Deep Learning for Uplink CSI-based Downlink Precoding in FDD massive MIMO Evaluated on Indoor Measurements

Florian Euchner, Niklas Süppel, Marc Gauger, Sebastian Dörner, Stephan ten Brink
Institute of Telecommunications, Pfaffenwaldring 47, University of Stuttgart, 70569 Stuttgart, Germany
{euchner,sueppel,gauger,doerner,tenbrink}@inue.uni-stuttgart.de

***Abstract*—When operating massive multiple-input multiple-output (MIMO) systems with uplink (UL) and downlink (DL) channels at different frequencies (frequency division duplex (FDD) operation), acquisition of channel state information (CSI) for downlink precoding is a major challenge. Since, barring transceiver impairments, both UL and DL CSI are determined by the physical environment surrounding transmitter and receiver, it stands to reason that, for a static environment, a mapping from UL CSI to DL CSI may exist. First, we propose to use various neural network (NN)-based approaches that learn this mapping and provide baselines using classical signal processing. Second, we introduce a scheme to evaluate the performance and quality of generalization of all approaches, distinguishing between known and previously unseen physical locations. Third, we evaluate all approaches on a real-world indoor dataset collected with a 32-antenna channel sounder.**

I. Introduction and Problem Statement

Massive multiple-input multiple-output (MIMO) is widely accepted to be a crucial technology for increasing the spectral efficiency of future cellular wireless systems through spatial multiplexing. At the multi-antenna basestation (BS), it relies on precoding in the downlink (DL) direction, which requires accurate channel state information (CSI) for the channel between BS and user equipment (UE). The BS estimates CSI for the uplink (UL) channel from pilots transmitted by the UE. In time division duplex (TDD) operation, thanks to channel reciprocity, DL CSI can be directly derived from UL CSI. If, however, UL and DL channels are at different frequencies (frequency division duplex (FDD) operation), acquisition of DL CSI is challenging. Sending downlink pilots and obtaining CSI feedback from UEs produces overhead that can become prohibitively large for high numbers of antennas [1].

Even as massive MIMO was originally conceived, it was conjectured that DL CSI feedback in FDD operation could be rendered unnecessary by exploiting relationships between UL and DL CSI [2, Section VII.J]. For example, in typical radio environments, measurements have indicated that angles of arrival and departure are similar for UL and DL channels [3]. However, in environments with many scatterers and potentially more than one strong propagation path, the relationship between UL and DL CSI is no longer this simple. Under the premise that the mapping from UL CSI to DL CSI is bijective, which is reasonable to assume for many practical environments

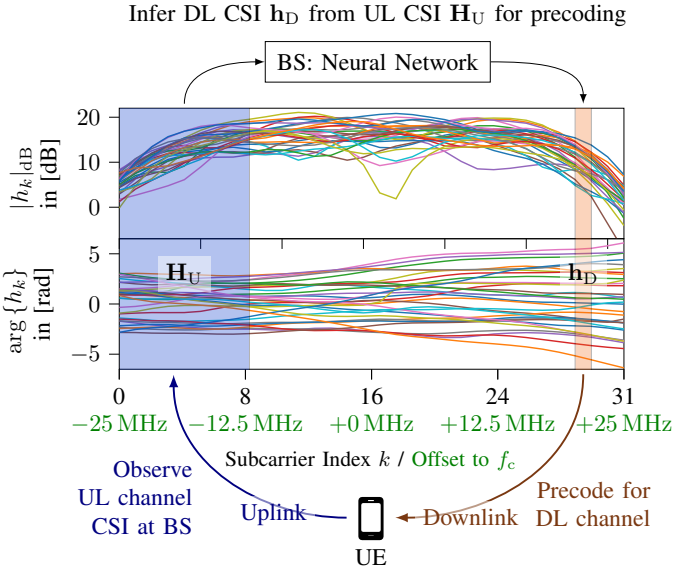


Fig. 1: Basic principle of operation

[4], a dense neural network (DNN) is capable of learning this mapping. This learning-based approach, illustrated in Fig. 1, has been proposed in several earlier publications and has been evaluated on simulated channel models [1] [5]. By contrast, experiments with measured channel data are rare [6] and cannot be replicated without the underlying datasets. We address these issues by making the following contributions:

- We derive upper and lower bounds for downlink precoding performance in Section II.
- In Section IV, we verify the concept of deep learning-based CSI estimation, which has primarily been developed with simulated channels, on a measured, publicly available CSI dataset, which is introduced in Section III.
- In Section V, we compare the quality of DL CSI estimates for different neural network (NN) architectures for our particular dataset and propose an evaluation framework for different network architectures that takes into account the difference in the quality of estimates in previously seen and unseen regions of the physical environment[1].

[1]A tutorial for a special case of DL CSI estimation is available at https://dichasus.inue.uni-stuttgart.de/tutorials/tutorial/downlinkcsi/

II. Model, Metrics and Baselines

In the context of this work, we always consider the case of a single BS antenna array with M co-located antennas and a single UE with one antenna. We assume orthogonal frequency division multiplex (OFDM)-modulated signals for both uplink and downlink, but restrict ourselves to estimating the CSI for a single subcarrier in the downlink channel. It is important to note that this approach can easily be extended to all subcarriers in the downlink channel by using one estimator per subcarrier.

We denote the unknown channel coefficient vector for this particular DL subcarrier by $\mathbf{h}_\mathrm{D} \in \mathbb{C}^M$. We assume that UL channel coefficients $\mathbf{H}_\mathrm{U} \in \mathbb{C}^{M \times N_\mathrm{sub}}$ for all antennas and all N_sub uplink subcarriers are known to the BS. We furthermore neglect hardware impairments and noise and assume that both UL and DL CSI are determined by some latent variable $\mathbf{x}$, which captures all properties of the radio environment such as location and orientation of transmitter, receiver and scatterers:

$$f_U : \mathbf{x} \mapsto \mathbf{H}_\mathrm{U} \quad \text{and} \quad f_D : \mathbf{x} \mapsto \mathbf{h}_\mathrm{D} \tag{1}$$

In Eq. (1), f_U and f_D are deterministic mappings from environment properties $\mathbf{x}$ to UL and DL channel coefficients, respectively. If f_U is bijective, which has been argued to be probable in practical environments [4] [7], and the BS is capable of learning $f_\mathrm{D} \circ f_\mathrm{U}^{-1}$, which NNs are theoretically capable of according to the universal approximation theorem, it can compute $\mathbf{h}_\mathrm{D}$ as $\mathbf{h}_\mathrm{D} = f_\mathrm{D} \circ f_\mathrm{U}^{-1}(\mathbf{H}_\mathrm{U})$. In practice, f_U may only be bijective on a (large) subset of the domain, the universal approximation theorem only holds for arbitrarily large NN sizes, only limited training data is available and UL channel estimates are noisy, hence the learned mapping $\hat{\boldsymbol{\theta}} : \mathbf{H}_\mathrm{U} \to \mathbf{w}$ will only produce an *estimate* $\mathbf{w} \in \mathbb{C}^M$, $\mathbf{w} \approx \mathbf{h}_\mathrm{D}$ for the true downlink channel $\mathbf{h}_\mathrm{D}$.

A suitable metric for the quality of the estimate $\mathbf{w}$ for one particular realization of the channel is given by the squared cosine similarity of $\mathbf{h}_\mathrm{D}$ and $\mathbf{w}$:

$$P = \frac{\left|\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}\right|^2}{\|\mathbf{h}_\mathrm{D}\|^2 \|\mathbf{w}\|^2} \tag{2}$$

In contrast to a mean squared error (MSE) metric, the expression for P in Eq. (2) has the advantage of being interpretable as the normalized received power on the considered downlink subcarrier at the UE when precoding with vector $\mathbf{w}^*$ and transmitting across the channel $\mathbf{h}_\mathrm{D}$. In this sense, a normalized received power of $P = 1$ corresponds to perfect knowledge of the downlink channel down to a global phase rotation, i.e., $\mathbf{w} = \mathrm{e}^{\mathrm{j}\varphi}\mathbf{h}_\mathrm{D}$ with arbitrary $\varphi \in \mathbb{R}$.

Eq. (2) refers to one particular downlink channel $\mathbf{h}_\mathrm{D}$ and estimate $\mathbf{w}$. In practice, $\mathbf{w}$ is estimated at the BS based on $\mathbf{H}_\mathrm{U}$, i.e., $\mathbf{w} = \hat{\boldsymbol{\theta}}(\mathbf{H}_\mathrm{U})$, and $\mathbf{H}_\mathrm{U}$ and $\mathbf{h}_\mathrm{D}$ are modelled as random variables that are jointly distributed over some distribution $\mathcal{H}$:

$$(\mathbf{H}_\mathrm{U}, \mathbf{h}_\mathrm{D}) \sim \mathcal{H}$$

To obtain a more universal indicator $\bar{P}$ for the performance of a DL CSI estimator $\hat{\boldsymbol{\theta}}$, we consider the expected value of P, i.e., the average normalized received power, over the whole distribution $\mathcal{H}$:

$$\bar{P} = \mathrm{E}_{(\mathbf{H}_\mathrm{U}, \mathbf{h}_\mathrm{D}) \sim \mathcal{H}} \left[\frac{\left|\mathbf{h}_\mathrm{D}^\mathrm{H}\hat{\boldsymbol{\theta}}(\mathbf{H}_\mathrm{U})\right|^2}{\|\mathbf{h}_\mathrm{D}\|^2 \|\hat{\boldsymbol{\theta}}(\mathbf{H}_\mathrm{U})\|^2} \right] \tag{3}$$

Without any knowledge about $\mathcal{H}$, it is still possible to achieve an average normalized received power of $\bar{P} = \frac{1}{M}$ through the use of random precoding vectors, as the following theorem will show.

Theorem 1 (Random Precoding Baseline)**.** *For any arbitrary distribution $(\mathbf{H}_\mathrm{U}, \mathbf{h}_\mathrm{D}) \sim \mathcal{H}$,* ***random precoding*** *with channel estimates $\mathbf{w} \in \mathbb{C}^M$, $\mathbf{w} := \frac{\mathbf{v}}{\|\mathbf{v}\|}$ where $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ and independent of $\mathbf{h}_\mathrm{D}$, $\mathbf{H}_\mathrm{U}$ leads to an expected received power*

$$\bar{P}_\mathrm{rand} := \mathrm{E}\left[\frac{|\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}|^2}{\|\mathbf{h}_\mathrm{D}\|^2}\right] = \frac{1}{M}. \tag{4}$$

Proof. Noticing that $\mathbf{h}_\mathrm{D}$ and $\mathbf{w}$ are independent and since $|\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}|^2 = \mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}\mathbf{h}_\mathrm{D}^\mathrm{T}\mathbf{w}^*$, we can exchange the order of the expectation operator and the scalar product sums in Eq. (4):

$$\bar{P}_\mathrm{rand} = \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{1}{\|\mathbf{h}_\mathrm{D}\|^2} \sum_{i=1}^{M}\sum_{j=1}^{M} h_{\mathrm{D},i}^* h_{\mathrm{D},j} \mathrm{E}_\mathbf{w}\left[w_i w_j^*\right]\right].$$

Next, we need to show that $\mathrm{E}[w_i w_j^*] = 0$ for $i \neq j$. For this, first note that the distribution of $\mathbf{v}$ is invariant under unitary transformations Q, and so is $\mathbf{w}$ since $Q\mathbf{w} = \frac{Q\mathbf{v}}{\|\mathbf{v}\|} = \frac{Q\mathbf{v}}{\|Q\mathbf{v}\|}$. In particular, this implies that the distributions of $\mathbf{w} = (\ldots, w_i, \ldots, w_j, \ldots)$ and $\mathbf{w}' = (\ldots, -w_i, \ldots, w_j, \ldots)$ are identical and hence $\mathrm{E}\left[w_i w_j^*\right] = -\mathrm{E}\left[w_i w_j^*\right] = 0$.

With this, $\bar{P}_\mathrm{rand}$ further simplifies to

$$\bar{P}_\mathrm{rand} = \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{1}{\|\mathbf{h}_\mathrm{D}\|^2} \sum_{i=1}^{M} |h_{\mathrm{D},i}|^2 \, \mathrm{E}_\mathbf{w}\left[|w_i|^2\right]\right].$$

For symmetry reasons, $\mathrm{E}_\mathbf{w}\left[|w_i|^2\right] = \mathrm{E}_\mathbf{w}\left[|w_j|^2\right]$ for any i, j. Hence, $\mathrm{E}_\mathbf{w}\left[\mathbf{w}^\mathrm{H}\mathbf{w}\right] = \mathrm{E}_\mathbf{w}\left[\sum_{i=1}^M |w_i|^2\right] = \sum_{i=1}^M \mathrm{E}_\mathbf{w}\left[|w_i|^2\right] = M\,\mathrm{E}_\mathbf{w}\left[|w_i|^2\right]$. Since $\mathbf{w}^\mathrm{H}\mathbf{w} = \|\mathbf{w}\|^2 = 1$, we find that $\mathrm{E}_\mathbf{w}\left[|w_i|^2\right] = \frac{1}{M}$ for any i:

$$\bar{P}_\mathrm{rand} = \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{\sum_{i=1}^M h_{\mathrm{D},i}^* h_{\mathrm{D},i}}{\|\mathbf{h}_\mathrm{D}\|^2} \, \frac{1}{M}\right]$$

Since $\sum_{i=1}^M h_{\mathrm{D},i}^* h_{\mathrm{D},i} = \|\mathbf{h}_\mathrm{D}\|^2$, all terms depending on $\mathbf{h}_\mathrm{D}$ cancel, which proves that Eq. (4) holds for arbitrary $\mathcal{H}$:

$$\bar{P}_\mathrm{rand} = \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{1}{M}\right] = \frac{1}{M} \qquad \square$$

For real channels, random precoding is not a fair benchmark to compare NN-generated estimates against, since it does not take the prior distribution of $\mathbf{h}_\mathrm{D}$ over a dataset into account. A badly designed NN could just learn the prior distribution of $\mathbf{h}_\mathrm{D}$ and not extract information from $\mathbf{H}_\mathrm{U}$. As another baseline, Theorem 2 describes a precoding technique with a constant DL channel estimate $\mathbf{w}$ that exploits a-priori information.

Theorem 2 (Principal Component Baseline)**.** *To maximize the mean normalized power $\bar{P}$ over the distribution $(\mathbf{H}_\mathrm{U}, \mathbf{h}_\mathrm{D}) \sim \mathcal{H}$ under the restriction that the DL channel estimate $\mathbf{w}$ is constant and $\|\mathbf{w}\| = 1$, $\mathbf{w}$ must be chosen such that $\mathbf{w} = \mathbf{w}_\mathrm{max}$, where $\mathbf{w}_\mathrm{max}$ is the eigenvector corresponding to the largest eigenvalue of the auto-correlation matrix $\mathbf{R} := \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{\mathbf{h}_\mathrm{D}\mathbf{h}_\mathrm{D}^\mathrm{H}}{|\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{h}_\mathrm{D}|}\right]$. We define*

$$\bar{P}_\mathrm{princ} := \max_{\|\mathbf{w}\|=1} \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{|\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}|^2}{\|\mathbf{h}_\mathrm{D}\|^2}\right].$$

Proof. The objective is to find $\mathbf{w}_\mathrm{max}$ according to

$$\begin{aligned}\mathbf{w}_\mathrm{max} &= \arg\max_{\|\mathbf{w}\|=1} \mathrm{E}_{\mathbf{h}_\mathrm{D}}\left[\frac{\mathbf{w}^\mathrm{H}\mathbf{h}_\mathrm{D}\mathbf{h}_\mathrm{D}^\mathrm{H}\mathbf{w}}{\|\mathbf{h}_\mathrm{D}\|^2}\right] \\ &= \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^\mathrm{H}\mathbf{R}\mathbf{w}.\end{aligned}$$

By taking the derivative of the Lagrange function $\mathcal{L}(\mathbf{w}) = \mathbf{w}^\mathrm{H}\mathbf{R}\mathbf{w} - \lambda(\mathbf{w}^\mathrm{H}\mathbf{w} - 1)$ with respect to $\mathbf{w}$, we find that $\mathbf{R}\mathbf{w} = \lambda\mathbf{w}$. Hence, $\mathbf{w}$ is an eigenvector of $\mathbf{R}$ and the function in

$$\mathbf{w}_\mathrm{max} = \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^\mathrm{H}\mathbf{R}\mathbf{w} = \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^\mathrm{H}\lambda\mathbf{w}$$

is maximized if $\mathbf{w}$ corresponds to the largest eigenvalue λ. □

III. Measurement Dataset

For evaluating our deep learning-based CSI estimation, we draw on a dataset measured with our own channel sounder called Distributed Channel Sounder by University of Stuttgart (DICHASUS) [8]. More specifically, we use a publicly available indoor dataset entitled *dichasus-015x* measured with an $M = 32$-antenna uniform planar array in a $6\,\mathrm{m} \times 6\,\mathrm{m}$ office room [9]. Overall, the dataset contains more than $85\,000$ position-tagged CSI datapoints captured at a carrier frequency of $f_\mathrm{c} = 1.272\,\mathrm{GHz}$. Each CSI datapoint was estimated from multiple OFDM symbols with $N_\mathrm{sub} = 1024$ subcarriers spread over a bandwidth of $50\,\mathrm{MHz}$. We averaged over batches of 32 neighboring subcarriers for the purpose of DL channel estimation, obtaining a total of 32 averaged channel coefficients.

From this large $50\,\mathrm{MHz}$ bandwidth, we collect channel coefficients within some ranges into *virtual* uplink and downlink channels. As shown in Fig. 1, we grouped the channel coefficients for averaged subcarriers 0-7 to be the virtual UL channel ($\mathbf{H}_\mathrm{U}$) and we call the channel coefficients for subcarrier 28 the virtual DL channel vector ($\mathbf{h}_\mathrm{D}$). Note that the CSI dataset was measured with all antennas in the array exclusively operated as receivers at carrier frequency f_c, but, thanks to channel reciprocity, the same channel coefficients can be assumed for the DL direction. Our choice corresponds to a virtual UL channel with a bandwidth of $12.5\,\mathrm{MHz}$ centered around $f_\mathrm{c,UL} \approx 1.2533\,\mathrm{GHz}$ and a virtual DL channel coefficient measured at carrier frequency $f_\mathrm{c,DL} \approx 1.2915\,\mathrm{GHz}$. The center frequencies of uplink channel and downlink subcarrier are separated by $f_\mathrm{c,DL} - f_\mathrm{c,UL} \approx 38.2\,\mathrm{MHz}$.

Through random precoding according to Theorem 1, it is always possible to achieve a mean received power of $\bar{P}_\mathrm{rand} = \frac{1}{32}$, $\bar{P}_\mathrm{rand}|_\mathrm{dB} \approx -15\,\mathrm{dB}$, i.e., approximately $15\,\mathrm{dB}$ less on average than is possible if the true channel vector $\mathbf{h}_\mathrm{D}$ was known by the BS. When precoding with the optimal constant DL channel estimate $\mathbf{w}_\mathrm{max}$ according to Theorem 2, we find that it is possible to achieve $\bar{P}_\mathrm{princ}|_\mathrm{dB} \approx -8.8\,\mathrm{dB}$ just by exploiting the prior distribution of the dataset. The distribution of received powers over the dataset's measurement area for this case is illustrated in Fig. 2a: Precoding with $\mathbf{w}_\mathrm{max}$ generates a single broad, forward-facing beam.

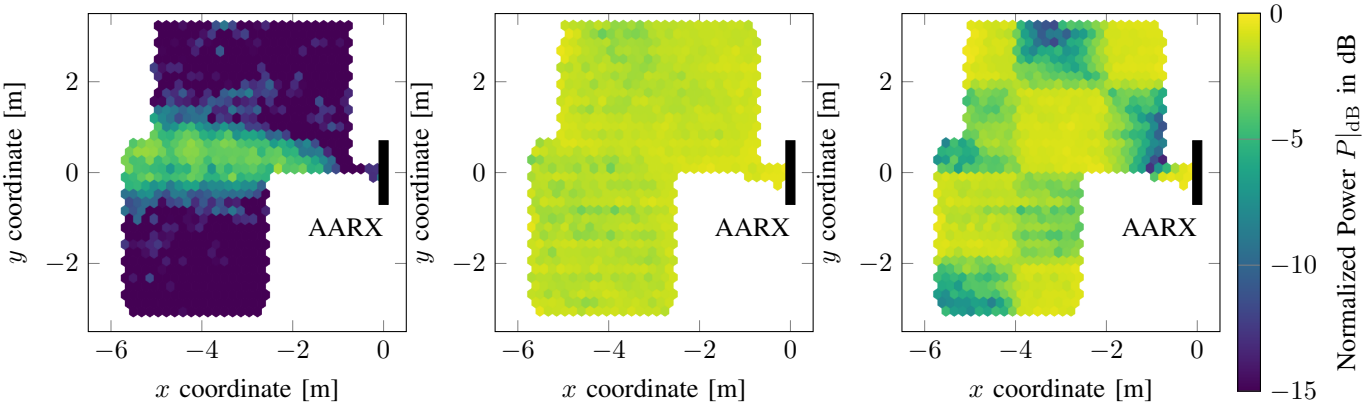
IV. Deep Learning-Based CSI Estimation

We evaluate five different deep learning-based downlink CSI estimators $\hat{\boldsymbol{\theta}}$ which produce an estimate $\mathbf{w}$ from $\mathbf{H}_\mathrm{U}$:

- **A DNN**: This simple architecture consists of four dense layers as shown in Fig. 3b.
- **A DNN with dropout**: Same as the DNN architecture, except for a dropout layer with dropout rate δ inserted between dense layers 2 and 3, to improve generalization.
- **An Encoder / Decoder structure** with arbitrary latent space: both encoder and decoder consist of three dense hidden layers each. The encoder reduces $\mathbf{H}_\mathrm{UL}$ to a latent space representation $\tilde{\mathbf{x}} \in \mathbb{R}^l$, that the decoder infers $\mathbf{h}_\mathrm{D}$ from, see Fig. 3a. This choice of network architecture is justified in the fact that, as explained in Section II, both $\mathbf{H}_\mathrm{U}$ and $\mathbf{h}_\mathrm{D}$ are entirely predetermined by a possibly sparser latent representation $\mathbf{x}$. The Encoder may be able to approximate f_U^{-1} whereas the decoder may approximate f_D.
- **An azimuth angle-based Encoder / Decoder structure**, i.e., the latent variable is forced to be an azimuth angle: Same as the previous architecture, except that encoder and decoder are now first trained separately: The encoder is trained to estimate the azimuth component α of the angle of arrival (AoA) from $\mathbf{H}_\mathrm{U}$ and the decoder is trained to generate $\mathbf{h}_\mathrm{D}$ from α, both supervised using position labels. The two NNs are then connected in series.
- **An azimuth and elevation angle-based Encoder / Decoder structure**, i.e., the two latent variables are forced to be elevation / azimuth angle estimates: Same as the previous architecture, except that the encoder now consists of two separate DNNs, for estimating both azimuth component α and elevation component β of the AoA. Again, the decoder is not trained on estimates, but on AoAs computed from position labels.

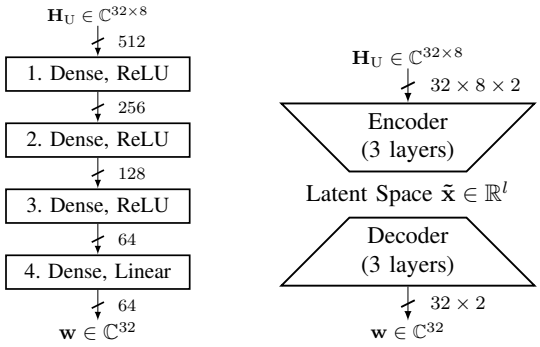
As a training loss function, we employ $\ell = 1 - P$, where P is the squared cosine similarity between estimated channel $\mathbf{w}$ and true channel $\mathbf{h}_\mathrm{D}$ as defined in Eq. (2). Instead of working with complex-valued channel coefficients, all NNs process channel coefficients in real / imaginary part representation. For Fig. 2b, we randomly assigned $50\,\%$ of all datapoints to the training set, trained the previously described DNN (without dropout) on this set and evaluated the DL CSI estimates on the complete dataset. The mean normalized received power over the complete dataset was found to be $\bar{P}|_\mathrm{dB} \approx -1.3\,\mathrm{dB}$.

For Fig. 2c, on the other hand, we partitioned the dataset into training set and test set in a checkerboard pattern with square side length $2\,\mathrm{m}$: All datapoints that were measured on



(a) Principal Component Baseline (b) DNN, trained on whole area (c) DNN, checkered training set

Fig. 2: Top view of normalized received powers P at different locations visualized over the approximately $6\,\mathrm{m} \times 6\,\mathrm{m}$ large measurement area in the dataset. The black box marked "AARX" indicates the location of the antenna array.



(a) DNN (b) Encoder / Decoder

Fig. 3: Architecture of different evaluated NNs

"white" checkerboard squares were assigned to the training set, all datapoints measured on "black" checkerboard squares to the test set. After evaluating the trained NN on both training and test set, it is easy to see that the performance on the two sets differs significantly, with $\bar{P}|_{\mathrm{dB}} \approx -0.9\,\mathrm{dB}$ on the training set and $\bar{P}|_{\mathrm{dB}} \approx -4.2\,\mathrm{dB}$ on the test set. Clearly, this indicates that the DNN is overfitting on the training set: It is not able to produce channel estimates $\mathbf{w}$ with comparable normalized DL power on unseen regions of the physical space.

This poses the question as to how this overfitting can be mitigated, either through standard methods such as adding dropout layers or by forcing the DNN to learn a sparser latent space representation. To formalize our enquiry into this topic into a quantifiable manner, we introduce a framework for evaluating the ability to generalize of different NN architectures.

V. A Framework for Evaluating Generalization

A. Defining Generalization

In the context of DL CSI estimation, we refer to generalization not as the ability to generalize from training set to test set in the same physical area (which a DNN can do well on our dataset, as is apparent from Fig. 2b), but as the ability to generalize from areas of the physical environment seen during training to areas that were not represented in the training set (which, considering the result in Fig. 2c, is much harder). Therefore, when talking about the quality of CSI estimates, it is insufficient to just measure a single performance metric: Some NN architectures perform well in seen areas, but worse in unseen locations whereas other architectures generalize better, but at the cost of a worse performance in seen areas.

To quantify this observation, as previously, the dataset is split into training and test set in a checkerboard pattern, with square side length a. If a is chosen to be small, the training set will contain a CSI datapoint in physical proximity of each (unseen) test set location. For large values of a, the NN needs to be able to generalize across larger distances. We define $\bar{P}_{\mathrm{seen}}$ to be the average received power (see Eq. (3)) when evaluating the trained NN on the training set, and $\bar{P}_{\mathrm{unseen}}$ to be the average received power after evaluation on the test set. $\bar{P}_{\mathrm{seen}}$ can be interpreted as the average loss in received power due to the suboptimal channel coefficient estimates. We expect $\bar{P}_{\mathrm{seen}} \geq \bar{P}_{\mathrm{unseen}}$, so $\bar{P}_{\mathrm{unseen}}|_{\mathrm{dB}} - \bar{P}_{\mathrm{seen}}|_{\mathrm{dB}} < 0\,\mathrm{dB}$ can be interpreted as the loss in average received power incurred in unseen areas due to lack of training data in physical proximity.

B. Seen/Unseen Loss Diagram and Baselines

To visualize NN performance, we propose a *seen/unseen loss* diagram as in Fig. 4, with losses $\bar{P}_{\mathrm{seen}}|_{\mathrm{dB}}$ on the horizontal axis and $\bar{P}_{\mathrm{unseen}}|_{\mathrm{dB}} - \bar{P}_{\mathrm{seen}}|_{\mathrm{dB}}$ on the vertical axis. In any case, the random precoding strategy from Thm. 1 provides a lower bound on the achievable performance (blue line and region). The best performance is achieved if perfect DL CSI is available at the receiver at all time, so that $\bar{P}_{\mathrm{seen}}|_{\mathrm{dB}} = \bar{P}_{\mathrm{unseen}}|_{\mathrm{dB}} = 0\,\mathrm{dB}$; this operating point is marked with "TDD", since, assuming perfect channel reciprocity, it is achievable by a TDD system. For all other estimators, the performance in seen and unseen areas depends on the partitioning of the dataset into training and test set. For Fig.

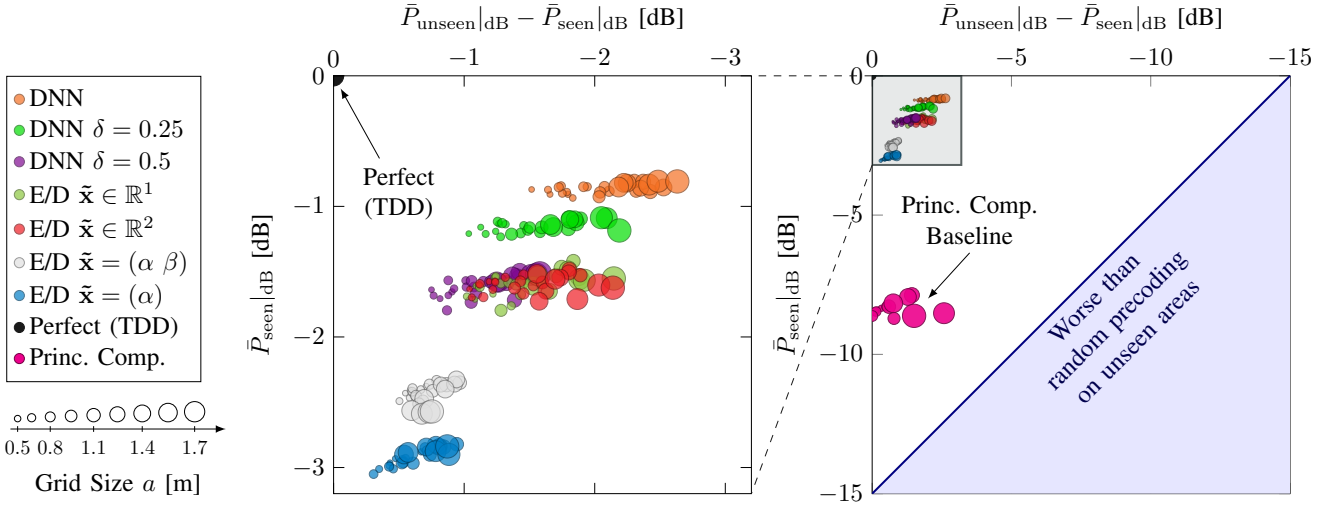


Fig. 4: *Seen/unseen loss* diagram with mean received power losses of different NN architectures and baselines on seen / unseen checkerboard fields, for grid sizes from $0.5\,\text{m}$ to $1.8\,\text{m}$. A larger grid size is indicated by a larger marker size.

4, this partitioning was performed in the afforementioned checkerboard pattern. The grid size parameter a was swept from $0.5\,\text{m}$ to $1.8\,\text{m}$ with a step size of $0.1\,\text{m}$. As an additional baseline, based on Thm. 2, we compute $\mathbf{w}_{\max}$ based on the training set and evaluate this vector for both training set ($\bar{P}_{\text{seen}}$) and test set ($\bar{P}_{\text{unseen}}$), yielding the principal component baseline (marked "Princ. Comp.") also illustrated in Fig. 4.

C. Discussion of Results

Among all tested NNs, the DNN without dropout performs best on previously seen data (i.e., with respect to $\bar{P}_{\text{seen}|\text{dB}}$). Increasing the dropout rate δ to $\delta = 0.25$ or $\delta = 0.5$ leads to a deteriorated performance with respect to $\bar{P}_{\text{seen}}$, but better generalization. Surprisingly, both encoder / decoder structures without predetermined latent space perform approximately equally well, regardless of the latent space dimensionality ($\mathbb{R}^1$ or $\mathbb{R}^2$), which may indicate that a sparse representation of CSI is indeed possible. A closer look at the learned latent representation would reveal that $\tilde{\mathbf{x}} \in \mathbb{R}^1$ is highly correlated with the azimuth angle. Despite this observation, encoder / decoder structures with predetermined azimuth α / elevation β latent spaces perform worse than all other NN architectures on previously seen physical areas, but generalize better.

We find that the performance of all evaluated NN architectures is significantly better than both *random precoding* and *principal component* baselines. In fact, $\bar{P}_{\text{seen}|\text{dB}} > -3.1\,\text{dB}$ and $\bar{P}_{\text{unseen}|\text{dB}} > -3.8\,\text{dB}$ for all NNs, which demonstrates that a NN-based approach is feasible and that some level of generalization to previously unseen physical areas is possible. However, Fig. 4 also clearly shows significant performance differences between the various NN architectures and the strong influence of the grid size on generalization.

VI. Summary and Outlook

We found that NN-based downlink channel estimation from available uplink CSI significantly outperformed the baselines and that generalization to physical areas not represented in the training set is one of the major challenges of the approach. With regards to generalization, we evaluated several different network architectures on measurement data. Thanks to the public data, our research may be reproduced on the same dataset or compared to other datasets captured in different types of environments or with different antenna configurations. The effect of the frequency separation between uplink and downlink channel may also be studied further.

References

[1] Y. Yang, F. Gao, G. Y. Li, and M. Jian, "Deep learning-based downlink channel prediction for FDD massive MIMO system," *IEEE Communications Letters*, vol. 23, no. 11, pp. 1994–1998, 2019.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE transactions on wireless communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[3] K. Hugl, K. Kalliola, J. Laurila *et al.*, "Spatial reciprocity of uplink and downlink radio channels in FDD systems," in *Proc. COST*, vol. 273, no. 2. Citeseer, 2002, p. 066.

[4] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency," in *2019 53rd asilomar conference on signals, systems, and computers*. IEEE, 2019, pp. 1465–1470.

[5] Y. Zhang, J. Wang, J. Sun, B. Adebisi, H. Gacanin, G. Gui, and F. Adachi, "CV-3DCNN: Complex-valued deep learning for CSI prediction in FDD massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 266–270, 2020.

[6] M. Arnold, S. Dörner, S. Cammerer, S. Yan, J. Hoydis, and S. ten Brink, "Enabling FDD massive MIMO through deep learning-based channel prediction," *arXiv preprint arXiv:1901.03664*, 2019.

[7] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–6.

[8] F. Euchner, M. Gauger, S. Dörner, and S. ten Brink, "A Distributed Massive MIMO Channel Sounder for "Big CSI Data"-driven Machine Learning," in *WSA 2021; 25th International ITG Workshop on Smart Antennas*, 2021.

[9] F. Euchner and M. Gauger, "CSI Dataset dichasus-015x: Indoor Line of Sight, Lab Room," 2021. [Online]. Available: https://doi.org/doi:10.18419/darus-2202