User Clustering for Rate Splitting using Machine Learning

Roberto Pereira[1], Anay Ajit Deshpande[2], Cristian J. Vaca-Rubio[3],
Xavier Mestre[1], Andrea Zanella[2], David Gregoratti[4], Elisabeth de Carvalho[3], Petar Popovski[3]

***Abstract*—Hierarchical Rate Splitting (HRS) schemes proposed in recent years have shown to provide significant improvements in exploiting spatial diversity in wireless networks and provide high throughput for all users while minimising interference among them. Hence, one of the major challenges for such HRS schemes is the necessity to know the optimal clustering of these users based only on their Channel State Information (CSI). This clustering problem is known to be NP hard and, to deal with the unmanageable complexity of finding an optimal solution, in this work a scalable and much lighter clustering mechanism based on Neural Network (NN) is proposed. The accuracy and performance metrics show that the NN is able to learn and cluster the users based on the noisy channel response and is able to achieve a rate comparable to other more complex clustering schemes from the literature.**

***Index Terms*—User grouping, latency reduction, machine learning, MIMO.**

I. Introduction

Multi-antenna radio technologies have shown to enhance spectral efficiency while ensuring connectivity to a large number of devices. Different encoding schemes such as Dirty Paper Coding (DPC) have been designed to achieve the multi-antenna channel capacity [1]. However, due to the high computational complexity as well as the need for precise Channel State Information (CSI), there has been much focus of research on sub-optimal solutions which combine Superposition Coding (SC) and spatial processing such as Non-Orthogonal Multiple Access (NOMA) [2]. Additionally, as these mechanisms tend to fully decode interference, the uncertainty over CSI directly affects interference cancellation among different users. Hence, the authors of [3] have recently proposed Rate Splitting Multiple Access (RSMA) as a non-orthogonal transmission scheme that partially decodes interference and partially treats it as noise thus further improving multiplexing gains. For 1-layer Rate Splitting (RS), the message intended to each user is divided into a common (s_c) and private (s_p) parts encoded separately. In order for this transmission scheme to work, it is necessary to ensure that every user perfectly decodes the common message. This is often tackled by allocating a larger fraction of the total power to the common message. In the presence of a large number of receivers, this condition limits the total rate by the minimal common rate achieved in the whole system[1]. Hence, in the presence of several users, the power assigned to each s_p is reduced, leading to a degradation in communication rate.

In these conditions, relying on multiple common streams (generalised rate splitting) leads to higher multiplexing gains, but at the cost of high complexity at the decoder caused by the several layers of Successive Interference Cancellation (SIC) [3]. To tackle the increasing complexity of generalised RS while having small loss in multiplexing, the authors in [4] consider a 2-layer Hierarchical Rate Splitting (HRS) transmission mechanism. In this scenario, users are considered to be divided into G groups and required to decode three messages: two common messages and a private message. One of the common messages (outer common - s_{oc}) is encoded using a codebook shared among all the user while the other one (inner common - $s_{ic,g}$) is encoded by a codebook share only among users in a specific group. But when the groups are orthogonal, i.e. the users are sufficiently separated spatially, optimal communication happens when inter-group and intra-group interference are reduced to a level that it can be completely distinguished from the intended signals.

But, to minimise the interference and maximise the rate using HRS, the Base Station (BS) is required to know what can be referred to as the optimal clustering scheme, i.e., the one that maximises the total communication rate. Unfortunately, finding this optimal clustering scheme is an NP hard problem which often requires an exhaustive search. Thus, it becomes extremely hard to come up with an optimisation mechanism that maximises the communication rate using HRS while also considering the clustering options as an optimisation variable. Hence, in this work, we propose a learning mechanism capable of directly learning (or approximating) the optimal clustering option from the imperfect CSI.

II. System Model

Consider a downlink transmission scenario where N single-antenna user equipment (UEs) receive messages from a base station (BS) over a spatially correlated Rayleigh-fading channel. We further assume this BS to be equipped with an antenna

The authors' affiliations and emails are as follows:
[1]ISPIC, Centre Tecnològic Telecomunicacions Catalunya, Barcelona, Spain. Emails: {rpereira, xmestre}@cttc.es
[2]Department of Information Engineering, University of Padova, Padova, Italy. Emails: {deshpande, zanella}@dei.unipd.it
[3]Department of Electronic Systems, Aalborg University, Aalborg, Denmark. Emails: {cjvr, edc, petarp}@es.aau.dk
[4]Software Radio Systems, Barcelona, Spain, Email: david.gregoratti@ieee.org
This work has been partially funded by the European Commission under the Windmill project (contract 813999) and the Spanish government under the Aristides project (RTI2018-099722-B-I00).

[1]This happens regardless of the number of antennas at the transmitter. Instead, this is a consequence of power allocation to reduce interference among different users.

with M isotropic antenna elements. Moreover, let these UEs be partitioned into $G \geq 1$ disjoint clusters. So, the signal $\mathbf{y} \in \mathbb{C}^N$ received by all the users is given by

$$\mathbf{y} = \mathbf{H}^{\mathrm{H}}\mathbf{x} + \mathbf{n} \tag{1}$$

where, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^{\mathrm{T}} \in \mathbb{C}^{M \times N}$ contains the stacked channels of all the $k = \{1, \ldots, N\}$ UEs, $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I}_N)$ is an additive white Gaussian noise vector and $\mathbf{x} \in \mathbb{C}^M$ is the combined signal

$$\mathbf{x} = \sqrt{p_{oc}}\mathbf{w}_{oc}s_{oc} + \sum_{g=1}^{G} \mathbf{B}_g \left(\sqrt{p_{ic,g}}\mathbf{w}_{ic,g}s_{ic,g} + \sqrt{p_{gk}}\mathbf{W}_g\mathbf{s}_g\right) \tag{2}$$

where p_{oc}, $p_{ic,g}$ and p_{gk} are the power allocated to the outer common message $s_{oc} \in \mathbb{C}$, inner common messages $\mathbf{s}_{ic} \in \mathbb{C}^G$ and the private messages $\mathbf{s}_g \in \mathbb{C}^{N_g}$, respectively. $\mathbf{B}_g \in \mathbb{C}^{M \times b_g}$ is the group outer precoder designed from the gth group channel's second order statistics and dependent on the integer design parameters b_g rank of the channel covariance matrix. By knowing the UEs that belong to the gth cluster, the matrix $\mathbf{H}_g = [\mathbf{h}_{g,1}, \ldots, \mathbf{h}_{g,N_g}]^{\mathrm{T}} \in \mathbb{C}^{M \times N_g}$ contains the stacked channels of all the N_g UEs that belong to the gth cluster. The downlink fading channel $\mathbf{h}_{g,k} \in \mathbb{C}^M$ associated to the kth user of the gth class can be factored out as

$$\mathbf{h}_{g,k} = \mathbf{R}_g^{\frac{1}{2}}\mathbf{g} = \mathbf{U}_g\mathbf{\Lambda}_g^{\frac{1}{2}}\mathbf{g}_k \tag{3}$$

where $\mathbf{R}_g \in \mathbb{C}^{M \times M}$ is the channel correlation matrix, $\mathbf{U}_g \in \mathbb{C}^{M \times M}$ a unitary matrix containing its eigenvectors, $\mathbf{\Lambda}_g \in \mathbb{C}^{M \times M}$ a diagonal matrix with its associated eigenvalues and $\mathbf{g}_k \in \mathbb{C}^M$ has Gaussian independent and identically distributed (i.i.d.) entries with zero mean and unit variance which describe the complex path gains.

In principle, the covariance matrices are directly dependent on the angular response of the channels [2]. Unfortunately, in a more realistic environment, due to limited feedback, the BS only observes an imperfect estimation of the channel (3). Following [5], we model this imperfection as the sum of a channel and a noise generated from the same subspace

$$\hat{\mathbf{h}}_{g,k} = \mathbf{U}_g\mathbf{\Lambda}_g^{\frac{1}{2}}\hat{\mathbf{g}}_k = \mathbf{U}_g\mathbf{\Lambda}_g^{\frac{1}{2}}\left(\sqrt{1-\tau^2}\mathbf{g}_k + \tau\mathbf{z}_k\right) \tag{4}$$

where $\mathbf{z}_k$ has i.i.d entries and $\tau \in [0, 1]$ indicates the quality of the instantaneous channel. For instance, $\tau = 0$ leads to a perfect channel estimation, i.e., $\hat{\mathbf{h}}_{g,k} = \mathbf{R}_g^{\frac{1}{2}}\mathbf{g}_k$ while $\tau = 1$ leads to an uncorrelated channel in the subspace spanned by $\mathbf{U}_g$, i.e., $\hat{\mathbf{h}}_{g,k} = \mathbf{R}_g^{\frac{1}{2}}\mathbf{z}_k$ for uncorrelated $\mathbf{g}_k$ and $\mathbf{z}_k$.

A. HRS Transmission Mechanism

HRS transmission design is defined based on the combined transmission signal $\mathbf{x}$ from (2). To determine the transmission signal $\mathbf{x}$, we obtain the precoder $\mathbf{B}_g$ following [4], [5], so that the group effective channel $\tilde{\mathbf{H}}_g^{\mathrm{H}} = \hat{\mathbf{H}}_g^{\mathrm{H}}\mathbf{B}_g \in \mathbb{C}^{b_g \times N_g}$ represents the projection of $\hat{\mathbf{H}}_g$ onto the b_g–dimensional subspace orthogonal to the $r^* = \sum_{l \neq g}^{G} r_g$ singular vectors associated to the r_g largest singular values of each of the interference groups. In order to distinguish all the N_g users in the group we must have $N_g \leq b_g$, i.e., enough degrees of freedom in the b_g–dimensional subspace. Unfortunately, it is not possible to choose b_g and r_g indiscriminately large as one constrains the growth of the other. Specifically, as there exists at most M singular vectors at each group, we have that $N_g \leq b_g \leq M - r^*$. Consequently, a large number of groups leads to less freedom on the choice of both b_g and r_g.

Moreover, $\mathbf{w}_{oc}$, $\mathbf{w}_{ic,g}$ and $\mathbf{w}_{gk} = [\mathbf{W}_g]_k$ are the unit norm precoders associated to the instantaneous outer common, inner common and private messages, respectively. We can design $\mathbf{W}_g = \xi_g \left(\tilde{\mathbf{H}}_g\tilde{\mathbf{H}}_g^{\mathrm{H}} + \varepsilon\mathbf{I}_{b_g}\right)^{-1}\tilde{\mathbf{H}}_g$, given a total transmission power P, as a Regularized Zero Forcing (RZF) precoder to allow distinguishing between the N_g users within the gth group by reducing the interference among the private messages in this group [5]. The parameter ξ_g is the power normalisation factor which normalizes $||\mathbf{W}_g||_2$ to the unit. Likewise, ε is also a normalisation parameter. Similarly, $\mathbf{w}_{ic,g} = \xi_{ic,g}\sum_{k=1}^{N_g}\mathbf{w}_{gk}$ is the equally weighted Matched Beamforming (MBF) built as a linear combination of the private precoders of the gth group where $\xi_{ic,g}$ is a normalisation parameter. Finally, the outer common precoder $\mathbf{w}_{oc} = \xi_{oc}\sum_{g=1}^{G}\sum_{k=1}^{N_g}\mathbf{B}_g\tilde{\mathbf{h}}_{gk}$ is also designed as a weighted MBF, but to handle inter-group power leakage where ξ_{oc} is a normalisation paramater. Notice that it is essential to reduce inter-group interference in order to guarantee communication. Specifically, when group leakage is completely nulled out, there is no need for $\mathbf{w}_{oc}$ and communication happens over G parallel 1-layer RS streams.

To allocate power among the different messages, we further design two parameters $\alpha, \beta \in (0, 1]$. The first one α represents the fraction of the total power P allocated to the outer common message. And the latter, the fraction of the remaining power allocated to the inner common message. Combining these, we have $p_{oc} = \alpha P$, $p_{ic,g} = \frac{(1-\alpha)\beta P}{G}$ and $p_{gk} = \frac{(1-\alpha)(1-\beta)P}{N_g}$. In this work we perform a brute force search to find the optimal α and β for every channel realisation.

As mentioned above, at the receiver side, the kth user associated to the gth group decodes its message in a 2-step successive interference cancellation fashion. In the first step, the user decodes the outer common message (s_{oc}) and removes it from the received signal. The group's inner common codeword is then decoded after applying SIC. After successfully decoding both common messages, each private message is extracted by considering all other private messages as interference. As a result, the Signal-to-Interference Plus-Noise Ratio (SINR) to each of these messages is written as

$$\gamma_{gk}^{oc} = \frac{p_{oc}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{oc}|^2}{1 + I_{gk}} \tag{5}$$

$$\gamma_{gk}^{ic} = \frac{p_{ic}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{ic,g}|^2}{1 + I_{gk} - p_{ic}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{ic,g}|^2} \tag{6}$$

$$\gamma_{gk}^{p} = \frac{p_{gk}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{gk}|^2}{1 + I_{gk} - \left(p_{ic}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{ic,g}|^2 + p_{gk}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{gk}|^2\right)} \tag{7}$$

where

$$I_{gk} = \sum_{l=1}^{G} p_{ic,l} |\mathbf{h}_{gk}^{\mathrm{H}} \mathbf{B}_l \mathbf{w}_{ic,l}|^2 + \sum_{l=1}^{G} \sum_{k=1}^{N_g} p_{lk} |\mathbf{h}_{gk}^{\mathrm{H}} \mathbf{B}_l \mathbf{w}_{lk}|^2$$

is the combination of all interference leaked from other users and groups. Finally, we can describe the achievable rate as the combination of the smallest achievable outer common rate among all users $R_{oc} = \min_{gk} \log_2(1 + \gamma_{gk}^{oc})$, the minimal inner common rate per group $R_{ic} = \sum_{g=1}^{G} \min_{k} \left(\log_2(1 + \gamma_{gk}^{ic}) \right)$ and the sum of the rate achievable at all private messages $R_p = \sum_{g=1}^{G} \sum_{k=1}^{N_g} \log_2(1 + \gamma_{gk}^{p})$. Then the total rate is the sum of these components, i.e, $R = R_{oc} + R_{ic} + R_p$.

III. User Clustering and Dataset Definition

As it becomes evident from the discussion above, and further supported in our results, choosing an appropriate clustering is crucial to take full advantage of two-tier precoding mechanisms, such as HRS [4], [5]. One can rely on extensive search in order to find the optimal clustering mechanism. However, this is an NP hard task as the number of ways that a set can be partitioned into nonempty sets is given by the Bell number which grows almost exponentially with N, i.e., the number of elements in the set. Moreover, in our scenario, many of these partitions lead to vanishing communication rates due to high interference. Therefore, in this work, we rely on (possible suboptimal) clustering options obtained from an agglomerative hierarchical clustering mechanism [6].

A. User Clustering

To devise the clustering mechanism, we define a bottom up approach where the objective is to combine clusters (groups of users in the wireless network) according to their similarity. Initially, each user is associated to a singleton cluster. At each step of the hierarchical clustering algorithm, the pair of users/clusters with highest similarity (according to a criterion discussed later) is then merged. As a result, after each merge we obtain a new clustering option and evaluate the rate achieved considering this new option. This process continues until we have evaluated all levels in the hierarchy. Notice that in this agglomerative mechanism there exist only $N+1$ (total number of users plus one) possible clustering options, one for each level in the hierarchy. These, however, are often relevant clustering options as each cluster only contains elements that are particularly similar to each other.

In [6], we consider the similarity measure between two channel matrices based on how close the principle angles of the subspaces spanned by their column-spaces are. Specifically, for two clusters of size N_k and N_j, we take the projection-Frobenius (PF) similarity

$$s_{k,j} = \frac{\mathrm{tr}(\hat{\mathbf{P}}_k \hat{\mathbf{P}}_j)}{\min(N_k, N_j)}, \tag{8}$$

where $\hat{\mathbf{P}}_j$ is the projection matrix given by,

$$\hat{\mathbf{P}}_j = \hat{\mathbf{H}}_j (\hat{\mathbf{H}}_j^{\mathrm{H}} \hat{\mathbf{H}}_j)^{-1} \hat{\mathbf{H}}_j^{\mathrm{H}} \tag{9}$$

which describes the first N_j left singular vectors of the kth group of channels. Moreover, to improve clustering results for $N_j \neq N_k$, we follow a statistical analysis of the quantity in (8) and further define the normalised similarity measure

$$\hat{s}_{k,j} = \frac{s_{k,j} - \eta_{k,j}}{\sigma_{k,j}} \tag{10}$$

based on its asymptotic mean $\eta_{k,j}$ and variances $\sigma_{k,j}^2$ defined as in [6]. However, this normalisation step is only possible for $M > N_j + N_k$, otherwise, we follow the projection-Frobenius similarity described in (8).

B. Dataset Definition

We design the dataset used for this work by devising channel matrices from (4) and clustering them according to the scheme described above. We consider four possible covariance matrices to which channels are randomly associated. Consequently, for different samples, we might obtain a different number of users associated to a specific covariance matrix. Notice that, this is not a cluster assignment, but merely a way to generate random channels. These covariance matrices are obtained by considering the azimuth angles $\theta_g = -\frac{\pi}{2} + \frac{\pi}{3}(g-1)$ and the constant angular spread $\Delta_g = \frac{\pi}{6}$. Moreover, we further assume the BS to be equipped with a Uniform Circular Array antenna.

Concretely, we design 3 different configurations based on the choices of the number of antennas at the BS: 1) $N > M$, 2) $N = M$ and 3) $N < M$. Moreover, we evaluate these configurations for two different system loads, based on the number of users $N \in \{8, 12\}$. Specifically, we have $M \in \{6, 8, 12\}$ for $N = 8$ and $M \in \{6, 12, 16\}$ for $N = 12$. As a result, we have 6 different scenarios. For each these we generate $S = 10.000$ random samples, each sample containing both imperfect and perfect CSIT of equal size $N \times M$ and the clustering scheme that maximises the rate based on the hierarchical clustering mechanism. As a result of this randomness, for each scenario we obtain more than $G^* = 200$ possible clustering options, thus, leading to very imbalanced datasets. To diminish this effect, for each scenario, we sub-sample the data such that only relevant classes are left, i.e., we discard classes that achieve less than 25% of the average rate of the scenario and have less than 50 samples. Moreover, to further balance the data, we crop the maximum number of samples in each class to be at most to 200. As a result, for each scenario, we still obtain an imbalanced dataset with approximately $G^* = 50$ classes, each containing at least 50 samples and at most 200 samples.

Finally, to compensate for this drop in the number of samples, we further augment the dataset of each configuration by randomly shuffling users that belong to the same cluster. This is a natural extension of this dataset as clustering should be indifferent to the ordering of the users.

IV. Machine learning model and training

We solve the classification problem presented in the previous section by designing a shallow neural network. We used the Keras library, so we describe the layers with their notation [7]. For each scenario, we divide our dataset into training,

validation and test sets in a proportion of 80/10/10. During the training procedure, we use the validation set to tune the corresponding hyper-parameters. Our model is defined as a shallow neural network following the parameters from Table I. The output layer consists of G^* neurons with a *softmax* activation that correspond to each cluster where G^* is the total number of classes in the scenario. The softmax function in the output layer is used to obtain the probability of a user belonging to a specific cluster and it is given by

$$\sigma(\mathbf{Z})_g = \frac{e^{z_g}}{\sum_{j=1}^{G} e^{z_j}} \tag{11}$$

where $\mathbf{Z}$ is the input vector from the previous hidden layer, z_g the g-th element and the denominator sum is the normalisation factor to ensure the output is into the range of $[0, 1]$. Then, by selecting the maximum, we can obtain the highest probability that users are clustered in a particular way. For the training procedure, we use the Adam optimiser with a learning rate of 10^{-3}, we train for 50 epochs and use a batch size of 128 samples. For our multi-class classification task, we aim to minimise the categorical cross-entropy loss [8] given by

$$\mathcal{L}(y_g, \hat{y}_g) = -\sum_{g=1}^{G^*} y_g \log \hat{y}_g. \tag{12}$$

where y_g and $\hat{y}_g$ are the groundtruth and NN score for each class. This loss is a very good measure of how distinguishable two discrete probability distributions are from each other. In this context, the vector $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_g] \in \mathbb{R}^{G^*}$ has entries which represent the probability that users are clustered in a specific manner and the sum of all entries is one. The accuracy of a model is often defined in terms of the entry with highest probability, this is often, called *top-1* accuracy. In our scenario, there exist several options which achieve the (close to) maximum rate. Therefore, it is also interesting to analyse the *top-k* accuracy of our model, i.e., if the desired clustering option is among the k most probable outputs.

Finally, we emphasise that we are applying a shallow neural network which contains only a small number of learnable parameters. This is designed as a consequence of our devised dataset. Recall that we have specifically defined it to be imbalanced and with a small number of samples to each class ($[50, 200]$). Nonetheless, as we present below, this network is capable of learning the relationship between the different channel matrices and directly output the desired clustering option that maximises transmission using HRS.

V. Performance Analysis

In this section, we evaluate the performance of the presented Neural Network (NN) method in comparison to RS under different scenario configurations. These numerical simulations are carried out in a MATLAB environment. The necessary configuration parameters are defined in Table I.

In order to validate the learning of the NN, we compare the rate achieved using the NN predicted classes and different RS clustering options. To perform a complete evaluation, we determine the rate achieved by the following solutions,

TABLE I: Parameters of the Simulations

Simulation Parameter	Simulation Value
Antenna Configuration	Uniform Circular Array
Angular Spread (Δ_g)	$\pi/6$
Number of Unique Distributions	4
Channel Quality (τ^2)	0.4
Dominant Eigenvectors ($b_g = r_g$)	$\lfloor M/G \rfloor$
Channel Quality (r_g)	0.4
Number Shuffling	10
Number of Neurons in NN	$\{256, 128\}$
NN Learning Rate	10^{-3}
NN Training Epochs	50
NN Training Batch Size	128
NN Input Layer Activation Function	ReLu Function
NN Hidden Layer Activation Function	ReLu Function
NN Output Layer Activation Function	Softmax Function
NN Loss Function	Categorical Cross-entropy Loss

- Hierarchical Clustering - Hierarchical Rate Splitting (HC): The users are clustered according to the clustering mechanism defined in Sec. III, the group with higher communication performance is selected;
- Neural Network - Hierarchical Rate Splitting (NN): Proposed NN based clustering;
- Universal Cluster (UNI): All users are clustered into one single cluster;
- Singleton Cluster (SING): Each cluster contains only single user.

As mentioned above, we consider three scenarios to evaluate the clustering solutions 1) $M < N$, 2) $M = N$ and 3) $M > N$. Hence, for $N = 8$, we determine the rate achieved for $M \in \{4, 8, 12\}$ and for $N = 12$, we determine the rate achieved for $M \in \{6, 12, 16\}$. Then, we compare the different clustering techniques mentioned before based on the rate achieved. Fig. 1 shows the rate achieved for all four clustering techniques for the different values of M and N. Each box plot shows the rate obtained for different realisations of the channel. The median rate is presented by a horizontal line through box and the top and bottom of the box are the 75th and 25th percentile rate (i.e. rate achieved by 75% and 25% of the scenarios). Lastly, the extremities of the boxplot refer to the 1% and 99% and the red plus indicators in the boxplot denote the outlier rate values. Notice that the rate achieved by HC-HRS and NN-HRS is approximately similar while both clustering techniques outperform UNI and SING. This is due to the fact that with a noisy channel, it is really difficult to generate accurate precoders that can maximise the rate achieved and minimise the inter-group and intra-group interferences. Additionally, the NN-HRS only receives the instantaneous noisy channel as an input and determines its clustering solution while HC-HRS needs to iteratively determine the similarity between different channels making it considerably slower when compared to the NN solution. Moreover, for SING, the choice of parameters b_g and r_g seems to harm the performance. We recall that both parameters are integers thus are susceptible to the trade-off between M and G. For instance, for $G = N = 8$ and $M = 12$, there exist only one viable option of r_g, i.e., $r_g = \lfloor M/G \rfloor = 1$. Alternatively, we could select four ($\mathrm{mod}(M, G)$) groups to have $r_g = 2$,

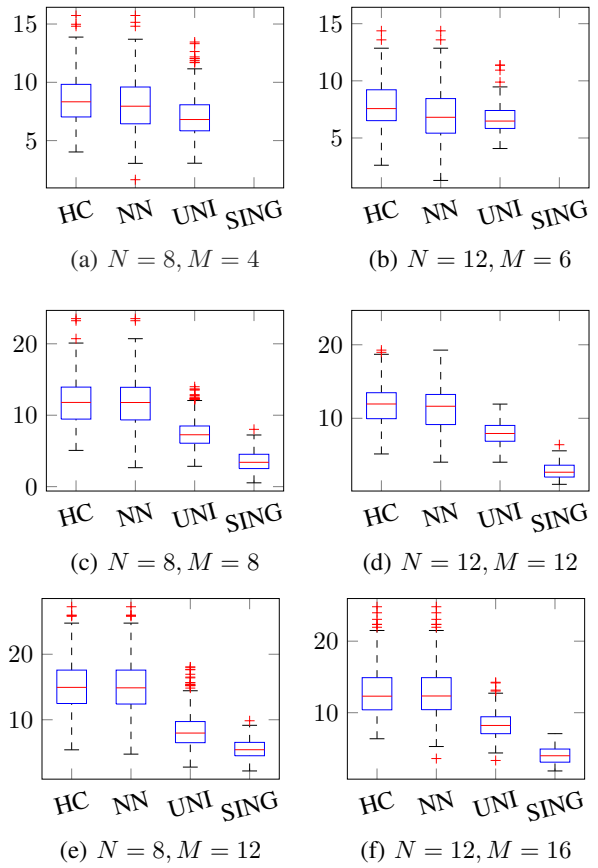

Fig. 1: Spectral efficiency (bps/Hz) achieved for clustering mechanisms using HRS.

but this requires further processing on the choice of these groups. As a consequence, we obtain similar rates for $N = 8$ users served with $M = 8$ or $M = 12$. Similar consequences are obtained for $N = 12$. Moreover, for $G = N > M$, we have $r_g = \text{mod}(M/N) = 0$ what makes impossible to derive meaningful precoders Fig 1(a)-(b). In contrast to that, the other three techniques, which consider clustering, do not suffer from this trade-off between G, r_g and M. Instead, even for $N > M$ we still achieve reasonable spectral efficiency.

Finally, we analyse the capability of the shallow NN to learn the grouping classification task as described above. To do so, we first analyse the accuracy of the network for class prediction. Recall that, here, a class represents a different clustering option. Table II presents, in percentage, the results obtained by training different NN according to the configuration parameters in Table I for different number of users (N) and antennas in the BS (M). The validation column contains the final classification accuracy in the validation dataset and indicates some learning capability in untrained data. During our experiments we noticed that different points of the same dendrogram might result in similar communication rates, i.e., there might exist different clustering options which achieve

TABLE II: Summary of Results

N / M	Validation *(top-1)*	Test *(top-1)*	Test *(top-3)*	Test *(top-5)*	Test Relative Rate
8 / 4	65.38%	65.37%	85.22%	90.48%	94.12%
8 / 8	98.3%	92.0%	96.3%	97.7%	99.0%
8 / 12	96.9%	92.2%	97.0%	98.2%	99.5%
12 / 6	71.45%	35.6%	65.62%	77.75%	89.99%
12 / 12	98.7%	86.2%	96.8%	98.9%	93.5%
12 / 16	99.18%	95.62%	98.32%	93.32%	99.77%

the same rate. Therefore, for the test dataset, we show the *top-1*, *top-3* and *top-5* classification accuracy. Despite the fact that performance in *top-1* accuracy might be considered poor, the *top-5* results are, often, above 90%. Finally, the last row compares the communication rate decay (in $\%$) if using the *top-1* option from the NN. Results show that, except in the cases where $N > M$, on average, the rate drops 2.5% which is an acceptable loss when compared to the complexity of the original problem. Moreover, we can infer from these results that the NN is capable of learning the maximum clustering option or clusters that approximate this option. In other words, it is capable to learn the relationship between different users directly from their channel matrices and cluster the users with a high degree of accuracy for most scenarios and finally achieve a rate comparable to more complicated similarity-based HC-HRS.

VI. Conclusion

In this work, we have proposed a NN based clustering technique that learns and clusters users based on instantaneous noisy channel to maximise the rate achieved using Hierarchical Rate Splitting mechanism. The proposed technique is defined based on a shallow NN architecture thereby making it extremely quick to learn and cluster the users based on the instantaneous noisy channel. The proposed technique is able to achieve a rate comparable with current works while being less complex compared to other techniques. Furthermore, this also helps to investigate further complex NN structures such as Graph NN which can learn covariances between different users to define clustering.

References

[1] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Transactions on information theory*, vol. 50, no. 9, pp. 1875–1892, 2004.
[2] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.
[3] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP journal on wireless communications and networking*, vol. 2018, no. 1, pp. 1–54, 2018.
[4] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A Rate Splitting Strategy for Massive MIMO With Imperfect CSIT," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4611–4624, 2016.
[5] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE transactions on information theory*, vol. 59, no. 10, pp. 6441–6463, 2013.
[6] R. Pereira, X. Mestre, and D. Gregoratti, "Subspace Based Hierarchical Channel Clustering in Massive MIMO," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021.
[7] F. Chollet *et al.*, "Keras," https://keras.io, 2015.
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.