GPU-Accelerated Machine Learning in Non-Orthogonal Multiple Access

Daniel Schäufele
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
daniel.schaeufele@hhi.fraunhofer.de

Guillermo Marcus
NVIDIA
Berlin, Germany
gmarcus@nvidia.com

Nikolaus Binder
NVIDIA
Berlin, Germany
nbinder@nvidia.com

Matthias Mehlhose
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
matthias.mehlhose@hhi.fraunhofer.de

Alexander Keller
NVIDIA
Berlin, Germany
akeller@nvidia.com

Sławomir Stańczak
Fraunhofer Heinrich Hertz Institute
Berlin, Germany
slawomir.stanczak@hhi.fraunhofer.de

***Abstract*—Non-orthogonal multiple access (NOMA) is an interesting technology that enables massive connectivity as required in future 5G and 6G networks. While purely linear processing already achieves good performance in NOMA systems, in certain scenarios, non-linear processing is mandatory to ensure acceptable performance.**

In this paper, we propose a neural network architecture that combines the advantages of both linear and non-linear processing. Its real-time detection performance is demonstrated by a highly efficient implementation on a graphics processing unit (GPU). Using real measurements in a laboratory environment, we show the superiority of our approach over conventional methods.

***Index Terms*—machine learning, neural networks, wireless communication, multi-user detection, NOMA, MIMO, massively parallel architectures, GPU, CUDA**

I. Introduction

In future 5G and 6G mobile networks, the demand for massive connectivity will not be satisfiable with traditional orthogonal multiple access (OMA) systems. This is especially the case in massive machine-type communication (mMTC) scenarios, e.g., in campus networks. For this reason a large body of research has been devoted to non-orthogonal multiple access (NOMA) systems [1]–[4]. In order to deal with multi-user interference in such systems, non-linear detection has been proposed (for example, see [5]–[7]).

While non-linear detectors can perform significantly better than linear detectors in scenarios with strong multi-user interference, they can be very sensitive to even small changes in a wireless environment (e.g., due to multi-path scattering and intermittent interference in mMTC scenarios). Therefore, the performance of non-linear detectors can degrade in dynamic environments. Theoretical studies [8] have shown that in massive multiple-input multiple-output (MIMO) systems linear detectors can achieve the spectral efficiency comparable to non-linear methods. For this reason, purely non-linear detectors may be inefficient in massive MIMO NOMA systems.

This work has been partially funded by the German Federal Ministry of Education and Research (BMBF, Germany) in the project Open Testbed Berlin - 5G and Beyond (OTB-5G+) under Grant 16KIS0980.

Contrary to our previous work which focused on using the adaptive projected subgradient method (APSM) [9], we propose a neural network (NN) architecture, which combines a linear and a non-linear branch. Additionally, we propose to use the linear least squares (LLS) algorithm to initialize the weights of the linear branch and to exploit symmetry in the IQ samples to improve the performance over a conventional NN. We then present a graphics processing unit (GPU)-based implementation, which is able to achieve very short execution times. In addition, we evaluate the performance on a real data set that has been acquired in a lab environment.

The remainder of this paper is organized as follows. In Section II we give a formal problem statement as well as a presentation of our proposed NN architecture and several tricks we used to increase the performance. In Sections III and IV we present the real-time implementation and the laboratory setup, which led to the results discussed in Section V.

II. Background

In this section we first give an overview of the system model and define the problem we intend to solve. Then we present our NN architecture and several tricks to improve its performance.

A. Preliminaries

In the following, scalars, column vectors and matrices are denoted by italic lower case letters x, bold lower case letters $\mathbf{x}$ and bold upper case letters $\mathbf{X}$, respectively. Matrix transposes and inverses are denoted by $\mathbf{X}^{\mathsf{T}}$ and $\mathbf{X}^{-1}$, respectively. The set of natural numbers, real numbers and complex numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$, respectively, while the real and imaginary parts of a complex number $c \in \mathbb{C}$ are denoted by $\Re(c)$ and $\Im(c)$, respectively. We define the range $\overline{N_1, N_2} := \{N_1, N_1 + 1, \ldots, N_2\} \subset \mathbb{N}$, where $N_1 \leq N_2$.

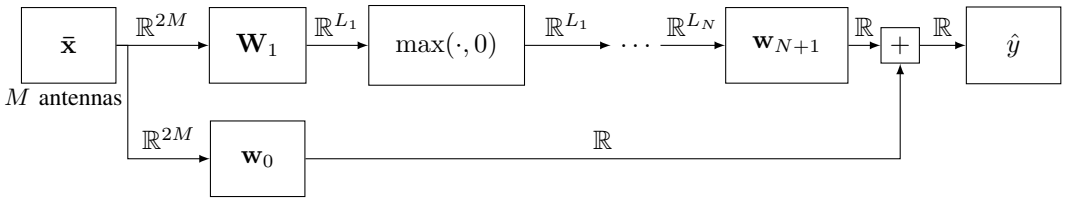

Fig. 1. Proposed structure of NN with N hidden layers.

B. System Model

We assume a multi-user uplink system with K users and M receive antennas and a non-dispersive channel. The received signal $\mathbf{r}(t) \in \mathbb{C}^M$ at the time $t \in \mathbb{N}$ is given by

$$\mathbf{r}(t) = \sum_{k=1}^{K} \sqrt{p_k} \cdot b_k(t) \cdot \mathbf{h}_k + \mathbf{n}(t), \tag{1}$$

where $p_k \in \mathbb{R}$ is the transmit power of user $k \in \overline{1, K}$, $b_k(t) \in \mathbb{C}$ is its information-bearing symbol, while the vectors $\mathbf{h}_k \in \mathbb{C}^M$ and $\mathbf{n}(t) \in \mathbb{C}^M$ stand for the channel signature of user k and additive noise, respectively. Note that we do not assume any distribution of the noise and structure of the receive antenna array. The objective of multi-user detection considered in this study is to design a filter $g^k : \mathbb{C}^M \to \mathbb{C}$ for a selected user k, such that $(\forall t \in \mathbb{N})\ \left|g^k(\mathbf{r}(t)) - b_k(t)\right| \leq \epsilon$, where $\epsilon > 0$ is a small predefined noise tolerance.

The receive process is split into two phases. In the first (training) phase all transmitters transmit a series of N_T uniformly distributed pseudo-random symbols, which are known by the receiver. The transmitted symbols for user k are then collected into a vector $\mathbf{y}_{\mathrm{T},k} = [b_{\mathrm{T},k}(1), b_{\mathrm{T},k}(2), \ldots, b_{\mathrm{T},k}(N_\mathrm{T})]^\mathsf{T} \in \mathbb{C}^{N_\mathrm{T}}$. The corresponding receive signal is also collected into a matrix $\mathbf{X}_\mathrm{T} = [\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(N_\mathrm{T})]^\mathsf{T} \in \mathbb{C}^{N_\mathrm{T} \times M}$. This training data can then be utilized to train the detection filter.

During the following detection phase a series of N_D data symbols is transmitted by each transmitter, which is to be reconstructed at the receiver. Again, we collect the receive signal into a matrix $\mathbf{X}_\mathrm{D} = [\mathbf{r}(N_\mathrm{T}+1), \mathbf{r}(N_\mathrm{T}+2), \ldots, \mathbf{r}(N_\mathrm{T}+N_\mathrm{D})]^\mathsf{T} \in \mathbb{C}^{N_\mathrm{D} \times M}$ and use this matrix as input for the detection algorithm.

C. Exploiting the Symmetry of IQ Samples

According to the argument given in [10], we can use the equivalence

$$g(\mathbf{r}(t)) = f(\mathbf{r_1}(t)) + \mathrm{i}\, f(\mathbf{r_2}(t)) \tag{2}$$

for linear functions $f : \mathbb{R}^{2M} \to \mathbb{R}$ and $g : \mathbb{C}^M \to \mathbb{C}$, where $\mathbf{r_1}(t) = \left[\Re(\mathbf{r}(t))^\mathsf{T}, \Im(\mathbf{r}(t))^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{2M}$ and $\mathbf{r_2}(t) = \left[\Im(\mathbf{r}(t))^\mathsf{T}, -\Re(\mathbf{r}(t))^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{2M}$. Applying this transformation to the training input matrix $\mathbf{X}_\mathrm{T}$ yields a new matrix $\bar{\mathbf{X}}_\mathrm{T} = [\mathbf{r_1}(1), \mathbf{r_2}(1), \ldots, \mathbf{r_1}(N_\mathrm{T}), \mathbf{r_2}(N_\mathrm{T})]^\mathsf{T} \in \mathbb{R}^{2N_\mathrm{T} \times 2M}$ and a new training target vector $\bar{\mathbf{y}}_{\mathrm{T},k} = [\Re(b_k(1)), \Im(b_k(1)), \ldots, \Re(b_k(N_\mathrm{T})), \Im(b_k(N_\mathrm{T}))]^\mathsf{T} \in \mathbb{R}^{2N_\mathrm{T}}$. The same transformation can be applied to the detection input matrix $\mathbf{X}_\mathrm{D}$ to obtain $\bar{\mathbf{X}}_\mathrm{D} \in \mathbb{R}^{2N_\mathrm{D} \times 2M}$ and $\bar{\mathbf{y}}_{\mathrm{D},k} \in \mathbb{R}^{2N_\mathrm{D}}$.

By using this approach, we can use twice as many training samples to train f. At the same time we enforce that both real and imaginary parts are predicted using the same function (with different inputs), which decreases the degrees of freedom of the learned function (compared to using two independent functions). This will speed up the training process given that the constraint is satisfied by the input data.

We later show experimentally that this approach is also beneficial in the case of non-linear NNs, although there is no strong theoretical background for this, yet.

D. Neural Network Structure

In many cases a purely linear estimator can already achieve good performance. However, in certain scenarios, non-linear processing (which we will provide with NNs) is necessary to achieve good performance. In order to combine both approaches, we propose to use the architecture shown in Figure 1. The linear (bottom) branch provides a dense layer without any activation functions, while the top branch contains N hidden, dense layers with L_n neurons each and rectified linear unit (ReLU) activation functions, where $n \in \overline{1, N}$, followed by a final dense layer without non-linearity. Both branches are summed element-wise to produce the final output. This approach resembles a single residual unit in deep residual networks (ResNets) [11]. However, in the proposed architecture, the skip connection contains a matrix multiplication, whose weights we efficiently initialize in a manner outlined in the next section. Due to this initialization, explicit orthogonalization of both branches is unnecessary, as the non-linear branch refines the output of the linear estimator.

E. LLS Initialization of Linear Weights

The optimal solution (under certain assumptions) for a linear estimator can be computed using the LLS algorithm [12, p. 83], where the weights for user k are computed as

$$\mathbf{w}^k = (\bar{\mathbf{X}}_\mathrm{T}^\mathsf{T} \bar{\mathbf{X}}_\mathrm{T})^{-1} \bar{\mathbf{X}}_\mathrm{T}^\mathsf{T} \bar{\mathbf{y}}_{\mathrm{T},k}. \tag{3}$$

We can then compute the solution by a matrix-vector multiplication as $\hat{\mathbf{y}}_{\mathrm{D},k} = \bar{\mathbf{X}}_\mathrm{D} \mathbf{w}^k$. By using the weights $\mathbf{w}^k$ for the weight vector $\mathbf{w}_0$ in the linear branch of the NN presented in Section II-D, we can incorporate the LLS solution into our NN. This approach is beneficial to reduce training time as the training of the NN by traditional stochastic gradient descent (SGD) methods already has a good starting point. The number of computations needed for training is further reduced by fixing the weights of the linear branch and thus avoiding the need for weight updates, resulting in a key performance improvement.

Under certain conditions, we can achieve better performance with minimum mean square error (MMSE) estimation [13], but we deem the added computational effort to be unnecessary, as the estimator will be further refined during the training of the neural network.

III. Real-Time Implementation

In order to achieve real-time performance for our proposed NN structure we propose to use an optimized implementation running on a GPU. For the purpose of this work we consider the fully fused implementation that was used in [14]. This implementation is heavily optimized to save all weights of the networks into the extremely fast registers and to save all intermediate results into shared memory, thus avoiding the need to access the relatively slow global memory. Additionally, the full network is fused into a single kernel, which avoids the need to copy intermediate results to global memory. However, due to the limited amount of registers and shared memory, this approach only works for up to 128 neurons per layer. To show results for layers with a larger number of neurons, we switch to an implementation that applies the general matrix multiplication (GEMM) routines provided by the CUTLASS template library when necessary.

In order to optimize the computation of the LLS solution we used the `cublasSgelsBatched` algorithm provided by the cuBLAS library.

IV. Experimental Setup

This section describes the hardware components and transmission signals of our real-time multi-user detection setup.

A. Hardware Components

The transmitter, receiver, and signal processing equipment are shown in Figure 2. All components are commercial off-the-shelf (COTS) devices.

The server, that was used for signal processing and the data transfer from and to the Software-Defined Radios (SDRs) is equipped with an Intel Xeon W-3245 central processing unit (CPU) and an RTX 2080 Ti consumer GPU.

The SDR equipment is composed of four Ettus USRP N310 SDRs. Furthermore, we use a single National Instruments (NI) OctoClock for a global positioning system (GPS) disciplined clock and timing source for our SDRs. With this setup, all four SDRs, each equipped with four ports on the transmitter (Tx) and receiver (Rx) path, behave like a single SDR system with sixteen synchronized physical antenna ports for both the Rx and Tx paths.

The 16 physical Rx antennas are arranged as a uniform circular array (UCA) with a radius of 6.5 cm. Uniform spacing is ensured by a ring retainer around the antennas. The antennas operate in the 2.4 GHz band with an omnidirectional radiation pattern and vertical polarisation.

For the transmitting users a single NAMC SDR module [15] is used, which is also synchronized to the OctoClock. Attached to this SDR module are six antennas, each of which represents a single transmitting user. Each user antenna is installed on a tripod, allowing for conveniently adjusting location and height, relative to the receiving antenna array. According to the data sheet, the antenna gain is 5 dBi at 2.5 GHz and the antenna polarization is vertical.

Fig. 2. System setup. The antennas in the front are used as transmitters, the receiver consists of one the circular antenna array in the background (only one is used for theses experiments), the Ettus SDRs below it and the server to the left of it.

B. Transmission signal

For data transmission we use Single-Carrier Modulation (SCM). Similar to the NOMA setup in [16] each of the six users is equipped with a single transmit antenna, and all users are perfectly synchronized because we use a single SDR module for all users.

The pseudo-random signal of each user is first quadrature amplitude modulation (QAM) modulated and then up-sampled by a factor of 16 and pulse shaped by a root-raised-cosine (RRC) filter to reduce the signal bandwidth to a total of 1.92 MHz and fit the system sampling rate of 30.72 MHz.

To allow for cross-correlation-based synchronization on the receiver side, a Frank-Zadoff-Chu sequence is added to the signal of each user.

At the receiver, we use the same RRC filter parameters in combination with the corresponding downsampling factor of 16 to recover the modulated signals.

In each transmission a total of 685 complex training symbols and 3840 complex data symbols are transmitted per user.

V. Results

In this section, we first demonstrate the performance of our proposed algorithm for different network dimensions. Then we evaluate the effectiveness of the measures proposed in Section II and finally we investigate the performance in the presence of noise.

In the following, each data point is an average of 50 transmissions, each consisting of 685 complex training symbols and 3840 complex data symbols modulated with quadrature phase-shift keying (QPSK). We only used the receive signal of 4 of the 16 antennas in our system. The reason is, that in a system where the number of users is smaller than the number of antennas, we can compute an optimal solution using

TABLE I
AVERAGE EXECUTION TIMES OF DIFFERENT ALGORITHMS.

Algorithm	Training Time	Detection Time
LLS	238.7 µs	13.4 µs
Fully-fused NN	20.4 ms	32.7 µs
Tensorflow NN	1.06 s	209.1 ms

a linear system, because the system is over-determined. Since the maximum number of users is limited by our available lab equipment, we instead limit the number of receiver antennas to 4 to assess a under-determined system.

The 6 users are consecutively numbered and each user has 3 dB less transmit power than the previous user. For example, the fourth user has 9 dB less transmit power than the strongest (i.e., first) user. By distributing the transmit power in this manner we can run simulations for users with a very low signal-to-interference-plus-noise ratio (SINR).

In addition to the (uncoded) bit error ratio (BER), we also give numbers for coded bit error ratio (CBER) which was measured after applying a low-density parity-check (LDPC) code with a block length of 7680 and a code rate of 0.7 (implemented in the Sionna library [17]). The demapping was performed using the a posteriori probability (APP) algorithm, which used the ground-truth signal-to-noise ratio (SNR) of the decoded user symbols as input.

The timings given in this section do not account for the copying of the data to and from the GPU. In a practical system the previous processing steps would also be performed on the GPU, so that the data would already reside on the GPU and no copying would be necessary.

By using manual parameter sweeps, we managed to achieve good training performance by employing the Adam optimizer [18] with a learning rate of 0.005, a training duration of 50 epochs and a batch size of 128. The relatively large batch size is influenced by the fact that the GPU cannot be fully loaded for smaller batch sizes, which would decrease the throughput. We tried to optimize the mean squared error (MSE) loss.

When choosing the optimal network dimensions, both detection performance in terms of achievable BER and execution time need to be taken into account. The simulation results for different network dimensions are given in Figure 3. A network with 3 hidden layers with 64 nodes each seems to provide a very good trade-off, as this achieves the lowest BER and still has a very short training time. Numerical values for training and detection times are given in Table I. We note that unlike for successive interference cancellation (SIC), where users are decoded sequentially [13], all users can be decoded in parallel and the execution time for our algorithm does not depend on the user that is to be decoded. The execution times given for Tensorflow were measured in graph mode on the same GPU, but are not directly comparable, as they include additional overhead for Python wrappers and data copying.

In Figure 4, we evaluate the performance gain that can be achieved by exploiting the symmetry of the IQ samples as proposed in Section II-C. For comparison, we also show

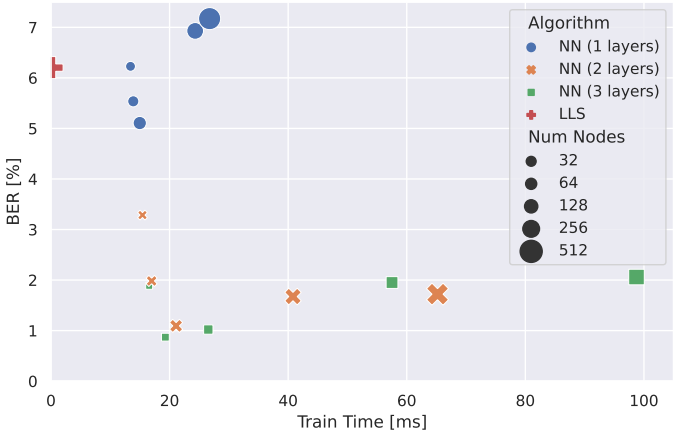

Fig. 3. Average BER over training time for different network dimensions.

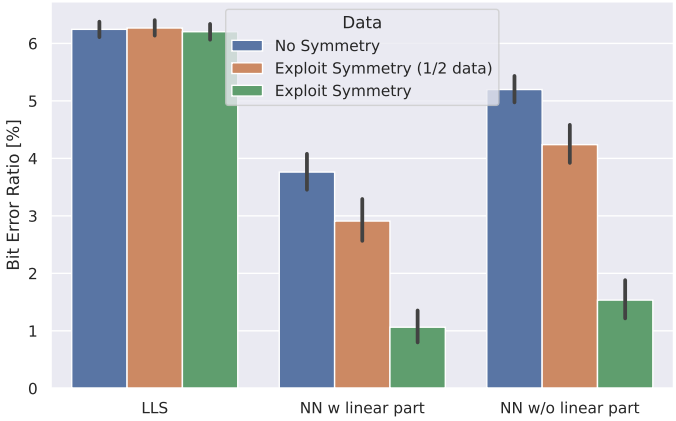

Fig. 4. Average BER of decoded user 4 when (not) exploiting complex symmetry. Error bars show the ±1SD confidence interval.

the performance of the model, when being trained with only half the data set, so that the number of training samples is identical to training without this scheme. For the LLS algorithm, the performance gain is negligible as the number of training samples is sufficient to achieve close to optimal performance for all schemes. For the NN algorithm however, there is a big performance gain, which clearly demonstrates the benefits of this approach. The fact that the simulations with half the data set also show significantly improved performance compared to the original simulation, clearly demonstrates that the equivalence (2) also holds for non-linear functions similar to our NN and that the imposed constraints improve the performance of the network. Additionally, the data shows that adding the linear part significantly improves the performance compared to a conventional NN structure with only a series of dense layers and activations.

In Figure 5, we examine the performance for different noise levels. To mimic the different noise levels we add artificial additive white Gaussian noise (AWGN) to our recorded samples. The SNR is defined here as the ratio of the power of the received signal (for all users and on all antennas) compared the power of the added noise. For each data point the simulation was run with 400 different noise realisations, giving a total of

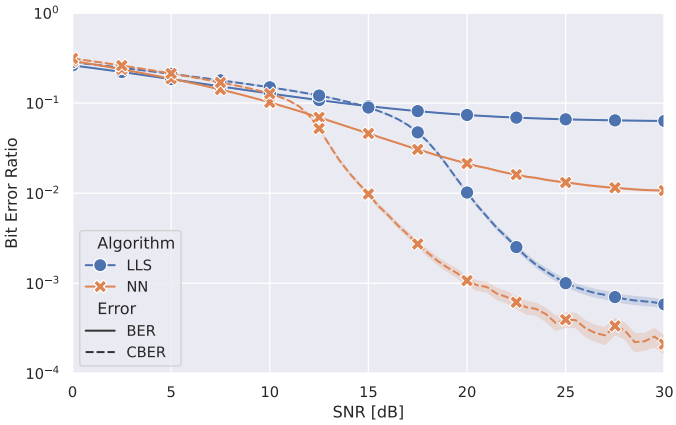

Fig. 5. Average BER and CBER of decoded user 4 for different noise levels. Shaded regions show the ±1SD confidence interval.

20 000 simulation runs per data point. The SNR of our physical system was measured to be 40 dB, which can be ignored for this experiment, as the artificially added noise is significantly stronger than the noise acquired by our measurement system. We note that the power of the signal of interest (SOI) (i.e., the received power of the fourth user, which we examine here) has a power of −12 dB relative to the received power of all users, so the SNRs for our SOI are actually 12 dB lower.

For SNRs above 5 dB the NN outperforms the LLS algorithm. The CBER of the NN drops towards an error floor about 5 dB earlier than LLS. In the high-SNR regime the NN algorithm achieves almost an order of magnitude less uncoded BER, and also a significantly lower CBER. The error floor of both algorithms is most likely due to residual interference, which can occur due to non-linearities in the radio receiver or algorithms that are not powerful enough to remove all interference (possibly due to an insufficient number of training samples).

A. Open Challenges

We empirically observed that in the low-SNR regime, the NN tends to overfit for some training cases. As a result, the learned NN can not properly remove the multi-user interference. This can have an impact on the demapper, leading to a very asymmetric distribution of log likelihood ratios (LLRs), which in turn will severely degrade the performance of the LDPC decoder. In these rare cases, the proposed NN can have a significantly higher CBER, while still having a lower BER than that of the LLS algorithm. We would like to emphasize that this challenge is mainly driven by the fact that the training must be done with a relatively small number of samples. However, it remains an open challenge to either find a more robust NN architecture or training techniques to reduce the impact of overfitting while maintaining the average performance.

VI. Conclusion

We proposed to use NNs for NOMA. By adding a linear part with LLS initialization and by exploiting the symmetry in the IQ samples, our proposed NN structure allows for multi-user detection with very small error despite its relatively small dimensions. By utilizing an optimized, fully-fused GPU implementation we achieve execution times that are several orders of magnitude smaller than those of a Tensorflow implementation and also small enough to perform the detection in the 1 ms time-frame of typical ultra-low latency (ULL) systems. Having shown the feasibility of ULL detection, in future research, we will address the training time, e.g., by tracking the channel and thus amortizing the training time across frames.

References

[1] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, October 2017.

[2] Y. Wang, B. Ren, S. Sun, S. Kang, and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Communications*, vol. 13, no. Supplement2, pp. 52–66, N 2016.

[3] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[4] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) in cellular uplink and downlink: Challenges and enabling techniques," *CoRR*, vol. abs/1608.05783, 2016.

[5] X. Su, H. Yu, W. Kim, C. Choi, and D. Choi, "Interference cancellation for non-orthogonal multiple access used in future wireless mobile networks," *EURASIP journal on wireless communications and networking*, vol. 2016, no. 1, pp. 1–12, 2016.

[6] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[7] D. A. Awan, "Robust learning in wireless networks: efficacy of models and prior knowledge in learning from small sample sets," Doctoral Thesis, Technische Universität Berlin, Berlin, 2021.

[8] E. Bjornson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. Now Foundations and Trends, 2017.

[9] M. Mehlhose, D. Schäufele, D. A. Awan, G. Marcus, N. Binder, M. Kasparick, R. L. Cavalcante, S. Stańczak, and A. Keller, "GPU-accelerated partially linear multiuser detection for 5G and beyond URLLC systems," *IEEE Access*, 2022.

[10] K. Slavakis, S. Theodoridis, and I. Yamada, "Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4744–4764, 2009.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[12] D. G. Luenberger, *Optimization by vector space methods*. John Wiley & Sons, 1997.

[13] S. Verdu *et al.*, *Multiuser detection*. Cambridge university press, 1998.

[14] T. Müller, F. Rousselle, J. Novák, and A. Keller, "Real-time neural radiance caching for path tracing," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 36:1–36:16, Aug. 2021.

[15] N.A.T. GmbH NAMC-SDR. [Online]. Available: http://www.nateurope.com/products/NAMC-SDR.html

[16] M. Mehlhose, D. A. Awan, R. L. G. Cavalcante, M. Kurras, and S. Stanczak, "Machine learning-based adaptive receive filtering: Proof-of-concept on an SDR platform," in *IEEE International Conference on Communications*, 2020, pp. 1–5.

[17] J. Hoydis, S. Cammerer, F. A. Aoudia, A. Vem, N. Binder, G. Marcus, and A. Keller, "Sionna: An open-source library for next-generation physical layer research," *arXiv preprint arXiv:2203.11854*, 2022.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.