the linear zero-forcing (ZF)detector;
- Numerical results from computational simulations compare the uncoded and coded error rates of the proposed detectors with other detectors under time-dispersive channels.

The remainder of this paper is organized as follows. In Sect. 2, we present the system model of the baseline orthogonal frequency division multiplexing (OFDM)-MIMO system. Section 3 then introduces the problem of signal detection for MIMO systems and gives a brief description of the PDA detector and of the deep unfolding learning. This is followed by a description of the proposed DU-PDA and an analysis on the computational complexity of all detectors discussed throughout this paper. Next, in Sect. 4, we provide numerical results to evaluate the performance of all detectors studied in this paper, including the optimum MLD. Finally, Sect. 5 concludes the paper.

1.2 Notation

Throughout this paper, italicized letters (e.g., x or X) represent scalars, boldfaced lowercase letters (e.g., $\mathbf{x}$) represent vectors, and boldfaced uppercase letters (e.g., $\mathbf{X}$) denote matrices. The nth entry of the vector $\mathbf{x}$ is represented by $x(n)$. The entry on the ith row and jth column of the matrix $\mathbf{X}$ is denoted by $X_{i,j}$. The superscript $\mathbf{x}^{(n)}$ denotes the nth instance of the vector $\mathbf{x}$, such that $\mathcal{X} = \{\mathbf{x}^{(n)}\}_{\forall n}$ forms a collection of vectors or a dataset. The sets of real and complex numbers are represented by $\mathbb{R}$ and $\mathbb{C}$, respectively. The absolute value of the scalar $x \in \mathbb{R}$ or the modulus of $x \in \mathbb{C}$ is denoted by $|x|$. The sets of vectors of dimension X with real and complex entries are, respectively, represented by $\mathbb{R}^X$ and $\mathbb{C}^X$. The sets of matrices of dimension $X \times Y$ with real and complex entries are correspondingly described by $\mathbb{R}^{X\times Y}$ and $\mathbb{C}^{X\times Y}$. The transposition operation of a vector or matrix is represented as $(\cdot)^{\mathrm{T}}$. The ℓ_p -norm, $p \geq 1$, of the vector $\mathbf{x}$ is given by $\|\mathbf{x}\|_p = \left(|x(0)|^p + |x(1)|^p + \cdots + |x(n-1)|^p\right)^{1/p}$. The expected value of the random variable z is denoted by $\mathrm{E}[z]$. The real and imaginary parts of $z \in \mathbb{C}$ are denoted by $\Re(z)$ and $\Im(z)$. The estimate of a scalar x, a vector $\mathbf{x}$ or a matrix $\mathbf{X}$ is represented by $\hat{x}$, $\hat{\mathbf{x}}$ and $\hat{\mathbf{X}}$, respectively. The number of elements in a set $\mathcal{X}$ is given by $\#\mathcal{X}$. Computational complexity is denoted by the asymptotic operator $\mathcal{O}(\cdot)$.

2 System model

Suppose that in a multiple antenna system we have N_{t} transmitting antennas and N_{r} receiving antennas, thereby constituting an $N_{\mathrm{t}} \times N_{\mathrm{r}}$ point-to-point baseband and fully digital MIMO system. Therefore, bits of data are demultiplexed into N_{t} substreams, which in turn are mapped to a sequence of complex symbols. These symbols are transmitted by its respective transmit antenna using an OFDM system, for which it is assumed that the cyclic prefix (CP) length is larger than the maximum delay spread for all $N_{\mathrm{t}}N_{\mathrm{r}}$ channels. Finally, after performing the discrete Fourier transform (DFT) we have the following representation of the received baseband signal at the kth subcarrier:

$$\tilde{\mathbf{r}}_k = \tilde{\mathbf{H}}_k \tilde{\mathbf{a}}_k + \tilde{\mathbf{n}}_k. \tag{1}$$

Here, $\tilde{\mathbf{H}}_k \in \mathbb{C}^{N_{\mathrm{r}}\times N_{\mathrm{t}}}$ is the matrix containing all channel frequency responses for the kth OFDM subcarrier; $\tilde{\mathbf{a}}_k \in \mathbb{C}^{N_{\mathrm{t}}}$ represents the symbol vector transmitted by the N_{t} transmit antennas on the kth subcarrier of the OFDM block and $\tilde{\mathbf{n}}_k \in \mathbb{C}^{N_{\mathrm{r}}}$ is the complex additive white Gaussian noise (AWGN) vector in the frequency domain at the kth subcarrier for the N_{r} receive antennas, with zero mean and covariance matrix given by $\sigma^2\mathbf{I}_{N_{\mathrm{r}}}$.

For convenience, henceforth we make use of the real-valued representation [3, 8, 9] for systems. Therefore, let the received signal (1) be represented by the concatenation of its real and imaginary parts, such that

$$\mathbf{r}_k = \mathbf{H}_k \mathbf{a}_k + \mathbf{n}_k, \tag{2}$$

where

$$\mathbf{r}_k = \left[\Re(\tilde{\mathbf{r}}_k)^{\mathrm{T}}\ \Im(\tilde{\mathbf{r}}_k)^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{2N_{\mathrm{r}}},\ \forall\, k, \tag{3}$$

$$\mathbf{H}_k = \begin{bmatrix} \Re(\tilde{\mathbf{H}}_k) & -\Im(\tilde{\mathbf{H}}_k) \\ \Im(\tilde{\mathbf{H}}_k) & \Re(\tilde{\mathbf{H}}_k) \end{bmatrix} \in \mathbb{R}^{2N_\mathrm{r} \times 2N_\mathrm{t}}, \ \forall\, k, \tag{4}$$

$$\mathbf{a}_k = \left[\Re(\tilde{\mathbf{a}}_k)^\mathrm{T} \ \Im(\tilde{\mathbf{a}}_k)^\mathrm{T} \right]^\mathrm{T} \in \mathbb{R}^{2N_\mathrm{t}}, \ \forall\, k, \tag{5}$$

$$\mathbf{n}_k = \left[\Re(\tilde{\mathbf{n}}_k)^\mathrm{T} \ \Im(\tilde{\mathbf{n}}_k)^\mathrm{T} \right]^\mathrm{T} \in \mathbb{R}^{2N_\mathrm{r}}. \tag{6}$$

Moreover, we assume that $\Re(\tilde{\mathbf{a}}_k) \in \mathbb{S}^{N_\mathrm{t}}$ and $\Im(\tilde{\mathbf{a}}_k) \in \mathbb{S}^{N_\mathrm{t}}$; that is, the real and imaginary parts of $\tilde{\mathbf{a}}_k$ can take on different values from the finite set of coordinates pertaining to the square M-quadrature amplitude modulation (QAM) constellation. Hence, let $\mathbb{S} = \{\pm E_0, \pm 3E_0, \ldots, \pm(\sqrt{M}-1)E_0\}$, for $E_0 = \sqrt{\frac{3}{2(M-1)}}$, such that the constellation energy is normalized to 1 (unity).

3 Detection in MIMO systems

A classical problem in the MIMO literature is to decide which symbols were transmitted by each antenna when only (2) is available at the receiver. This detection problem can be solved optimally, however at great computational effort, by the MLD for MIMO as follows

$$\hat{\mathbf{a}}_k = \underset{\mathbf{a}_k \in \mathbb{S}^{2N_\mathrm{t}}}{\arg\min} \|\mathbf{r}_k - \mathbf{H}_k \mathbf{a}_k\|_2^2, \tag{7}$$

for which $\hat{\mathbf{a}}_k \in \mathbb{S}^{2N_\mathrm{t}}$ is the estimated vector of symbols' coordinates.

It is known that the prohibitive complexity presented by the MLD motivated the research of several alternative detectors for MIMO throughout the last decades [3]. The PDA detector is one of these alternatives that presents significantly lower complexity when compared with the MLD, with an affordable bit error rate(BER) performance loss under specific conditions, as will be detailed in Sects. 3.5 and 4. In Sect. 3.1, the PDA detectors' algorithm first proposed in [9] is briefly revisited, followed by our proposed DU-PDA, for which the PDA is the underlying algorithm.

3.1 Probability data association detector

Before the detection task is carried out by the PDA detector, the received signal, $\mathbf{r}_k$, is preprocessed or equalized using the ZF principle as follows [2, 3, 9]

$$\mathbf{z}_k = \mathbf{H}_k^\dagger \mathbf{r}_k = \mathbf{a}_k + \mathbf{v}_k, \tag{8}$$

wherein $\mathbf{H}_k^\dagger = (\mathbf{H}_k^T \mathbf{H}_k)^{-1} \mathbf{H}_k^T$ is the left Moore–Penrose pseudoinverse and $\mathbf{v}_k = \mathbf{H}_k^\dagger \mathbf{n}_k$ is the enhanced AWGN. Let us rewrite (8), such that

$$\mathbf{z}_k = \mathbf{e}_i a_k(i) + \underbrace{\sum_{j \neq i} \mathbf{e}_j a_k(j) + \mathbf{v}_k}_{\mathcal{V}_i}, \ \forall\, i, j \in \{0, 1, \ldots, 2N_\mathrm{t} - 1\}, \tag{9}$$

where $\mathbf{e}_i$ is the vector with 1 (one) at its ith entry and 0 (zero) otherwise, and $\mathcal{V}_i$ is a multivariate random variable (RV) that can be seen as the effective interference-plus-noise contaminating $a_k(i)$ [9]. Therefore, the crux is at detecting the symbol transmitted by the ith antenna, while considering that all other $j \neq i$ transmitted symbols are interference added to the noise term, which is described by $\mathcal{V}_i$.

Therefore, the PDA detector associates, for each $a_k(i)$, a probability vector $\mathbf{p}_i \in \mathbb{R}^{\sqrt{M}}$, which is given by the evaluation of $P_m(a_k(i) = q(m) \mid \mathbf{z}_k, \{\mathbf{p}_j\}_{\forall j \neq i})$; $q(m) \in \mathbb{S}$ being a coordinate of the M-QAM constellation and $m \in \{0, 1, \ldots, \sqrt{M} - 1\}$. It is important to remark that the PDA detector uses all $\{\mathbf{p}_j\}_{\forall j \neq i}$ associated with interfering symbols already detected, thanks to the incorporation of a strategy similar to that of successive interference cancellation (SIC) detectors. This significantly reduces the computational complexity for calculating $\mathbf{p}_i$, since otherwise $P_m(a_k(i) = q(m) \mid \mathbf{z}_k)$ would have to be evaluated. The problem here is the requirement of computing multiple integrals for each received symbol, rendering this evaluation prohibitive in practice. Dropping the subscript $(\cdot)_k$ in order to simplify the notation and assuming that $\mathcal{V}_i$ has a Gaussian distribution [9, 11], then the likelihood function of $\mathbf{z} \mid a(i) = q(m)$ can be defined as

$$P_m(\mathbf{z} \mid a(i) = q(m)) \propto \exp\left(\alpha_m(i)\right), \tag{10}$$

for which

$$\alpha_m(i) = \left(\mathbf{z} - \boldsymbol{\mu}_i - 0.5\mathbf{e}_i q(m)\right)^{\mathrm{T}} \boldsymbol{\Omega}_i^{-1} \mathbf{e}_i q(m), \tag{11}$$

wherein $\mathrm{E}[\mathcal{V}_i] = \boldsymbol{\mu}_i$ and $\mathrm{COV}[\mathcal{V}_i] = \boldsymbol{\Omega}_i$ are given by

$$\boldsymbol{\mu}_i = \sum_{j \neq i} \mathbf{e}_j \left(\mathbf{q}^{\mathrm{T}} \mathbf{p}_j\right), \tag{12}$$

$$\boldsymbol{\Omega}_i = \sum_{j \neq i} \mathbf{e}_j \mathbf{e}_j^{\mathrm{T}} \left(\left(\mathbf{q}^2\right)^{\mathrm{T}} \mathbf{p}_j - \boldsymbol{\mu}_j^2 \right) + 0.5\sigma^2 \mathbf{G}^{-1}, \tag{13}$$

where $\mathbf{q} = [q(0)\ q(1)\ \ldots\ q(\sqrt{M} - 1)]^{\mathrm{T}}$ and $\mathbf{G}^{-1} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}$ is the inverse of the Gram matrix [2] that accounts for the noise enhancement caused by the ZF. To evaluate the posteriors probabilities associated with each symbol, we compute

$$P_m(a(i) = q(m) \mid \mathbf{z}, \{\mathbf{p}_j\}_{\forall j \neq i}) \approx \frac{P_m(\mathbf{z} \mid a(i) = q(m))}{\sum_{m=0}^{\sqrt{M}-1} P_m(\mathbf{z} \mid a(i) = q(m))}, \tag{14}$$

which can be seen as an approximate form of the Bayesian theorem [11]. Then, substituting (10) into (14) yields

$$p_i(m) = \frac{\exp\left(\alpha_m(i)\right)}{\sum_{m=0}^{\sqrt{M}-1} \exp\left(\alpha_m(i)\right)}. \tag{15}$$

Finally, the PDA detector procedure is given in Algorithm 1.

Algorithm 1 The PDA detector

Require: $\tilde{\mathbf{z}}$ via (8)
Require: k_i (see (16)), $\epsilon > 0$
Ensure: $p_i(m) \leftarrow \frac{1}{\sqrt{M}},\ \forall\ m\ \forall\ i$
repeat
  for $i = 1, 2, \ldots, 2N_t$ **do** ▷ outer iteration
    $\mathbf{p}'_i \leftarrow \mathbf{p}_i$
    Compute $\boldsymbol{\mu}_{k_i}$ from (12) and $\boldsymbol{\Omega}_{k_i}$ from (13) with $\{\mathbf{p}_j\}_{\forall j \neq k_i}$
    for $m = 1, 2, \ldots, \sqrt{M}$ **do** ▷ inner iteration
      Calculate $\alpha_m(k_i)$ from (11)
      Evaluate:
      $P_m(\tilde{a}(k_i) = q(m) \mid \tilde{\mathbf{z}}, \{\mathbf{p}_j\}_{\forall j \neq k_i}) \approx p_{k_i}(m)$, from (15)
    end for
  end for
until $|\mathbf{p}_i - \mathbf{p}'_i| \leq \epsilon,\ \forall\ i$ ▷ convergence iteration
$l_i \leftarrow \arg\max_m \{p_i(m)\},\ \forall\ i$
Decide transmitted symbols $\hat{a}(i) \leftarrow q_{l_i},\ \forall\ i$

Note that the optimal detection sequence [9] used in Algorithm 1 can be found with the aid of the following operation:

$$\rho(i) = \frac{1}{\mathbf{f}_i^{\mathrm{T}} \mathbf{H} \mathbf{f}_i} \max \left\{ 0, \mathbf{f}_i^{\mathrm{T}} \mathbf{h}_i - \sum_{j \neq i} |\mathbf{f}_i^{\mathrm{T}} \mathbf{h}_j| \right\}^2, \tag{16}$$

where $\mathbf{f}_i^{\mathrm{T}}$ represents the ith row of $\mathbf{F} = \mathbf{H}^{\dagger}$ and $\mathbf{h}_j$ denotes the jth column of $\mathbf{H}$. Note that larger magnitudes for $\rho(i)$ mean that the ith antenna suffers less IAI [3]. In other words, the off-diagonal entries of the ith row from $\mathbf{FH}$ have, combined, smaller magnitudes than its ith diagonal entry. It is easy to show that the optimal sequence is defined by sorting $\boldsymbol{\rho} = [\rho(0)\ \rho(1)\ \ldots\ \rho(2N_{\mathrm{t}} - 1)]^{\mathrm{T}}$ in a descending order, denoted as $\{k_i \in \{1, \ldots, 2N_{\mathrm{t}}\} \mid \rho(k_0) > \rho(k_1) > \ldots > \rho\big(k_{2N_{\mathrm{t}}}\big)\}$.

3.2 Deep unfolding

Prior to presenting our proposed DU-PDA detector, a brief description of NNs and deep unfolding is provided in this section.

In general, the NN architecture has shown great potential for detecting signals, but its design and parameterization, among other problems, impose limitations [4]. Alternatively, this architecture can be adapted such that iterations of an given algorithm are unfolded on its layers [5, 6, 12], hence the term "unfolding." It is also commonly assumed that the NN employs several layers and, consequently, the term "deep" is added.

More specifically, consider an algorithm with an input vector denoted by $\mathbf{x} \in \mathbb{R}^N$, for which its output is given by $\mathbf{y} \in \mathbb{R}^S$, then this algorithm can be expressed by [12]

$$y(s) = g(\mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\Theta}),\ \forall\ s \in \{0, 1, \ldots, S-1\}, \tag{17}$$

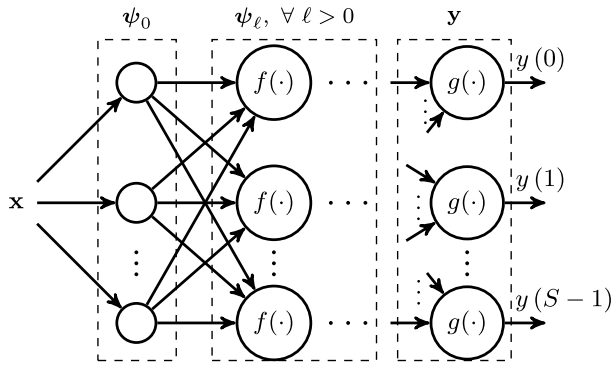

Fig. 1 Deep unfolding architecture. It is based on an underlying algorithm with an input vector given by **x** and an output determined by **y**. Each hidden layer unfolds the ℓth iteration of this algorithm and its input–output relationship is expressed by (18), whereas the output layer is represented by (17)

wherein $\boldsymbol{\Theta}$ is the set of all parameters used by the algorithm, $g(\cdot)$ represents a mapping function, usually nonlinear, and $\boldsymbol{\psi}$ is iteratively updated as follows

$$\psi_\ell(s) = f(\mathbf{x}, \psi_{\ell-1}(s), \boldsymbol{\Theta}), \tag{18}$$

where the ℓth iteration also involves an operation with a mapping function $f(\cdot)$ and $\boldsymbol{\psi}_0$ denotes the initial value.

Therefore, in the deep unfolded context, $\boldsymbol{\psi}_\ell$ can be understood as the input-output relationship at the ℓth layer of a NN architecture, as illustrated in Fig. 1.

Note that dimensions of learnable parameters $\boldsymbol{\Theta}$ are defined according to the underlying algorithm after which (17), (18) and the architecture depicted in Fig. 1 are based. This includes weights and bias, for example, which are optimized by the NN training algorithm [4, 12]. In other words, this means that the number of layers and neurons is fixed, thereby simplifying considerably the process of defining what is commonly known as the NN hyperparameters.

Moreover, improvements are also obtained by using the aforementioned learnable parameters directly into the iterative algorithm. That way, learning capabilities of NNs can be applied for optimizing algorithms such that its global performance, computational complexity, or even both, are improved. In Sect. 3.3, the PDA detector, reviewed in Sect. 3.1, is implemented using the deep unfolded architecture for NNs, unveiling our proposed DU-PDA detector for MIMO systems.

3.3 Proposed deep unfolded PDA detector

Aiming to take advantage of the iterative algorithm of the PDA detector, we propose the DU-PDA detector. Firstly, in the DU-PDA detector, the received signal, $\mathbf{r}$, is preprocessed at the ℓth layer by the following operation [8]; [13, §IV-B, p. 1706]

$$\mathbf{z}_\ell = \hat{\mathbf{a}}_\ell + w_\ell \mathbf{H}^{\mathrm{T}}\left(\mathbf{r} - \mathbf{H}\hat{\mathbf{a}}_\ell\right),\ \forall\ \ell \in \{0, 1, \ldots, L-1\}, \tag{19}$$

where $\hat{\mathbf{a}}_\ell \in \mathbb{R}^{2N_\mathrm{t}}$ is the estimated transmitted symbol vector and the scalar $w_\ell \in \mathbb{R}$ represents a learnable parameter. Note that this preprocessing principle differs from the ZF, which is used by the PDA detector, as defined in (8). In contrast, for the proposed

DU-PDA, it is employed a preprocessing based on the approximate message passing (AMP)algorithm [14], which also bear similarities with the Richardson method [2, §IV-6, p. 9]. In this way, $\hat{\mathbf{a}}_\ell$ is updated iteratively until it converges to an acceptable approximation of the transmitted symbol vector. Interestingly, when we have $\hat{\mathbf{a}}_\ell \to \mathbf{a}$, then the so-called residual term $\left(\mathbf{r} - \mathbf{H}\hat{\mathbf{a}}_\ell\right) \to \mathbf{n}$, which give us a result in (19) similar to (8).

The preprocessed signal of (19) is then fed into the following operation[1]:

$$\begin{aligned}\psi_{\ell^*}(m) = \operatorname{softm}\Big(\left(\mathbf{z}_\ell - \boldsymbol{\mu}_{\ell^*} - 0.5\mathbf{e}_{\ell^*}q(m)\right)^{\mathrm{T}}\boldsymbol{\Omega}_{\ell^*}^{-1}\mathbf{e}_{\ell^*}q(m)\Big)\\ \forall\, m \in \{0, 1, \ldots, \sqrt{M} - 1\},\end{aligned} \tag{20}$$

where

$$\operatorname{softm}(x_\ell(m)) = \frac{e^{x_\ell(m)}}{\sum_{m=0}^{L-1} e^{x_\ell(m)}}. \tag{21}$$

Note that the nonlinear function $\operatorname{softm}(\cdot)$ is applied at each layer. This makes (20) identical to (15) except that it is unfolded on successive layers and that $\boldsymbol{\psi}_j = \mathbf{p}_j$. Notably, this also distinguishes the proposed DU-PDA from other architectures [5, 8, 13] that use instead optimal denoisers at each layer, which do not account for interfering symbols as the underlying PDA algorithm of the DU-PDA does. Moreover, since the preprocessing is modified, then it is necessary to redefine the covariance matrix, $\boldsymbol{\Omega}_{\ell^*}$, as follows [15, §III-D, p. 2023], [8]

$$\boldsymbol{\Omega}_{\ell^*} = \sum_{j \neq \ell^*} \mathbf{e}_j\mathbf{e}_j^{\mathrm{T}}\left(\left(\mathbf{q}^2\right)^{\mathrm{T}}\boldsymbol{\psi}_j - \boldsymbol{\mu}_j^2\right) + \mathbf{e}_{\ell^*}\mathbf{e}_{\ell^*}^{\mathrm{T}}\mathrm{COV}[\mathbf{z}_\ell - \mathbf{a}], \tag{22}$$

where

$$\mathrm{COV}[\mathbf{z}_\ell - \mathbf{a}] = \frac{[\epsilon_\ell]_+\|\mathbf{I}_{2N_\mathrm{t}} - w_\ell\mathbf{H}^{\mathrm{T}}\mathbf{H}\|_2^2 + 0.5\sigma^2\|w_\ell\mathbf{H}^{\mathrm{T}}\|_2^2}{2N_\mathrm{t}}, \tag{23}$$

wherein $[x]_+ = \max(0, x)$ and for which

$$\epsilon_\ell = \frac{\|\mathbf{r} - \mathbf{H}\hat{\mathbf{a}}_\ell\|_2^2 - N_\mathrm{r}\sigma^2}{\|\mathbf{H}\|_2^2}. \tag{24}$$

Equation (23) can be understood as the empirical mean-squared error (MSE) estimator of the covariance matrix originated from the residual and noise terms of (19). More importantly, note that $\boldsymbol{\Omega}_{\ell^*}$ is now a diagonal matrix. This means that computing $\boldsymbol{\Omega}_{\ell^*}^{-1}$ is not as costly as its counterpart in (11), that is, in the PDA detector. More details about such implications are given in Sect. 3.4.

Therefore, by considering developments presented in this subsection and the general model described in Sect. 3.2, we have

$$\psi_{\ell^*+1}(m) = \operatorname{softm}\left(\mathbf{z}_\ell, \psi_{\ell^*}(m), \{w_\ell, \boldsymbol{\mu}_{\ell^*}, \boldsymbol{\Omega}_{\ell^*}\}\right), \tag{25}$$

[1] $\{\ell^* \in \{0, 1, \ldots, 2N_\mathrm{t} - 1\},\ k \in \{1, 2, \ldots, \lceil L/2N_\mathrm{t}\rceil - 1\} \mid \ell^* = \ell - k2N_\mathrm{t};\ k2N_\mathrm{t} \leq \ell < (k+1)2N_\mathrm{t}\}$

which is similar to what is evaluated in (15) with the addition, however, of a learnable parameter and a different preprocessing of the received signal. Note also that $\boldsymbol{\psi}_L = \mathbf{y}$, meaning that the last layer output is also given by (25). Furthermore, let

$$\hat{\mathbf{a}}_{\ell+1} = \sum_{j \neq \ell} \mathbf{e}_j z_\ell(j) + \mathbf{e}_\ell \left(\mathbf{q}^{\mathrm{T}} \boldsymbol{\psi}_{\ell^*} \right), \tag{26}$$

such that the convergence of (19) might be improved, given that the soft combining of symbols' coordinates and their estimated associated probabilities are fed forward to the next layer.

In Algorithm 2,

Algorithm 2 The DU-PDA detector.

function LAYER($\mathbf{r}$, $\mathbf{H}$, $\boldsymbol{\psi}_{\ell^*-1}$, $\hat{\mathbf{a}}_{\ell-1}$)
  Evaluate (19) and (22), followed by (20) and (26)
  return $\boldsymbol{\psi}_{\ell^*}$, $\hat{\mathbf{a}}_\ell$
end function

Ensure: $N_{\mathsf{TR}} > 0$
Ensure: $\psi_{\ell^*}(m) \leftarrow \frac{1}{\sqrt{M}}$, $\forall\ m\ \forall\ \ell^*$
Ensure: $\hat{\mathbf{a}}_0 = \sum_{\ell^*} \mathbf{e}_{\ell^*} \left(\mathbf{q}^T \boldsymbol{\psi}_{\ell^*} \right)$
Require: $\mathcal{L}(\mathcal{I}, \boldsymbol{\psi})$ (see (27))

procedure TRAIN($\mathcal{L}(\mathcal{I}, \boldsymbol{\psi})$, $\boldsymbol{\psi}$, $\hat{\mathbf{a}}_0$, N_{TR})
  for all Epochs **do**
    Generate set of training samples:
    $\mathcal{S}_{\mathsf{TR}} = \left\{ \left(\mathbf{r}^{(1)}, \mathbf{I}^{(1)} \right), \ldots, \left(\mathbf{r}^{(N_{\mathsf{TR}})}, \mathbf{I}^{(N_{\mathsf{TR}})} \right) \right\}$
    for $\ell = 1, 2, \ldots, L+1$ **do**
      Train: LAYER($\mathbf{r}^{(1,2,\ldots,N_{\mathsf{TR}})}$, $\mathbf{H}^{(1,2,\ldots,N_{\mathsf{TR}})}$, $\boldsymbol{\psi}_{\ell^*-1}$, $\hat{\mathbf{a}}_{\ell-1}$)
    end for
  end for
end procedure

procedure DETECT($\mathbf{r}$, $\mathbf{H}$)
  Execute foward-pass: LAYER($\mathbf{r}$, $\mathbf{H}$, $\boldsymbol{\psi}_{\ell^*-1}$), $\forall\ \ell$
  $d_{\ell^*} \leftarrow \arg\max_m \{\psi_{\ell^*}(m)\}$, $\forall\ \ell^*$
end procedure

Decide transmitted symbols $\hat{a}(\ell^*) \leftarrow q_{d_{\ell^*}}$, $\forall\ \ell^*$

we detail the general procedure carried out by the proposed DU-PDA detector.

The ground truth used for training the NN is defined by $\mathbf{I}_{\ell^*} = [I(0)\, I(1)\, \ldots\, I(\sqrt{M}-1)]^{\mathrm{T}}$, such that $\mathcal{I} = \{\mathbf{I}_{\ell^*}\}_{\forall \ell^*}$. It indicates the known constellation coordinates that are transmitted for the training procedure; thus, $I_{\ell^*}(m) \in \{0, 1\}\ \forall\ m$. Observe also that the PDA detector outputs approximate posteriors, as shown in (15), which is leveraged by our proposed DU-PDA detector in Algorithm 2 when employing the categorical cross-entropy loss function:

$$\mathcal{L}(\mathcal{I}, \boldsymbol{\psi}) = \frac{-1}{\sqrt{M}} \sum_{\ell^*} \mathbf{I}_{\ell^*} \log\left(\boldsymbol{\psi}_{\ell^*}\right) + (1 - \mathbf{I}_{\ell^*}) \log\left(1 - \boldsymbol{\psi}_{\ell^*}\right). \tag{27}$$

Bear in mind that the loss is calculated considering the output of all L unfolded layers and not only the last one. Also, note that the use of (27) contrasts with the popular choice of the MSE loss function [5]. Additionally, it is a well-known fact that the cross-entropy loss function is more appropriate for classification tasks.

3.4 Simplified DU-PDA

The model of the DU-PDA presented in the previous subsection can be simplified even further if some assumptions are made. Therefore, a new variation of the proposed DU-PDA detector, namely the simplified DU-PDA detector, is presented in this subsection. For this detector, the calculations performed in (23) are simplified and the scalar $0.5\sigma^2$ is applied directly in (22). The reasoning behind this approach lies in the asymptotic case, that is, when $N_\mathrm{t} \to \infty$ and $N_\mathrm{r} \to \infty$. For this case, the first term of (23) vanishes, since[2]

$$\mathbf{H}^\mathrm{T}\mathbf{H} \to \mathbf{I}_{2N_\mathrm{t}}, \tag{28}$$

and similarly for the second term we have

$$\|w_\ell \mathbf{H}^\mathrm{T}\|_2^2 \to 2N_\mathrm{t}, \tag{29}$$

which yields

$$\begin{aligned} \mathrm{COV}[\mathbf{z}_\ell - \mathbf{a}] &\to \frac{[\epsilon_\ell]_+ \|\mathbf{I}_{2N_\mathrm{t}} - \mathbf{I}_{2N_\mathrm{t}}\|_2^2 + N_\mathrm{t}\sigma^2}{2N_\mathrm{t}} \\ &\to 0.5\sigma^2, \end{aligned} \tag{30}$$

wherein, for the sake of simplicity, the learnable parameter w_ℓ is omitted. This is analogous to the channel hardening effect present in massive MIMO systems [2, 3], where values for N_t and N_r are large. Although we demonstrate via computational simulations in Sect. 4 that the simplified DU-PDA only presents marginal losses in performance, it is still unknown if other similar architectures proposed in the literature [5, 6, 8, 13] are robust enough to allow such simplifications.

3.5 Computational complexity

According to the guidelines presented in [4, §IV-C, p. 122404], the global computation complexity of the PDA detector is approximately given by

$$\mathcal{O}(16N_\mathrm{t}^4 + 8\sqrt{M}N_\mathrm{t}^3 + 8N_\mathrm{t}^2(N_\mathrm{r} + \sqrt{M}) + 4N_\mathrm{t}N_\mathrm{r}). \tag{31}$$

However, if we let $N_\mathrm{r} \gg \sqrt{M}$ and simplify constants, then it can be written more compactly as

$$\mathcal{O}(N_\mathrm{t}^4 + \sqrt{M}N_\mathrm{t}^3 + N_\mathrm{t}^2 N_\mathrm{r} + N_\mathrm{t}N_\mathrm{r}). \tag{32}$$

[2] We adopt the normalization of the channel matrix by $1/\sqrt{N_\mathrm{r}}$ as detailed in Sect. 4.

Note that $\mathcal{O}(8N_t^3 + 16N_t^2N_r + 4N_tN_r)$ refers to the local cost of (8), where the inverse of $\mathbf{G}$ costs $\mathcal{O}(8N_t^3)$[3] and $\mathcal{O}(16N_t^4 + 8\sqrt{M}(N_t^3 + N_t^2))$ is the complexity due to computing (11), for which $\mathbf{\Omega}_i^{-1}$ costs $\mathcal{O}(8N_t^3)$ [9] per outer iteration in Algorithm 1.

Moreover, the DU-PDA detector has an approximate global complexity of

$$\mathcal{O}(4LN_t^2 + 4LN_t(4N_r + \sqrt{M}) + LN_r). \tag{33}$$

Consider again that all constants are simplified and that $N_r \gg \sqrt{M}$ is simplified (33) to

$$\mathcal{O}(LN_t^2 + LN_tN_r + LN_r). \tag{34}$$

The global complexity is composed mainly by the local cost of (19), given by $\mathcal{O}(8N_tN_r)$ per layer, and the local cost of (23), expressed by $\mathcal{O}(4N_t^2 + 8N_tN_r + N_r)$ for each layer[4]. The NN training stage cost is not taking into account when calculating the computational complexity of the detection stage, since the training stage is assumed to be computed offline as discussed in [4]. Nevertheless, in general, the backpropagation algorithm used for training NNs has a complexity that scales linearly with the number of training samples, N_{TR}, and training iterations, say N_{TI}. More importantly, it scales exponentially with the number of layers L because of the chain rule derivatives calculated during backpropagation. In principle, this is a high complexity when compared with the detection or forward-pass complexity, but once trained, the NN-based detector may serve multiple users during a prescribed timeline [16]. This means that the training complexity cost is distributed over time and users, whereas the detection complexity is fixed for each user and transmission cycle. Hence, since training is not performed in the detection cycle, its complexity is not considered, enabling a fair comparison with other detectors.

Furthermore, recall that the simplified form of calculation demonstrated by (30) reduces even further the global complexity of the proposed DU-PDA detector. More specifically, the global complexity of the simplified DU-PDA detector is given approximately by $\mathcal{O}(LN_tN_r)$, meaning that the cost is reduced to one order-of-magnitude when compared to the DU-PDA detector.

From the computational complexity associated with each detector, it is possible to conclude that the PDA is more complex than the proposed DU-PDA. More specifically, this cost difference is due to the higher-order term N_t^4, included in the PDA global complexity. This is expected because of the inversion of matrices performed by the PDA detector, which are not necessary for both the DU-PDA and simplified DU-PDA detectors. Also, notice that for both of these detectors, the total number of layers L might significantly increase its global complexity. It is demonstrated in Sect. 4, however, that this number is a multiple of N_t, thus still implying in a lower global complexity for the DU-PDA when compared to the PDA. In fact, the simplified DU-PDA complexity becomes even lower than that of the ZF in the aforementioned case. Additionally, an optimal detection sequence, such as (16), is not a general requirement for the DU-PDA, which further reduces its global complexity in relation to the PDA.

[3] For the sake of brevity, we assume that the inverse of a matrix, say $\mathbf{X} \in \mathbb{R}^{N\times N}$, is computed by the well-known Gaussian elimination, whose cost is approximately $\mathcal{O}(N^3)$.

[4] Note that the squared norm of a matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ can be written as $\|\mathbf{X}\|_2^2 = \sum_{\forall i}\sum_{\forall j} X_{i,j}^2$; thus, its cost is $\mathcal{O}(MN)$.

Table 1 Global computational complexity of detectors studied in this work. Note that they are given in the most compact form and are also ranked in an ascending order, that is, from less to more costly as lines progress to the bottom of the table

Detector	Global Computational Complexity
Simplified DU-PDA	$\mathcal{O}(LN_tN_r)$
Approximate Message Passing (AMP)	$\mathcal{O}(N_IN_tN_r + N_IN_t\sqrt{M})$
Zero-Forcing (ZF)	$\mathcal{O}(N_t^3 + N_t^2N_r + N_tN_r)$
Deep Unfolded PDA (DU-PDA)	$\mathcal{O}(LN_t^2 + LN_tN_r + LN_r)$
Probability Data Association (PDA)	$\mathcal{O}(N_t^4 + \sqrt{M}N_t^3 + N_t^2N_r + N_tN_r)$
Sphere Detector (SD)	$\mathcal{O}(M^{\sqrt{N_t}})$
Maximum Likelihood Detector (MLD)	$\mathcal{O}(M^{N_t}(N_tN_r + N_r))$

Despite shedding light on how detectors' computational complexity compares to each other, these are only asymptotic predictions of complexity. A detailed evaluation of system end-to-end latency [17, 18], for example, is out of scope in this work. However, it can be verified for a typical 4×8 MIMO considered in Sect. 4, that the symbol detection (see Line 15 of Algorithm 2) of the DU-PDA takes approximately 50 milliseconds in average with neglectable variance. Note that this time value heavily depends on the implementation of the proposed detector, which in this work is based on the TensorFlow library [19] not yet optimized for a full-fledged hardware implementation. Indeed, implementations using hardware description language (HDL) can provide a more reliable analysis on the end-to-end latency of the proposed detector.

For convenience, Table 1 summarizes the global computational complexity for all detectors of interest. Observe that the AMP detector and the sphere detector (SD) are also included for the sake of completeness. For the AMP, N_I refers to the number of iterations or updates executed, whereas for the SD we considered the fixed-complexity SD [3, §VIII-D, p. 20], since its performance is near-optimum. To conclude, note also in Table 1 how the complexity of all detectors increases polynomially with the number of transmitting antennas N_t. The exceptions, however, are the MLD and the SD, whose complexity increases exponentially with N_t and $\sqrt{N_t}$, respectively, as expected.

4 Numerical results and discussion

Before presenting numerical results about the detectors performances, we list important system parameters in the following subsection.

4.1 System parameters

In this work, the following system parameters are adopted: (i) Before transmission, a frame of n_b data bits is encoded using the polar encoder [20] with a code rate of $R < 1$. Thus, n_b/R bits now represent the coded frame that is effectively transmitted; (ii) entries of the channel frequency response matrix, $\mathbf{H}$, are drawn from a complex Gaussian random process for all k subcarriers at each transmission of an OFDM frame and are normalized by $1/\sqrt{N_r}$. Hence, we have $H_{i,j} \sim \mathcal{CN}(0, 1/N_r)$, $\forall\, i, j$ and, consequently, the system signal-to-noise ratio (SNR) per bit can be expressed as follows

Table 2 Hyperparameters of interest for the proposed DU-PDA

Hyperparameters	Values
Training set size	10^5 samples
Layers	$L = 4N_t$
Input dimension	$\mathbb{R}^{2N_r}$, $\mathbb{R}^{2N_r \times 2N_t}$, $\mathbb{R}^{2N_t \times \sqrt{M}}$, $\mathbb{R}^{2N_t}$
Output dimension	$\mathbb{R}^{2N_t \times \sqrt{M}}$
Number of learnable parameters	$\#\{w_\ell\}_{\forall \ell} = 4N_t$
Activation function	softmax$(\cdot)$, $\forall \ell$
Learning rate	10^{-3}
Solver	Adam

$$\Gamma_k = \left(\sqrt{M}R\right)^{-1} \frac{\mathrm{E}\left[\|\mathbf{H}_k \mathbf{a}_k\|_2^2\right]}{N_r \sigma^2}, \ \forall\, k, \tag{35}$$

which is henceforward assumed to be identical for all subcarriers.

The BER is employed for measuring coded detectors' performances, which is obtained by averaging bit decision errors over multiple Monte Carlo experiments. Each experiment is generated using a computational simulation that involves: (i) the generation of $n_b = 256$ equiprobable data bits; (ii) the encoding of data bits by the polar encoder, resulting in a codeword of $\frac{256}{R}$ bits; (iii) mapping of coded bits into complex symbols $\mathbf{a}_k \in \mathbb{S}^{N_t}$ for all k subcarriers; (iv) transmission of the OFDM frame; (v) the generation of normalized channel coefficients to form entries of the channel matrix $\mathbf{H}_k$; (vi) the generation of complex AWGN samples present in the receiver; (vii) the final decision in favor of the symbol coordinate associated with the higher probability value; and (viii) the subsequent decoding of decided symbols into bits via the polar decoder [21]. More specifically, we implement a tree-based architecture of a successive cancellation list decoding [22], with code rate equal to R.

For the sake of brevity, some algorithmic procedures[5] were omitted from Algorithm 2. However, it is worth mentioning that the DU-PDA training is performed considering that SNR values are drawn from a uniform distribution $\mathcal{U} \sim [\min(\text{SNR}), \max(\text{SNR})]$, as discussed in [4, §VI-A, p. 122405]. Additionally, it was decided heuristically to use a total number of $N_{\text{TR}} = 10^5$ samples for training and also that the DU-PDA should include $L = 4N_{\text{t}}$ layers[6]. More details about the proposed DU-PDA hyperparameters can be verified in Table 2. These parameters are used for all scenarios demonstrated in Sect. 4.2.

Furthermore, note that in this work we employ hard decoding for all detectors analyzed. However, in principle, soft decoding could also be integrated to the proposed DU-PDA since soft outputs are available via (25) [11]. Nonetheless, for the proposed DU-PDA, the hard decoding approach attains a better performance-complexity

[5] As mentioned earlier, we used the TensorFlow library [19] to implement a customized deep unfolded NN model. The implementation code can be found at https://github.com/PedroSouza-INATEL/DU-PDA-coded.git.

[6] It was verified that the PDA algorithm converges within an average of 2 convergence iterations in Algorithm 1 (with $\epsilon = 10^{-3}$), for all scenarios of interest. Therefore, there is no loss of generality when comparing both detectors costs in the context of results presented in this section.

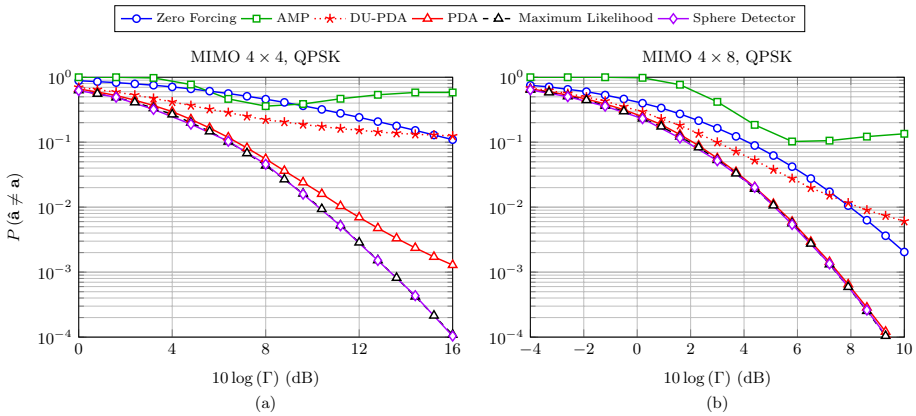

Fig. 2 Performance of the ZF, AMP,DU-PDA,PDA, MLD and SD detectors for the uncoded $N_t \times N_r$ MIMO system. The performance metric is the probability of symbol vector error, $P(\hat{\mathbf{a}} \neq \mathbf{a})$, which is given as a function of a range of SNR values. The scenario of **a** 4 × 4 MIMO is illustrated, followed by the **b** 4 × 8 MIMO, both considering the QPSK modulation

trade-off, which is more aligned with the general aim of the work of proposing a low-complexity detector with affordable performance losses. This also allows for a fair comparison with algorithms that provide hard decoding sequences.

4.2 Performance results

Figure 2 brings the uncoded detection performance for all detectors presented in Table 1, considering a square 4 × 4 MIMO (Fig. 2a) system and a underloaded [3] 4 × 8 MIMO (Fig. 2b), all of which employ the quadrature phase shift keying (QPSK) ($M = 4$) modulation.

The detection performance is given as a function of multiple SNR values, and it is defined as the probability of occurrence of any error in the received symbol vector. This is done because bits are not encoded for the scenarios analyzed in Fig. 2.

Firstly, observe in Fig. 2a that the performance of the PDA detector adheres closely with that reported in the seminal work of [9], thus validating the simulation model. Moreover, notice that the DU-PDA detector has shown a prohibitive performance for the 4 × 4 MIMO scenario, which was also verified to be the case for other square MIMO systems. However, for the underloaded scenario demonstrated in Fig. 2b, where $N_r \gg N_t$, the DU-PDA detector presents better performance. All the same, if the relative performance of the DU-PDA against the ZF and, particularly, the AMP detectors is taken into account, then Fig. 2a and b shows that the DU-PDA outperforms these detectors for most of the SNR range analyzed, while presenting a comparable detection complexity[7]. It was verified, however, that for the underloaded scenario of 4 × 8 MIMO, the DU-PDA detector reaches a performance floor of $P(\hat{\mathbf{a}} \neq \mathbf{a}) \approx 3 \times 10^{-3}$, from which no improvement can be obtained irrespective of how high are the SNR values.

[7] Note here that we consider $L = 4N_t$ as stated in Sect. 4.1, making N_t^3 the highest-order term within the DU-PDA complexity. Additionally, we also considered $N_l = 50$ [8, §IV-A, p. 5] for the AMP detector, which clearly implies $N_l \gg N_t$ and, consequently, also a highest cubic-order polynomial.

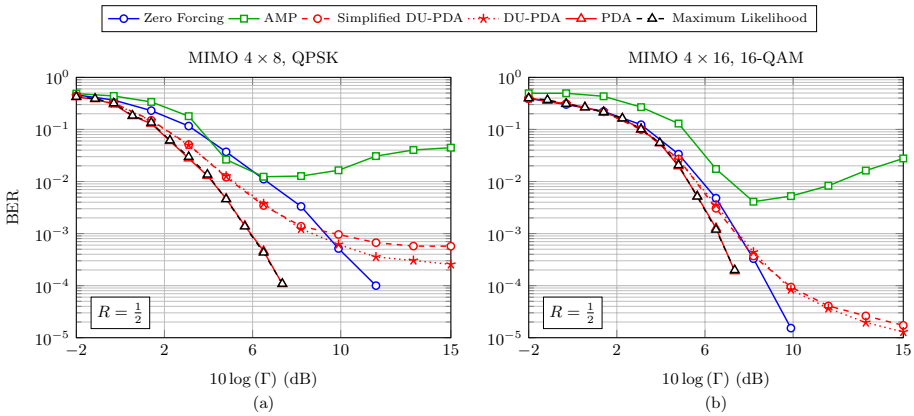

Fig. 3 Performance of the ZF, AMP, simplified DU-PDA, DU-PDA, PDA and MLD detectors for the coded $N_t \times N_r$ MIMO system. Here, the performance metric is the BER, which is given as a function of a range of SNR values. The scenario presented is of the **a** 4×8 MIMO with a code rate of $R = 1/2$ and QPSK modulation, followed by the **b** 4×16 MIMO also with $R = 1/2$ and considering now the 16-QAM modulation. Note that we have omitted the SD curves here because it achieves the MLD performance

This motivated the integration of the Polar encoder as described in Sect. 4.1, also with a aim at potentially improving the proposed DU-PDA performance relative to other detectors. Note in Fig. 3a that the 4×8 MIMO scenario is illustrated again as in Fig. 2, however, considering now the Polar encoding with a code rate of $R = 1/2$.

This is accompanied in Fig. 3b, for which the 4×16 MIMO scenario with a 16-QAM ($M = 16$) modulation is presented, considering the same aforementioned code rate.

We begin by pointing out that the performance floor observed in Fig. 3a and b, although undesirable, is not so much detrimental to the overall performance as in Fig. 2b. This happens because the introduced channel coding improves the performance for all the SNR range under analysis. Therefore, the BER values where the DU-PDA is better than the ZF and AMP consist of the more interesting region of values for which SNR < 10 (dB). It is granted that the performance floor is still presented in Fig. 3a and b, but now at low values of BER $\approx 2 \times 10^{-4}$ and BER $\approx 2 \times 10^{-5}$, respectively. These observations support the conjecture that the uncoded DU-PDA detector is interference limited for high SNR values. In this SNR range, the distribution of (19) ceases to be approximately Gaussian because of the low AWGN levels and becomes defined in most part by the non-Gaussian IAI distribution. This in turn violates the Gaussian distribution assumption mentioned in Sect. 3.1, regarding the PDA detector, which is the underlying algorithm of the proposed DU-PDA detector. Hence, we have the performance floor shown in Fig. 2b, but which is partially mitigated by a robust coding scheme in Fig. 3. Furthermore, to elaborate on the detection performance of the AMP detector in Figs. 2 and 3, one can see that this detector suffers from a severe performance floor for high SNR. This behavior is also explained by the reasoning described for the DU-PDA, which means that the violation of the Gaussian distribution assumption also severely affects the AMP detection performance [23].

Moreover, note also that Fig. 3 depicts the detection performance of the simplified DU-PDA detector. For this detector, the calculations performed in (23) are simplified, yielding (30). Although the dimensions of MIMO systems illustrated in Fig. 3 are

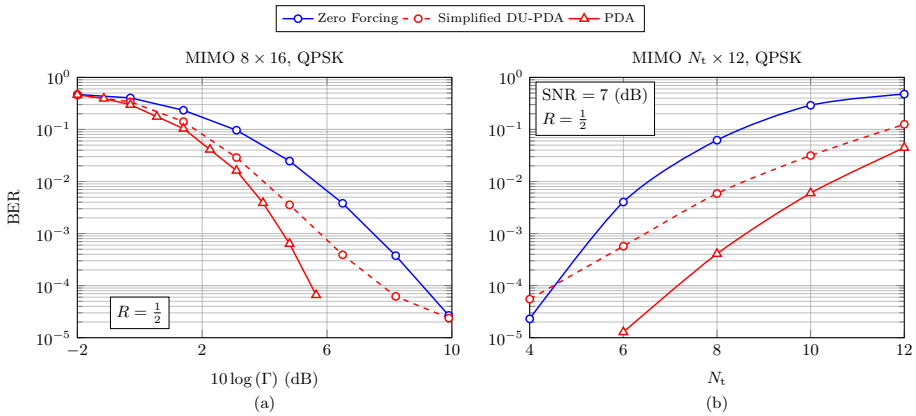

Fig. 4 Performance of the ZF, simplified DU-PDA and PDA detectors for the coded (code rate of $R = 1/2$) MIMO system. The performance given in terms of BER values is plotted against a range of SNR values for the **a** 8×16 MIMO, and as a function of multiple values for the number of transmitting antennas when considering the **b** $N_t \times 12$ MIMO scenario. All scenarios presented consider the QPSK modulation. The remaining detectors described in Table 1 are not analyzed either because their performance is too prohibitive or identical to the performance of detectors already shown here

not large, numerical BER results presented here show that conclusions from Sect. 3.4 may still hold for a small number of antennas. Note in Fig. 3 that the detection performance of the simplified DU-PDA detector is practically identical to the DU-PDA detectors' performance, except at the high SNR region where the simplified DU-PDA is marginally worse than the DU-PDA detector.

Finally, note also that the simplified DU-PDA complexity becomes even lower than that of the ZF and AMP detectors, especially when the number of $L = 4N_t$ layers used is considered. This makes the simplified DU-PDA detector the less costly of all detectors analyzed in this work, as can be verified in Table 1. Yet it performs approximately 2 dB better than the ZF in Fig. 3a, for values of SNR < 10 (dB), for example. More importantly, the simplified DU-PDA largely improves upon the performance of the AMP detector, in spite of using similar operations as described in (19).

Additionally, Fig. 4a

shows the performance of relevant detectors for the 8×16 MIMO scenario considering the QPSK modulation. Figure 4b in turn illustrates detectors performances, also considering the QPSK modulation, for multiple values of transmitting antennas, N_t, for which the number of receiving antennas, $N_r = 12$, and the SNR $= 7$ (dB) are fixed. Note that for this scenario we still assume the number of layers, L, of the DU-PDA detector, to be restrained by N_t, such that $L = 2cN_t$. This is adopted since each layer in the DU-PDA architecture outputs the posterior associated with one transmitted symbol, a by-product of the underlying PDA algorithm employed by the DU-PDA detector. However, we verified through experiments that for $c > 2$ no improvement was obtained in detection performance, yet at the cost of increased training and detection complexity. Therefore, the value $L = 4N_t$ defined in Sect. 4.1 was shown to be the most suitable one.

In Fig. 4a, it can be observed with the larger MIMO system that the proposed simplified DU-PDA detector outperforms the ZF detector, particularly for the low BER

$< 10^{-3}$ region. It is important to remark that for higher values of SNR the performance floor of the coded simplified DU-PDA is still present, remaining, however, at low BER values of approximately 10^{-5}. Moreover, note that the simplified DU-PDA performance becomes worse relative to the PDA detectors' performance as the SNR values get higher, but recall that the simplified DU-PDA presents the lowest complexity (see Table 1). In addition to that, Fig. 4b shows that the simplified DU-PDA detector performance varies approximately linearly with the number of transmitting antennas N_t, while the performance of the ZF detector changes more abruptly with N_t. This means that the proposed simplified DU-PDA detector not only outperforms the more complex ZF, but it is also more robust for all considered MIMO system dimensions, assuming a target BER of 10^{-3}.

5 Conclusion

In this work, we proposed a detector for MIMO systems based upon the deep unfolded architecture for NNs, namely the DU-PDA detector. This detector unfolds iterations of the PDA algorithm in its layers, enhancing the model-driven PDA detector with the aid of its data-driven architecture.

It was shown that the DU-PDA detector, as well as its simplified form, outperforms both the AMP and ZF detectors, considering most of the SNR range evaluated. This can be particularly verified, for instance, in coded detection for the 8×16 MIMO system. However, the global computational complexity of the simplified DU-PDA detector is orders-of-magnitude less than the ZF detector. Furthermore, the lack of matrix inverses computations in the DU-PDA architecture not only reduces its cost, but also simplifies its implementation in practical systems. This is the case when, for example, channels are correlated, increasing the condition number of $\mathbf{G}$ and making impractical its inverse computation.

For future research endeavors, it would be interesting to increase the scenarios and dimensions of MIMO systems analyzed, by increasing the number of transmitting and receiving antennas, also evaluating practical underloaded and square MIMO systems alike. Moreover, the integration of soft decoding to the proposed DU-PDA can improve its performance and can be regarded as a natural progression of the research done in this work. The applicability of the proposed detector in MIMO systems that employ precoding is also an interesting research topic for future works. Finally, given the flexibility of the deep unfolding architecture, we maintain that other MIMO detection schemes could benefit greatly from the principles laid out in this work, becoming thus a promising topic for future research.

Abbreviations

4G	Fourth generation of mobile networks
5G	Fifth generation of mobile networks
6G	Sixth generation of mobile networks
AMP	Approximate message passing
AWGN	Additive white Gaussian noise
BER	Bit error rate
CP	Cyclic prefix
DFT	Discrete Fourier transform
DNN	Deep neural network
DU-PDA	Deep unfolded PDA
HDL	Hardware description language

IAI	Inter-antenna interference
iid	Independent identically distributed
LTE	Long-term evolution
MF	Matched filter
MIMO	Multiple-input multiple-output
MLP	Multilayer perceptron
MMSE	Minimum mean square error
ML	Machine learning
MLD	Maximum likelihood detector
MSE	Mean-squared error
NN	Neural network
OFDM	Orthogonal frequency division multiplexing
PDA	Probability data association
QAM	Quadrature amplitude modulation
QPSK	Quadrature phase shift keying
RV	Random variable
SD	Sphere detector
SIC	Successive interference cancellation
SISO	Single-input single-output
SNR	Signal-to-noise ratio
ZF	Zero-forcing

Acknowledgements
Not applicable.

Author contributions
Both authors contributed equally for this publication.

P. H. C. de Souza
P. H. C. S. was born in Santa Rita do Sapucaí, Minas Gerais, MG, Brazil, in 1992. He received the BS and MS degrees in telecommunications engineering from the National Institute of Telecommunications - INATEL, Santa Rita do Sapucaí, in 2015 and 2017, respectively, and is currently working toward the PhD degree in telecommunications engineering at INATEL. During the year of 2014, he was a Hardware Tester with the INATEL Competence Center - ICC. His main interests are: digital communication systems, mobile telecommunications systems, 6G, cognitive radio, convex optimization for telecommunication systems, compressive sensing/learning, embedded systems and embedded hardware/firmware.

L. L. Mendes
L. L. M. received the BSc and MSc degrees from Inatel, Brazil, in 2001 and 2003, respectively, and the Doctor degree from Unicamp, Brazil, in 2007, all in electrical engineering. Since 2001, he has been a Professor with Inatel, where he has acted as the Technical Manager of the Hardware Development Laboratory from 2006 to 2012. From 2013 to 2015, he was a Visiting Researcher with the Technical University of Dresden in the Vodafone Chair Mobile Communications Systems, where he has developed his postdoctoral. In 2017, he was elected Research Coordinator of the 5G Brazil Project, an association involving industries, telecom operators and academia which aims for funding and build an ecosystem toward 5G in Brazil. He is also the technical coordinator of the Brazil 6G Project.

Funding
This work was partially supported by RNP, with resources from MCTIC, Grant No. 01245.010604/2020-14, under the 6G Mobile Communications Systems project of the Radiocommunication Reference Center (Centro de Referência em Radiocomunicações - CRR) of the National Institute of Telecommunications (Instituto Nacional de Telecomunicações - Inatel), Brazil, FAPESP Grant No. 20/05127-2 under the SAMURAI project, CNPq-Brazil and CAPES.

Received: 10 March 2022 Accepted: 28 July 2022
Published online: 09 August 2022

References

1. J. Jeon, G. Lee, A.A.I. Ibrahim, J. Yuan, G. Xu, J. Cho, E. Onggosanusi, Y. Kim, J. Lee, J.C. Zhang, MIMO evolution toward 6G: modular massive MIMO in low-frequency bands. IEEE Commun. Mag. **59**(11), 52–58 (2021). https://doi.org/10.1109/MCOM.211.2100164
2. M.A. Albreem, M. Juntti, S. Shahabuddin, Massive MIMO detection techniques: a survey. IEEE Commun. Surveys Tutor. **21**(4), 3109–3132 (2019). https://doi.org/10.1109/COMST.2019.2935810
3. S. Yang, L. Hanzo, Fifty years of MIMO detection: the road to large-scale MIMOs. IEEE Commun. Surveys Tutor. **17**(4), 1941–1988 (2015). https://doi.org/10.1109/COMST.2015.2475242
4. P.H.C. De Souza, L.L. Mendes, M. Chafii, Compressive learning in communication systems: a neural network receiver for detecting compressed signals in OFDM systems. IEEE Access **9**, 122397–122411 (2021). https://doi.org/10.1109/ACCESS.2021.3108061
5. A. Balatsoukas-Stimming, C. Studer, Deep unfolding for communications systems: a survey and some new directions. In: 2019 IEEE International Workshop on Signal Processing Systems (SiPS), pp. 266–271 (2019). https://doi.org/10.1109/SiPS47522.2019.9020494
6. Q.-V. Pham, N.T. Nguyen, T. Huynh-The, L. Le Bao, K. Lee, W.-J. Hwang, Intelligent radio signal processing: a survey. IEEE Access **9**, 83818–83850 (2021). https://doi.org/10.1109/ACCESS.2021.3087136

7. C. Liu, J. Thompson, T. Arslan, A deep unfolding network for massive multi-user MIMO-OFDM detection. In: 2022 IEEE Wireless Communications and Networking Conference (WCNC), pp. 2405–2410 (2022). https://doi.org/10.1109/WCNC51071.2022.9771554
8. M. Khani, M. Alizadeh, J. Hoydis, P. Fleming, Adaptive neural signal detection for massive MIMO. IEEE Trans. Wireless Commun. **19**(8), 5635–5648 (2020). https://doi.org/10.1109/TWC.2020.2996144
9. D. Pham, K.R. Pattipati, P.K. Willett, J. Luo, A generalized probabilistic data association detector for multiple antenna systems. IEEE Commun. Lett. **8**(4), 205–207 (2004). https://doi.org/10.1109/LCOMM.2004.823405
10. M.A. Albreem, A.H.A. Habbash, A.M. Abu-Hudrouss, S.S. Ikki, Overview of precoding techniques for massive MIMO. IEEE Access **9**, 60764–60801 (2021). https://doi.org/10.1109/ACCESS.2021.3073325
11. S. Yang, T. Lv, R.G. Maunder, L. Hanzo, From nominal to true a posteriori probabilities: an exact Bayesian theorem based probabilistic data association approach for iterative MIMO detection and decoding. IEEE Trans. Commun. **61**(7), 2782–2793 (2013). https://doi.org/10.1109/TCOMM.2013.053013.120427
12. A. Zappone, M. Di Renzo, M. Debbah, Wireless networks design in the era of deep learning: model-based, AI-based, or both? IEEE Trans. Commun. **67**(10), 7331–7376 (2019). https://doi.org/10.1109/TCOMM.2019.2924010
13. H. He, C.-K. Wen, S. Jin, G.Y. Li, Model-driven deep learning for MIMO detection. IEEE Trans. Signal Process. **68**, 1702–1715 (2020). https://doi.org/10.1109/TSP.2020.2976585
14. D.L. Donoho, A. Maleki, A. Montanari, Message-passing algorithms for compressed sensing. Proceedings of the National Academy of Sciences **106**(45), 18914–18919 (2009). https://doi.org/10.1073/pnas.0909892106. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0909892106
15. J. Ma, L. Ping, Orthogonal AMP. IEEE Access **5**, 2020–2033 (2017). https://doi.org/10.1109/ACCESS.2017.2653119
16. S. Ali, W. Saad, N. Rajatheva, K. Chang, D. Steinbach, B. Sliwa, C. Wietfeld, K. Mei, H. Shiri, H.-J. Zepernick, T.M.C. Chu, I. Ahmad, J. Huusko, J. Suutala, S. Bhadauria, V. Bhatia, R. Mitra, S. Amuru, R. Abbas, B. Shao, M. Capobianco, G. Yu, M. Claes, T. Karvonen, M. Chen, M. Girnyk, H. Malik, 6G White Paper on Machine Learning in Wireless Communication Networks. arXiv:2004.13875 (2020). arXiv:2004.13875
17. J. Chen, X. Ran, Deep learning with edge computing: a review. Proc. IEEE **107**(8), 1655–1674 (2019). https://doi.org/10.1109/JPROC.2019.2921977
18. C. Zhang, P. Patras, H. Haddadi, Deep learning in mobile and wireless networking: a survey. IEEE Commun. Surveys Tutor. **21**, 2224–2287 (2019)
19. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A system for large-scale machine learning. arXiv (2016). https://doi.org/10.48550/ARXIV.1605.08695. arXiv:https://arxiv.org/abs/1605.08695
20. Guimarães, D.A.: Digital Transmission: A Simulation-Aided Introduction with VisSim/Comm. Springer, Berlin Heidelberg (2009). https://doi.org/10.1007/978-3-642-01359-1
21. K. Besser, Digcommpy 0.9. https://pypi.org/project/digcommpy/ Accessed 2022-06-07
22. I. Tal, A. Vardy, List decoding of polar codes. IEEE Trans. Inf. Theory **61**(5), 2213–2226 (2015). https://doi.org/10.1109/TIT.2015.2410251
23. E. Beck, C. Bockelmann, A. Dekorsy, CMDNet: learning a probabilistic relaxation of discrete variables for soft detection with low complexity. IEEE Trans. Commun. **69**(12), 8214–8227 (2021). https://doi.org/10.1109/TCOMM.2021.3114682

Publisher's Note