Active User Blind Detection Through Deep Learning

Cyrille Morin[1(✉)], Diane Duchemin[1], Jean-Marie Gorce[2], Claire Goursaud[2], and Leonardo S. Cardoso[2]

[1] Univ Lyon, Inria, INSA Lyon, CITI, Lyon, France
{cyrille.morin,diane.duchemin}@inria.fr
[2] Univ Lyon, INSA Lyon, CITI, Lyon, France
{jean-marie.gorce,claire.goursaud,leonardo.cardoso}@insa-lyon.fr

Abstract. Active user detection is a standard problem that concerns many applications using random access channels in cellular or *ad hoc* networks. Despite being known for a long time, such a detection problem is complex, and standard algorithms for blind detection have to trade between high computational complexity and detection error probability. Traditional algorithms rely on various theoretical frameworks, including compressive sensing and bayesian detection, and lead to iterative algorithms, e.g. orthogonal matching pursuit (OMP). However, none of these algorithms have been proven to achieve optimal performance.

This paper proposes a deep learning based algorithm (NN-MAP) able to improve on the performance of state-of-the-art algorithm while reducing detection time, with a codebook known at training time.

Keywords: Non-coherent active user detection · Machine learning · Massive random access

1 Introduction

Massive access in internet of things (IoT)-dedicated radio networks, especially in 5G, brings several challenges. In this setting, a huge number of sensor nodes is to be sporadically served within the specific constraints of machine type communication (MTC). These networks will be implemented mainly with low cost devices, thus having restricted radio functionalities as well as scarce power and computational resources. Besides, latency, spectrum and energy, must be held to the same efficiency demands of current communication standards, sometimes even higher, to fulfil the requirements of the foreseen tactile internet [1]. Uplink data transmissions from simplified sensor nodes aim at minimising transmission duration as well as the amount of transmissions, given their finite power resources, but also the small amount of data to be transmitted. A high signalling overhead would drastically reduce the operational life-time of such devices as they would spend more time and energy to transmit protocol-related messages than useful payload data. Unlike current 4G based access procedures, as planned in

G. Caso et al. (Eds.): CrownCom 2020, LNICST 374, pp. 3–15, 2021.
https://doi.org/10.1007/978-3-030-73423-7_1

narrowband IoT (NB-IoT) (even though new releases provide a shorter access procedure [2]), an "all in one" grant-free uplink message encapsulating access request, device identifier, and data, would be ideal.

To achieve this grant-free uplink reality, the main challenge is the detection of the active subset of sensor nodes by the base station (BS), also referred to as active user detection (AUD). To enable the transmission of users' identities despite a high network density, the usage of a dedicated spectrum sharing technique is required that must be compatible with the rapidly evolving traffic load within a high number of potential users. Non orthogonal multiple access (NOMA) [3], and in particular code-domain NOMA, is a good candidate [4] for such a spectrum sharing technique as it limits collisions for simultaneous transmissions without requiring to use an extremely long access sequence for each user, given the network density.

As a result, all the complexity of the AUD task is pushed to the BS. Avoiding a handshake procedure requires efficient detection algorithms to retrieve active users' identity from a "one shot" access message with limited channel state information at the receiver (CSIR). The optimal AUD as described in [5] suffers from high complexity which does not seem compatible with real time implementation. An iterative version of the optimal detector, having a lower complexity -but also lower performance-, is also introduced therein. Developing a high performance though low complexity detector is crucial for a realistic and efficient AUD implementation. Most efficient algorithms proposed in the literature to cope with this problem exploit either a Bayesian estimation formalism or the compressive sensing formulation [6,7]. Both have many similarities but lead to different iterative algorithms. Despite their efficiency, none of these algorithms can guarantee to achieve the optimal solution as they have to trade their accuracy with complexity. Therefore, the competition is still open.

With the recent and growing interest of the community toward machine learning, and particularly deep learning (DL), it has been shown that its usage can help to solve complex problems, mainly when defining good models is difficult, or when the models exist but provide solutions too complex for their exploitation. The scenario presented here falls into the second category, and appear to be a good candidate to exploit DL. The objective of this paper is to design a DL receiver for massive NOMA, and more specifically, the AUD in non-coherent channels. Related studies have been done around this subject, for instance in [1,8]. These works are focused on the resource optimisation problem for code domain NOMA and employ auto-encoder based solutions in both cases. They show that the encoding and decoding performance can be improved through end-to-end optimisation. The metrics used there are symbol error rate (SER), sum rate and convergence rate. In [9], the question of imperfect CSIR is addressed for power domain NOMA. This paper is also addresses resource allocation optimisation. The authors of [10], while also dealing with power domain NOMA, focus on channel estimation and signal detection in the context of orthogonal frequency-division multiplexing (OFDM). They propose a comparison with a successive interference cancellation (SIC) based algorithm and show the interest of the DL approach. The model is nevertheless restricted to two users and the channel realisation is fixed in training and testing phases. A preamble and collision detection scheme based on DL is proposed in [11], where the study is performed on pre-processed long term evolution (LTE) random access preamble signals: the correlations with the possible Zadoff-Chu sequences are directly provided to the network. The objective of the authors include the detection of multiple collisions in order to improve contention resolution, and therefore, access probability. Whereas the collision study is realised in a massive access scenario, the detection evaluation is performed with a single user scenario only, by comparing the missed detection performance of the proposed fully connected neural network (NN) with other more classic preamble detectors. All these works are closely related to the use-case of the present paper, but none of them directly address the task at hand: to the best of our knowledge, this work is the first to apply DL on non-coherent AUD with code domain NOMA.

The rest of the paper is organised as follows: Sect. 2 presents our model and the reference schemes to which our approach is compared. Section 3 provides details on the implementation choice regarding the DL scheme we propose, while Sect. 4 is dedicated to the evaluation of the solution. Section 5 concludes the paper.

2 System Model for the Massive Random Access

2.1 Non-coherent AUD

The random access channel in massive MTC is important to guarantee a fair radio medium access. It is herein assumed that the sensor nodes, henceforth referred to simply as *nodes*, receive a random code in advance which is used to send a resource request to the BS they are associated with. In a standard approach, if two nodes request a resource in the same slot, a collision occurs and at least one of the two messages is lost. However, with the knowledge of the codes distributed to the nodes, the BS tries to determine the identity of all the nodes involved in a request. Such an approach, referred as coded random access [6], allows to reduce the number of resources reserved for the random access mechanism and can accelerate the handshake mechanism. Indeed, unlike the 4G access protocol which the NB-IoT is based on and relying on a pool of available Zadoff-Chu access sequences, this approach ensures the uniqueness of the codes employed by the nodes. This fact allows to avoid access code collisions but also additional steps, known as the contention resolution, dedicated to the identification of the users in the handshake procedure. It can also be used as a standalone mechanism in cases where the only one bit is to be transmitted, then the binary value of the nodes' activity suffices without needing further handshake.

In our model, the BS transmits a beacon allowing the nodes to be roughly synchronised and to control their power such that in average the received power at the BS is constant for all nodes. However, the instantaneous channel states are not known, and no pilots are used in this detection phase. The detector thus operates in non-coherent detection mode [5].

We adopt the following notation for the remainder of this work: $(\mathcal{U}, \Phi)$ is a measurable space where $\mathcal{U}$ denotes the total set of nodes, with cardinality $K = |\mathcal{U}|$ and $\Phi = \mathcal{P}(\mathcal{U})$ the powerset of $\mathcal{U}$. A node subset is denoted by $\mathcal{A} \in \Phi$. For a given random access slot, we note $\underline{\mathcal{A}} \in \Phi$ a set of *active* nodes. The activity rate is assumed low (less than 0.5) implying a sparse transmission set. We further assume that the node activity follows a Poisson distribution with mean parameter λ (thus the node activity probability is $\theta = \lambda/K$). As stated previously, a unique codebook $\mathcal{C}$ is generated and shared among the network (the transmitters and BS both agree on the codes during an initial association phase). As a result, each node k owns a dedicated complex Gaussian code $\mathbf{c}_k$ of codelength M and unit power.

As mentioned previously, the received messages are considered synchronous and a perfect average power control allows the messages to be received with an average signal-to-noise ratio (SNR) ρ. The BS possesses N antennas while the nodes have a single antenna. Transmissions are subject to a flat Rayleigh block fading channel, modelled as a random vector $\mathbf{h}_k \sim \mathcal{N}_{\mathscr{C}}(0, \mathbf{I}_N)$ of size N where $\mathbf{I}_N$ is the identity matrix of dimension n and $\mathcal{N}_{\mathscr{C}}(0, \cdot)$ indicates a complex standard Gaussian distribution. The receiver noise introduces an additive white Gaussian noise (AWGN), modelled as a random vector $\mathbf{z} \sim \mathcal{N}_{\mathscr{C}}(0, \mathbf{I}_{NM})$ of size NM. It should be noted that neither the BS nor the transmitting nodes are aware of the actual channel realisations, but only know the channel statistics, as described above. For a given active node k, the channel coefficients $h_{m,n}$ are constant with respect to (w.r.t.) to m and are independent and identically distributed (i.i.d.) w.r.t. n. This means that the message is sent over a narrowband channel, typically a single carrier in an OFDM frame, as defined in NB-IoT for MTC. The proposed model is similar to the one used in [5,7].

Let $\mathbf{y} \in \mathbb{C}^{NM}$ denote the received signal, ρ the targeted SNR and $\otimes$ the Kronecker product. The received signal is then given by:

$$\mathbf{y} = \sum_{k \in \underline{\mathcal{A}}} \sqrt{\rho} (\mathbf{I}_N \otimes \mathbf{c}_k) \mathbf{h}_k + \mathbf{z}. \tag{1}$$

The BS performs an AUD given $\mathbf{y}$ and prior knowledge, restricted to the codebook, the activity probability law and the statistical CSIR. The AUD algorithm is performed on a non-coherent channel, since no pilots are used for prior channel estimation. Let $\hat{\mathcal{A}}$ denote the detected active node subset. To evaluate the performance of the algorithm, the following metrics will be used: codeset error rate (CER), user error rate (UER), misdetection rate (MDR) and false alarm rate (FAR), according to the following definitions:

$$\text{MDR}: \quad \bar{\epsilon}_{md} = \mathbb{E}_k \left[\mathbb{P}[k \notin \hat{\mathcal{A}} | k \in \underline{\mathcal{A}}] \right] \tag{2}$$

$$\text{FAR}: \quad \bar{\epsilon}_{fa} = \mathbb{E}_k \left[\mathbb{P}[k \in \hat{\mathcal{A}} | k \notin \underline{\mathcal{A}}] \right] \tag{3}$$

$$\text{UER}: \quad \bar{\epsilon}_{s} = \bar{\epsilon}_{md} \cdot \theta + \bar{\epsilon}_{fa} \cdot (1 - \theta) \tag{4}$$

$$\text{CER}: \quad \bar{\epsilon}_{C} = p[\hat{\mathcal{A}} \neq \underline{\mathcal{A}}]. \tag{5}$$

The MDR (resp. FAR) corresponds to the false negative (resp. false positive) rate. The UER combines these errors to compute an average individual error rate. In addition, the CER is a system level error rate, that counts the rate of non-perfect codeset detection.

2.2 MAP Detectors

Let $\mathbf{y}_n \in \mathbb{C}^M$ denote the received signal on antenna n and $\mathbf{C}_{\mathcal{A}} \in \mathbb{C}^{M\times\omega}$ the codeset of a given node subset $\mathcal{A}$ whose cardinality is ω. Its singular value decomposition (SVD) is written $\mathbf{C}_{\mathcal{A}} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{U}$, where $\mathbf{V} \in \mathbb{C}^{M\times M}$ and $\mathbf{U} \in \mathbb{C}^{\omega\times\omega}$ are unitary matrices. $\boldsymbol{\Gamma} \in \mathbb{C}^{M\times\omega}$ is composed of the singular values γ on its diagonal. From (2.1), following [5], the likelihood of a codeset is given by:

$$p(\mathbf{y}|\mathcal{A}) = \prod_{n=1}^{N} \frac{1}{\pi^M |\sigma|} \exp\left(\|\tilde{\mathbf{y}}_n\|_2^2 - \|\mathbf{y}_n\|_2^2\right), \tag{6}$$

where $\sigma \in \mathbb{C}^{M\times M}$ is $\sigma = \rho\mathbf{C}_{\mathcal{A}}\mathbf{C}_{\mathcal{A}}^{\mathsf{H}} + \mathbf{I}_M$ and $\tilde{\mathbf{y}}_n$ is the projection of $\mathbf{y}_n$ onto the codeset $\mathbf{C}_{\mathcal{A}}$ space, and is defined as:

$$\tilde{\mathbf{y}}_n = diag\left(\sqrt{\frac{\rho|\gamma_1|^2}{1+\rho|\gamma_1|^2}}, \cdots, \sqrt{\frac{\rho|\gamma_M|^2}{1+\rho|\gamma_M|^2}}\right)\mathbf{V}^{\mathsf{H}}\mathbf{y}_n. \tag{7}$$

The maximum likelihood estimate (MLE) has been used in [5] to estimate the active set. Since we know the prior probability on $(\mathcal{U}, \Phi)$, related to the Poisson distribution, a maximum a posteriori (MAP) detector can be defined and is optimal w.r.t. to the Bayes risk minimisation, when defined from the CER.

Definition 1 (C-MAP estimate). *The C-MAP estimate of the codeset detection problem is given by:*

$$\begin{aligned}\hat{\mathcal{A}}_C &= \arg\min_{\mathcal{A}\in\Phi} p[\underline{\mathcal{A}} \neq \mathcal{A}|\mathbf{y}] && (8)\\ &= \arg\max_{\mathcal{A}\in\Phi} p(\mathbf{y}|\mathcal{A})p(\mathcal{A}), && (9)\end{aligned}$$

where $\hat{\mathcal{A}}_C$ is the detected subset, given the received signal $\mathbf{y}$, from the knowledge of the codebook $\mathcal{C}$, the Gaussian distributions of the channels $\mathbf{h}_k$ and noise $\mathbf{z}$.

In addition, the prior probability is given by $\mathbb{P}(\mathcal{A}) = \lambda^{|\mathcal{A}|}\cdot(1-\lambda)^{|\mathcal{U}|-|\mathcal{A}|}$.

But if the objective is to minimize the user error rate, the Bayes risk is modified and leads to the following estimate.

Definition 2 (U-MAP estimate). *The U-MAP estimate of the codeset detection problem is given by:*

$$\hat{\mathcal{A}}_U = \cup_{k\in\mathcal{U}}\{k|\delta_k(\mathbf{y}) = 1\}, \tag{10}$$

with δ the delta function and where $\delta_k(\mathbf{y}) = 1$ is given by:

$$\delta_k(\mathbf{y}) = \begin{cases} 1 \ \textit{if} \sum\limits_{\substack{\mathcal{A}\in\Phi; \\ k\in\mathcal{A}}} p(\mathbf{y}|\mathcal{A})p(\mathcal{A}) > \sum\limits_{\substack{\mathcal{A}\in\Phi; \\ k\notin\mathcal{A}}} p(\mathbf{y}|\mathcal{A})p(\mathcal{A}) \\ 0 \ \textit{else} \end{cases}. \tag{11}$$

Let us prove that U-MAP is optimal with respect to the UER metric. $P_{MD}(k|\mathbf{y})$ and $P_{FA}(k|\mathbf{y})$, the probability of Missed Detection, respectively False Alarm, of a user k given a received signal $\mathbf{y}$ are given by:

$$P_{MD}(k|\mathbf{y}) = \sum_{\substack{\mathcal{A}\in\Phi; \\ k\in\mathcal{A}}} \mathbb{1}_{[k\notin\hat{\mathcal{A}}(\mathbf{y})]} p(\mathcal{A}|\mathbf{y}) = \mathbb{1}_{[k\notin\hat{\mathcal{A}}(\mathbf{y})]} \sum_{\substack{\mathcal{A}\in\Phi; \\ k\in\mathcal{A}}} p(\mathcal{A}|\mathbf{y}), \tag{12}$$

and

$$P_{FA}(k|\mathbf{y}) = \sum_{\substack{\mathcal{A}\in\Phi; \\ k\notin\mathcal{A}}} \mathbb{1}_{[k\in\hat{\mathcal{A}}(\mathbf{y})]} p(\mathcal{A}|\mathbf{y}) = \mathbb{1}_{[k\in\hat{\mathcal{A}}(\mathbf{y})]} \sum_{\substack{\mathcal{A}\in\Phi; \\ k\notin\mathcal{A}}} p(\mathcal{A}|\mathbf{y}). \tag{13}$$

Minimising the UER thus corresponds to performing a binary test for each user, by comparing $P_{MD}(k|\mathbf{y})$ and $P_{FA}(k|\mathbf{y})$:

$$\delta k(\mathbf{y}) = 1 \text{ if } \sum_{\substack{\mathcal{A}\in\Phi; \\ k\in\mathcal{A}}} p(\mathcal{A}|\mathbf{y}) > \sum_{\substack{\mathcal{A}\in\Phi; \\ k\notin\mathcal{A}}} p(\mathcal{A}|\mathbf{y}). \tag{14}$$

Then, having $p(\mathcal{A}|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{A})p(\mathcal{A})$, the decision given in (11) is optimal.

To compute the estimate given either by Eq. (9) or (11), each element of the power set Φ has to be evaluated, making such kind of AUD non feasible for computational reasons.

2.3 It-MAP Detector

As an alternative to these solution with prohibitive computational complexity, many iterative algorithms have been proposed in the literature [5,7,12]. Following the work presented in [5], the Iterative-MAP (It-MAP) is herein proposed as a reference solution, built as an approximation of C-MAP. The philosophy of It-MAP is similar to that of a Successive Interference Cancellation (SIC) as it processes the received signal $\mathbf{y}$ iteratively and retrieves a new detected user at each iteration i based on the assumption of the previous $\hat{\mathcal{A}}$ at $i-1$. Even if the detected subset is built iteratively, the detection rule is based on the MAP criteria given by Eq. (9) for each i, with a search restricted on some elements of Φ. More precisely, the evaluated subsets $\mathcal{A}_i \in \Phi_i$ are built from the previously detected subset $\hat{\mathcal{A}}_{i-1}$ as follows:

$$\Phi_i = \cup_{k\in\left\{\mathcal{U}\setminus\hat{\mathcal{A}}_{i-1},\emptyset\right\}} \left\{\hat{\mathcal{A}}_{i-1}, k\right\} \tag{15}$$

The It-MAP detection stops as soon as two successive iterations provide the same detected subset, i.e., $\hat{\mathcal{A}}_{i-1} = \hat{\mathcal{A}}_i$.

The architecture of this MAP-based AUD algorithm makes its complexity lower than the computation requirement of the MAP, but at the cost of a reduced accuracy since the iterative detection makes the It-MAP prone to error propagation. This fact let room for other AUD algorithms seeking for a better complexity-accuracy trade-off. As DL is envisioned to accommodate well to large scale problems, a blind AUD based on DL is thus presented in the following section and will be compared to the C-MAP, U-MAP and It-MAP detectors.

3 A Neural Network Based Algorithm

3.1 The NN-MAP Estimate

Definition 3 (NN-MAP estimate). *A NN-MAP architecture for the AUD problem is defined as follows. The inputs to the NN-MAP detector come from* $\mathbf{y}$*, from Eq.* (1)*, and* ρ *(in dB) as a side information. It outputs a vector* $\mathbf{p}$ *of length* K *containing the estimated probability that each node is active. That probability is obtained by using a* sigmoid *activation function at the end of the network and is compared with a vector of ground truth labels* $\mathbf{t}$ *of the same length through a binary cross-entropy loss function* $\mathcal{L}$ *to optimise the network's parameters:*

$$\mathcal{L}(\mathbf{t}, \mathbf{p}) = -\sum_{k=1}^{K} t_k \cdot \log(p_k) + (1 - t_k) \cdot \log(1 - p_k) \tag{16}$$

During the training phase, the average cost over all tuples $(\underline{\mathcal{A}}_i, \mathbf{y}_i, \mathbf{p}_i)$ *is:*

$$\bar{\mathcal{L}} = \frac{1}{I} \sum_i \mathcal{L}(\mathbf{t}(\underline{\mathcal{A}}_i), \mathbf{p}_i), \tag{17}$$

where I *is the mini-batch size.*

After training, the soft probabilities are converted to hard decisions using a threshold: a user k *is considered active if* $p_k > 0.5$.

This choice is justified by the following theorem.

Theorem 1. *For a non-coherent AUD problem, the solution that minimizes the cost function given by* (17) *converges to the U-MAP estimate of Definition 2 if the dataset is large enough.*

Proof. By incorporating (16) into (17), and permuting the sums w.r.t. i and k, one can write:

$$\bar{\mathcal{L}} = \sum_{k \in \mathcal{U}} \bar{\mathcal{L}}(k), \tag{18}$$

with :

$$\bar{\mathcal{L}}(k) = -\sum_{i=1}^{I} \mathbb{1}_{[k \in \underline{\mathcal{A}}_i]} \log(p_k(\mathbf{y}_i)) + \mathbb{1}_{[k \notin \underline{\mathcal{A}}_i]} \log(1 - p_k(\mathbf{y}_i)). \tag{19}$$

Then if the tests are randomly selected according to the prior probability $\mathbb{P}(\underline{\mathcal{A}})$, one have:

$$\bar{\mathcal{L}}^*(k) = \lim_{I\to\infty} \bar{\mathcal{L}}(k) \tag{20}$$
$$= -\mathbb{E}_{\underline{\mathcal{A}},Y}\left[\mathbb{1}_{[k\in\underline{\mathcal{A}}]}\log\left(p_k(\mathbf{y})\right) + \mathbb{1}_{[k\notin\underline{\mathcal{A}}]}\log\left(1-p_k(\mathbf{y})\right)\right]. \tag{21}$$

Finally, using the decomposition $\bar{\mathcal{L}}^*(k) = \int_{\mathcal{Y}} \bar{\mathcal{L}}^*(k|\mathbf{y}) \cdot f_Y(\mathbf{y}) \cdot d\mathbf{y}$, one gets:

$$\bar{\mathcal{L}}^*(k|\mathbf{y}) = -P(k|\mathbf{y}) \cdot \log\left(p_k(\mathbf{y})\right) + (1 - P(k|\mathbf{y})) \cdot \log\left(1 - p_k(\mathbf{y})\right). \tag{22}$$

The minimum of (17) is asymptotically achieved if the NN returns for each observation $\mathbf{y}$, the output $\mathbf{p}$ which minimizes (22) for any user k, independently.

Thanks to the convexity of this function w.r.t. $p_k(y)$ it is straightforward to show that the global cost function is minimal when $p_k(\mathbf{y}) = P(k|\mathbf{y})$, which is nothing but the posterior probability. Therefore, if the learning phase succeeds to find a set of parameters for the NN such that the output probability vector converges to the posterior probability distribution, the U-MAP estimate is achieved by selecting at the output all the nodes k with $p_k \geq 0.5$. Clearly, the NN architecture approximates the U-MAP.

It is worth mentioning that there is no guarantee that a NN can achieve this global optimum. This theorem only claims that the objective function based on the cross-entropy is well-posed w.r.t. the U-MAP problem.

3.2 The NN-MAP System Parameters

In this paper, the NN architecture and hyper parameters have been empirically optimised as follows, for a scenario chosen with $K = 10$, $M = 8$, $N = 4$, $\lambda = 4$. One random codebook is generated and reused for all subsequent training for a fair comparison of the performance thus obtained. Unless specified otherwise, all figures correspond to this scenario.

Architecture Type: The observation vector $\mathbf{y}$ is a combination of random Gaussian variables related to the properties of the codebook, channel realisations and noise. The unique correlation may come from the codebook, which is selected randomly. As such, the correlations are low and convolutional layers are not mandatory. In addition, we don't assume in this model any correlation in the data transmissions, neither between nodes, nor over time. Under these assumptions, a recurrent layer is not necessary. Consequently, the chosen architecture is a fully-connected neural network.

Layers: A set of networks was trained with an increasing amount of dense layers and a constant amount of units in each one, from 3 to 12. The search was not expanded above due to lack of significant improvement. Finally, the 5 layers network was selected.

Units: The number of units per layer was set to be proportional to K, M, and N so it can scale according to scenario complexity. The proportionality factor was chosen by increasing it by powers of 2 from 1 to 128. Performance stopped increasing after 4, so this is what is used in the following.

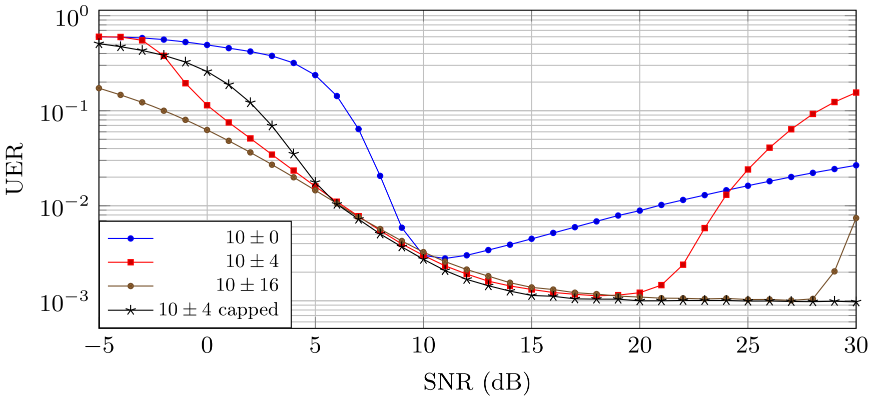

Fig. 1. Network trained on examples with uniform SNR. Mean is kept at 10 dB and range is varied. For the capped line, the SNR value input to the NN at test time is limited to the training range.

SNR Range: Training data is generated with a given ρ, which is also input to the NN. It is therefore necessary to determine the values that will be used to train the network with. Indeed, ρ can have a big impact on the overall performance of the trained network: a too low value could generate a very noisy signal from which the NN would fail to learn anything. A too high SNR would not help the NN to learn to cope with noise. In our approach, the training dataset is generated with a range of SNR values, distributed uniformly over a specified interval. To find out a good interval, the impact of two parameters is evaluated: the range of the interval (Fig. 1) and the mean of that interval. In those two cases, NN shows good performance only inside the SNR interval used for training, with a rapid decrease outside.

Note that when a NN has to work above its learned range, it appears more efficient to limit the SNR values input to the range bounds. This is highlighted in the curve (—⋆—), where the NN learned in the range 10 ± 4 and performs well at SNR above 14 dB. This is not the case for low SNR values, where the NN learned in the range 10 ± 16 outperforms all other curves.

The most important parameters of the NN used in this paper are summarized in Table 1.

Table 1. Summary of network and training parameters

Parameter	Value
Layers	5
Units	$4 \times K \times M \times N$
Learning rate	1×10^{-3}
Optimiser	Adam
Batch size	4096
Training iterations	100000
Training SNR values	10 dB $\pm$ 16 dB

4 Results

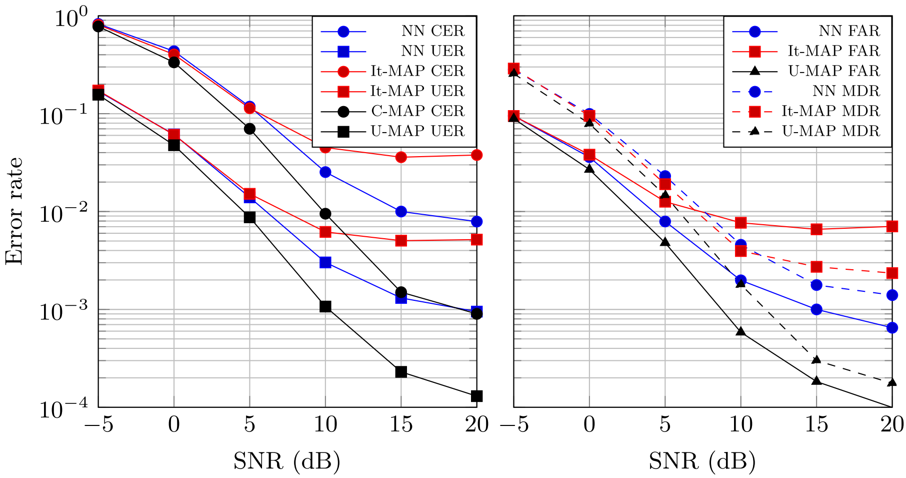

Fig. 2. Performance comparison between C-, U-, It- and NN-MAP algorithms for codeset and user error rates.

Comparison Between the Different Algorithms: The different algorithms defined in the previous section are compared for the chosen scenario ($K = 10$, $M = 8$, $N = 4$, $\lambda = 4$).

As shown in Fig. 2, the trained NN-MAP outperforms It-MAP, especially at high SNR w.r.t. UER and CER metrics as well. The performance of NN-MAP bridges half of the gap with the MAP's optimal given either with C-MAP or U-MAP. In addition, as can be seen in Fig. 2, NN-MAP outperforms It-MAP for both MDR and FAR criteria. It is worth noting however that It-MAP achieves a MDR lower than FAR while it is the opposite for NN-MAP. Note that this trade-off may be easily tuned with NN-MAP. Indeed, in (22), we proved the

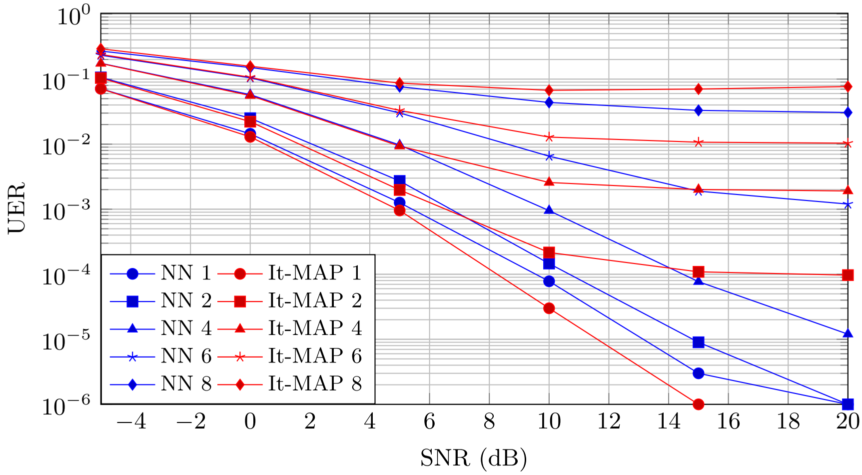

Fig. 3. It-MAP and NN-MAP are set and trained with $\lambda = 4$. A test time, active user number is not random and varied from 1 to 8.

relationship between the MAP and the loss function. It is known that the MAP solution is optimal only if mis-detection and false alarm errors have the same cost. However, it is known that the relative weight of these errors can be tuned to achieve any point of the ROC curve of the detector [13]. For the NN-MAP, one have two options to balance MDR and FAR: the cross-entropy can be modified or the hard decision threshold can be tuned.

All algorithms tested in this paper assume the knowledge of λ as a prior information. Figure 3 shows how It-MAP and NN-MAP perform when the actual number of users deviates from the expected value (the actual number of active users is indicated in the legend). It-MAP outperforms NN-MAP only for 1 active user. It is interesting to mention that a method to increase the performance of NN-MAP in that regard could be to either train it with several values of λ, or to use a uniform distribution instead of the distribution associated to the Poisson distribution assumption.

Analysis on Larger Scenarios: In Fig. 4, the performance results of It-MAP and NN-MAP are given when $K = 20$. Note that in this scenario, both U-MAP and C-MAP are not computable in reasonable time. There, NN-MAP outperforms It-MAP, since it provides a gain in terms of UER without compromising on the CER.

Computational Considerations: On top of performance increases, the NN approach also provides a reduction of computation times as shown in Table 2, where CPU executions are single thread and the GPU execution allows a batch size of 10000. The source of this reduction is twofold: a less complex computation through simple add and multiply operations, without branching leads to a reduced load and smoother execution on the system, and it also creates the

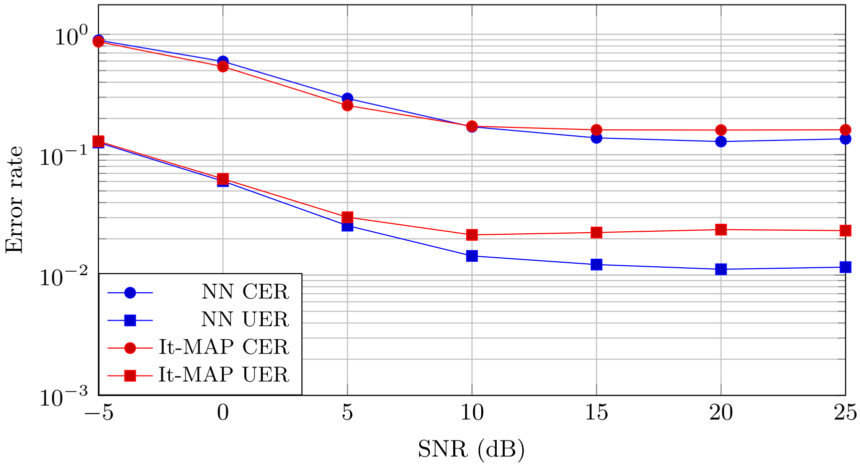

Fig. 4. Performance comparison between NN-MAP and It-MAP algorithms for codeset and user error rates on a scenario with more potential users: $K = 20$.

capability to massively parallelise batches of computations in an efficient manner. The second aspect is most prevalent in the present case: the It-MAP rests on sequential operations with nested loops that create inefficiencies through CPU branch prediction misses and are not trivial to convert to parallel computation, and so, to offload to high performance accelerators such as a GPU.

Table 2. Average example processing time for a scenario with $K = 10$, $M = 8$, $N = 4$, $\lambda = 4$.

MAP CPU	It-MAP CPU	NN CPU	NN GPU
159 ms	8.25 ms	3.6 ms	542 ns

5 Conclusion

In this paper, we have proposed to use DL for the non-coherent AUD problem based on coded domain NOMA. We have proposed a NN-MAP detector which minimises asymptotically the MAP cost function. We have shown that this approach improves on the performance of state-of-the-art iterative algorithms by a factor of 5 in some scenarios, especially w.r.t. the UER metric. Moreover, as it also reduces the algorithmic complexity, this work shows the interest of using DL for such a task even though more work still needs to be done to improve the scalability of the architecture to very large sets of nodes as expected for massive access IoT.

Acknowledgement. This work has been supported by the French National Agency for Research (ANR) under grant n∘ANR-16-CE25-0002 - EPHYL and Nokia Bell Labs in the framework of the Inria-Nokia Bell Labs common lab.

References

1. Ye, N., Li, X., Yu, H., Wang, A., Liu, W., Hou, X.: Deep learning aided grant-free NOMA toward reliable low-latency access in tactile internet of things. IEEE Trans. Ind. Inf. **15**(5), 2995–3005 (2019)
2. 3GPP, Further NB-IoT enhancements (RP-171428) (2017)
3. Ding, Z., Lei, X., Karagiannidis, G.K., Schober, R., Yuan, J., Bhargava, V.K.: A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. IEEE J. Sel. Areas Commun. **35**(10), 2181–2195 (2017)
4. Paolini, E., Stefanovic, C., Liva, G., Popovski, P.: Coded random access: applying codes on graphs to design random access protocols. IEEE Commun. Mag. **53**(6), 144–150 (2015)
5. Duchemin, D., Chetot, L., Gorce, J., Goursaud, C.: Coded random access for massive MTC under statistical channel knowledge. In: IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2019, pp. 1–5 (2019)
6. Wunder, G., Stefanović, Č., Popovski, P., Thiele, L.: Compressive coded random access for massive MTC traffic in 5G systems. In: 2015 49th Asilomar Conference on Signals, Systems and Computers, pp. 13–17. IEEE (2015)
7. Ke, M., Gao, Z., Wu, Y., Gao, X., Schober, R.: Compressive sensing-based adaptive active user detection and channel estimation: massive access meets massive MIMO. IEEE Trans. Signal Process. **68**, 764–779 (2020)
8. Kim, M., Kim, N., Lee, W., Cho, D.: Deep learning-aided SCMA. IEEE Commun. Lett. **22**(4), 720–723 (2018)
9. Liu, M., Song, T., Gui, G.: Deep cognitive perspective: resource allocation for NOMA-based heterogeneous IoT with imperfect SIC. IEEE Internet Things J. **6**(2), 2885–2894 (2019)
10. Narengerile, Thompson, J.: Deep learning for signal detection in nonorthogonal multiple access wireless systems. In: 2019 UK/China Emerging Technologies (UCET), pp. 1–4, August 2019. https://doi.org/10.1109/UCET.2019.8881888. ISSN null
11. Magrin, D., Pielli, C., Stefanovic, C., Zorzi, M.: Enabling LTE RACH collision multiplicity detection via machine learning (2018). arXiv: 1805.11482 [cs.IT]
12. Cirik, A.C., Mysore Balasubramanya, N., Lampe, L.: Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA. IEEE Wirel. Commun. Lett. **7**(1), 46–49 (2018)
13. Poor, H.V.: An Introduction to Signal Detection and Estimation. Springer, New York (2013). https://doi.org/10.1007/978-1-4757-2341-0