



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY



# Channel state information prediction with limited UE feedback in 5G NR

Analyzing existing and enhanced codebooks

Master's thesis in Communication Engineering

JOAKIM EDBY & THIAGO PUPPIN ROMANO

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2023

[www.chalmers.se](http://www.chalmers.se)



MASTER'S THESIS 2023

# Channel state information prediction with limited UE feedback in 5G NR

Analyzing existing and enhanced codebooks

JOAKIM EDBY & THIAGO PUPPIN ROMANO



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering  
*Division of Communication, Antennas, and Optical Networks*  
Communication Systems Group  
CHALMERS UNIVERSITY OF TECHNOLOGY  
Gothenburg, Sweden 2023

Channel state information prediction  
with limited UE feedback in 5G NR  
Analyzing existing and enhanced codebooks  
JOAKIM EDBY & THIAGO PUPPIN ROMANO

© JOAKIM EDBY & THIAGO PUPPIN ROMANO, 2023.

Supervisors: Xinlin Zhang, Ericsson; Hao Guo, Chalmers; Mehdi Sattari, Chalmers  
Examiner: Tommy Svensson, Electrical Engineering

Master's Thesis 2023  
Department of Electrical Engineering  
Division of Communication, Antennas, and Optical Networks  
Communication Systems Group  
Chalmers University of Technology  
SE-412 96 Gothenburg  
Telephone +46 31 772 1000

Cover: Antennas on a rooftop in Gothenburg, Sweden. Photo taken by the authors.

Typeset in L<sup>A</sup>T<sub>E</sub>X  
Printed by Chalmers Reproservice  
Gothenburg, Sweden 2023

Channel state information prediction with limited UE feedback in 5G NR  
Analyzing existing and enhanced codebooks  
JOAKIM EDBY & THIAGO PUPPIN ROMANO  
Department of Electrical Engineering  
Chalmers University of Technology

## Abstract

Accurate channel state information (CSI) is a key component for efficient and reliable communications in wireless networks. Utilizing CSI enables the system to adapt to various channel conditions and increases the total throughput. In 5G New Radio (NR), codebooks determine how the user equipment (UE) reports CSI to the gNodeB (gNB). More specifically, a precoder matrix indicator (PMI) is included in the CSI report, suggesting the recommended precoder matrix that the gNB may use for downlink transmission. Since the channel is time-varying, the CSI report may be outdated when received at the gNB, which degrades system performance. A potential solution is to predict the CSI at the gNB. Thus, this study investigates CSI prediction based on the current codebooks in 5G NR. Moreover, a new reporting format is proposed where the throughput is increased with only a minor increase in overhead. An autoregressive (AR) model combined with a Kalman filter is used for prediction. The results indicate that prediction based on existing codebooks is possible, but by slightly increasing the overhead using the new reporting format, the channel can be recreated at the gNB and more accurately predicted. Simulations based on a standardized channel model show that implementing a Kalman filter on existing codebooks provide a gain of around 1.9 dB. By using the new reporting method, a gain of approximately 2.2 dB can be obtained.

Keywords: 5G, channel prediction, gNB, CSI, codebook, PMI.



# Acknowledgements

We would like to express our gratitude to the following individuals who have helped and supported us with our thesis:

First and foremost, we would like to thank our dedicated supervisors at Ericsson: Xinlin Zhang, Johan Wingses, and Keerthi Kumar Nagalapur. Your expertise and guidance have been instrumental and greatly enriched the quality of the thesis.

We would like to extend our gratitude to our supervisors at Chalmers: Hao Guo and Mehdi Sattari, for providing insightful feedback and constructive criticism on the report and helping to shape the thesis into its final form.

Finally, we would like to thank our hiring manager Anders Aronsson at Ericsson, for his support and belief in our potential, and our examiner Tommy Svensson at Chalmers, for suggesting the thesis and guiding us throughout the research process.

Joakim Edby & Thiago Puppini Romano, Gothenburg, June 2023



# List of Acronyms

Below is a list of acronyms that have been used throughout this thesis, listed in alphabetical order:

3GPP	Third Generation Partnership Project
5G	Fifth generation
AR	Autoregressive
CDL	Clustered delay line
CQI	Channel quality indicator
CSI	Channel state information
CSIR	CSI at the receiver
CSIT	CSI at the transmitter
DFT	Discrete Fourier transform
eType II	Enhanced Type II
FDD	Frequency division duplex
feType II	Further Enhanced Type II
gNB	gNodeB
ISI	Intersymbol interference
LC	Linear combining
LI	Layer indicator
LMMSE	Linear minimum mean squared error
LOS	Line of sight
MCS	Modulation and coding scheme
MIMO	Multiple input multiple output
MISO	Multiple input single output
MP	Multi-panel
MRT	Maximum ratio transmission
MSE	Mean squared error
MU-MIMO	Multi-user MIMO
NLOS	Non line of sight
NR	New Radio
OFDM	Orthogonal frequency-division multiplexing
PMI	Precoding matrix indicator
PRB	Physical resource block

---

PS	Port-selection
RI	Rank indicator
RS	Reference signal
SIMO	Single input multiple output
SINR	Signal-to-interference-plus-noise ratio
SNR	Signal-to-noise ratio
SP	Single-panel
SRS	Sounding reference signals
SU-MIMO	Single-user MIMO
SVD	Singular value decomposition
TDD	Time division duplex
UE	User equipment
ULA	Uniform linear array
WSS	Wide sense stationary





# Contents

<b>List of Acronyms</b>	<b>ix</b>
<b>List of Figures</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Aim and scope . . . . .	2
1.2 Contributions . . . . .	2
1.3 Structure of the report . . . . .	3
<b>2 Theory</b>	<b>5</b>
2.1 Basics of the wireless channel . . . . .	5
2.2 Beamforming . . . . .	6
2.3 Spatial multiplexing in MIMO . . . . .	8
2.4 CSI reports . . . . .	10
2.5 Codebook structure . . . . .	10
2.5.1 Release 15 codebooks . . . . .	12
2.5.1.1 Example of a PMI calculation . . . . .	13
2.5.2 Release 16 codebooks . . . . .	14
2.6 AR modeling . . . . .	16
2.7 The Kalman Filter . . . . .	16
<b>3 Methods</b>	<b>19</b>
3.1 Channel model . . . . .	19
3.2 Codebook design . . . . .	19
3.3 Kalman filter and AR model implementation . . . . .	20
3.4 Performance evaluation of precoders . . . . .	20
<b>4 Results</b>	<b>21</b>
4.1 Simulation setup . . . . .	21
4.2 Prediction-based on the eTypeII codebook . . . . .	22
4.3 Prediction based on augmented reporting scheme . . . . .	24
<b>5 Discussion</b>	<b>31</b>
5.1 Key findings . . . . .	31
5.2 Simulation parameters . . . . .	31
5.3 Outlook . . . . .	32

<b>6 Conclusion</b>	<b>33</b>
6.1 Future work . . . . .	33
<b>Bibliography</b>	<b>35</b>

# List of Figures

2.1	A narrowband LOS MISO system. Under the assumption that $d_1, \dots, d_{N_{\text{tx}}} \gg \Delta_t \lambda_c$ , the outgoing signals can be approximated as parallel. . . . .	7
2.2	Visualization of the transmission and beamforming patterns. The red dots correspond to the antennas. The antennas are not to scale. . . . .	8
2.3	An illustration of how independent channels can be created by the singular value decomposition method. . . . .	9
2.4	A timeline of the codebooks from Release 15 to 18. . . . .	11
2.5	Illustration of a single-panel and multi-panel UPA, also showing the cross-polarized antennas and how $N_1$ and $N_2$ are defined. . . . .	12
2.6	Visualization of the possible beams to select from when calculating $\mathbf{W}_1$ . The figure shows an example where $N_1 = 4, N_2 = 2, O_1 = 4$ , and $O_2 = 4$ . . . . .	13
2.7	(a) Energy of $\mathbf{W}_2$ in the beam-frequency and (b) beam-delay domain. The beam-delay domain is much sparser and allows for compression. . . . .	15
4.1	Throughput versus SNR for prediction methods based on the eType II codebook. The KF eType II with preprocessing and the KF eType II lines correspond to methods with prediction, while eType II is the baseline model where no prediction is performed. . . . .	23
4.2	Phase over time for a beam-delay pair, with and without the preprocessing step. Applying the preprocessing step decreases the phase jumps, making the phase more continuous and increasing the predictability. . . . .	24
4.3	Throughput over time for prediction methods based on the eType II codebook. . . . .	25
4.4	Block Diagram of the augmented method. . . . .	25
4.5	Phase normalization of $\mathbf{U}$ and $\mathbf{V}$ with the objective of making the second receiver port coefficients with phase zero. Here, only the first layer is presented, other layers must be normalized independently. In these plots, the color represent the amplitude of the coefficient while the red arrow represent the phase. . . . .	26
4.6	Applying a DFT matrix on $\mathbf{U}$ shows that the energy of is concentrated in the first delay, allowing for compression. . . . .	27
4.7	The singular values in $\mathbf{\Sigma}$ are almost the same for all subbands, allowing them to be reported in a wideband manner. . . . .	28

4.8	Comparison between the regular KF eType II with preprocessing against the augmented method, with and without compression. . . . .	28
4.9	Throughput as a function of SNR. Note that the gain of the augmented method does not decrease significantly by compressing $\mathbf{U}$ and $\mathbf{\Sigma}$ . . . . .	29
4.10	Throughput as a function of bits per PMI report. . . . .	29

# 1

## Introduction

Mobile networks are an integral part of society. Providing mobile telephony and broadband, they enable fast and reliable communication and connect users globally. As global mobile network traffic is expected to grow, the fifth generation (5G) of mobile communications will play a crucial role in handling the data growth. In 2022, the mobile network traffic was approximately 90 exabyte (EB) per month and the 5G share was 17 %. By 2028, mobile network traffic is predicted to exceed 300 EB per month, which is more than three times the number in 2022, and 5G is expected to have a 70 % share of that traffic [1]. In addition to handling more data, 5G envisions use cases with high network requirements, such as enhanced mobile broadband (eMBB), which enables even higher data rates and larger volumes [2].

As the data demand grows and the number of use cases increase, there is an increased need for reliable and efficient communications. A key component is accurate channel state information (CSI) between the user equipment (UE) and the gNodeB (gNB)<sup>1</sup>. By utilizing CSI at the transmitter (CSIT), beamforming can be used to send narrow beams with high gain to the UE, increasing the signal-to-noise ratio (SNR) and reducing interference to other UEs. Beamforming is accomplished by a precoder that changes the phase and amplitude of the signal generated by the transmit antennas in such a way that the signals experience constructive interference at the receiver.

In frequency division duplex (FDD) systems, the gNB obtains CSI by sending a reference signal (RS) to the UE, which then estimates the channel and sends back a CSI report to the gNB. A CSI report in current 5G standards typically includes a channel quality indicator (CQI), a rank indicator (RI), and a precoding matrix indicator (PMI) [2,3]. The PMI is the main focus of this thesis. It contains information on how the gNB should select a precoder to best serve the UE. The contents of the PMI and how it is calculated at the UE is determined by a codebook.

Since the wireless channel is time-varying, CSI can suffer from an aging problem, especially when the UE is moving [4, 5]. That is, the CSI obtained for one time instance may be outdated when the CSI report is received at the gNB at a later time instance. Applying an outdated CSI report not only degrades the performance of served UEs, but also cause interference to other UEs. A remedy for getting an up-to-date CSI is to send more frequent downlink RS and uplink CSI reports over time, however, this incurs large overhead for both uplink and downlink.

---

<sup>1</sup>The term gNodeB (gNB) is used to denote the base station in 5G.

Instead of sending CSI reports more frequently, a solution to the aging problem could be to predict the CSI. CSI prediction at the UE is an ongoing work item in Release 18, the latest Third Generation Partnership Project (3GPP) 5G standardization [6–8]. However, gNB-based prediction could be an interesting alternative with some advantages. Firstly, a gNB can handle a higher computational complexity, such that more sophisticated prediction algorithms can be used [9]. Secondly, different UE vendors may have different implementations with varying quality. Since the gNB controls scheduling and resource allocation, maintaining a uniform prediction quality at the gNB might be beneficial in terms of overall system performance. Thirdly, a potential limitation of UE-based prediction is that the CSI reports would be transmitted for specific time slots. Downlink transmissions between these time slots would then have reduced quality. The gNB, however, could potentially predict for the specific time where downlink transmission would occur.

The topic of channel prediction is a well-studied area with an extensive list of literature. For example, [9] and [10] study channel prediction using the Kalman filter and autoregressive modeling, [11] and [12] use convolutional neural networks, and [5] use predictor antennas placed in front of the receiver antennas on moving vehicles. While the literature typically performs prediction based on the full channel information, the CSI reports need to be limited in terms of overhead. Therefore, this thesis focuses on prediction where the available CSI is limited by the data reported in the current 5G codebooks.

### 1.1 Aim and scope

The aim of this thesis is to study the feasibility of CSI prediction at the gNB, constrained by information reported by the existing codebooks. Furthermore, the aim is to propose an augmented CSI reporting method that increases the prediction quality while limiting overhead.

As for scope, the thesis focuses on the Enhanced Type II (eType II) codebook from Release 16 [13] as a baseline for the studies. Note that our results can be extended to other codebooks. Moreover, while there are many different models and methods that can be used for prediction, this thesis uses an autoregressive (AR) model, which is a proven solution to the channel prediction problem. More specifically, a Kalman filter will be used and the AR parameters will be estimated using the Yule-Walker equations. Thus, investigating alternative prediction methods is outside the scope of this thesis.

### 1.2 Contributions

The report has two main contributions. Firstly, CSI prediction using the Release 16 codebook is investigated. It is shown that with a processing step at the gNB, where one beam is used as a reference for the phase over all time slots, prediction using the current codebook is possible and provides a throughput gain. Secondly, an augmented method for reporting the PMI is proposed. The main idea behind

the new method is that the PMI includes not only the gNB-side singular vectors but also the UE-side singular vectors and the singular values. This information allows the gNB to reconstruct the channel and perform prediction on it. The additional information is also compressed to limit the reporting overhead.

### 1.3 Structure of the report

The rest of the report is structured as follows.

Section 2 provides the theoretical background for the report. It includes the basics behind the wireless channels, a mathematical description of beamforming and precoding, the principles behind CSI acquisition, and the structure of Release 15 and Release 16 codebooks. The theory behind AR models and the Kalman filter is also presented.

Section 3 includes a description of the channel model used in the simulations. The implementation of the codebook, AR model, and the Kalman filter are also described. Additionally, it includes the measurement used to compare different precoders based on different CSI reporting formats.

Section 4 includes the results for predicting the CSI report based on the current Release 16 codebook and the augmented CSI reporting method. The details behind the augmented method are motivated and explained.

Section 5 discusses the obtained results with respect to their significance, validity, and implications. It also includes suggestions for improvements and future work.

Finally, Section 6 concludes the report and highlights the key findings.



# 2

## Theory

This section presents the theoretical framework for the report. Sections 2.1-2.4 include the basics of the wireless channel, beamforming, precoding and CSI reporting. This information is useful for section 2.5, which explains the codebooks in 5G. Section 2.6 and 2.7 present the theory behind AR modeling and the Kalman filter, which are used for prediction.

### 2.1 Basics of the wireless channel

Wireless communication use signals that propagate through the wireless channel. The signals reflect, scatter, and diffract due to objects in the environment. Thus, a receiver can receive multiple copies of the transmitted signal as it may take different paths and arrive with varying signal strengths and delays [14].

The time between the first received signal and the last copy is known as the delay spread. If the delay spread exceeds the symbol time, which is the time during which a signal is transmitted, subsequent signals may overlap, causing intersymbol interference (ISI). The inverse of the delay spread is known as the coherence bandwidth and is the bandwidth for which a channel remains constant. If the signal's bandwidth exceeds the coherence bandwidth, which is also known as wideband communication, the signal experiences frequency selective fading. In frequency selective fading, the channel gain changes with respect to frequency over the communication bandwidth [14].

To combat ISI and frequency selective fading, multicarrier modulation formats can be used. Specifically, 5G New Radio (NR) uses orthogonal frequency-division multiplexing (OFDM) [2]. The main advantage of OFDM is that it divides a broadband signal into multiple narrowband signals, called subcarriers. Each subcarrier has a lower data rate, but experiences something closer to flat fading, thus reducing the effects of frequency selective fading. An OFDM symbol consists of multiple subcarriers. It is transmitted during a certain time and occupies a certain frequency bandwidth. Furthermore, a resource element is the smallest resource unit that carries one data symbol and occupies a single subcarrier in an OFDM symbol. In 5G NR, a physical resource block (PRB) comprises 12 consecutive resource elements in the frequency domain. In turn, a subband consists of one or multiple contiguous PRBs. The minimum resource allocation unit to any single user is a PRB over 14 OFDM symbols in the normal cyclic prefix regime.

## 2.2 Beamforming

In 5G, gNBs are equipped with many antennas, which allows for beamforming. Beamforming is used to increase the SNR and reduce interference. A mathematical description and explanation of beamforming is provided below.

Imagine a narrowband line of sight (LOS) multiple input single output (MISO) system with multiple antennas at the transmitter and one antenna at the receiver. The transmit antennas are aligned in a uniform linear array (ULA) and simultaneously transmit the same signal. The propagating signals will interfere with each other and create an interference pattern through the principle of superposition. If the phase of the outgoing signals are adjusted such that they interfere constructively at the receiver, it will result in a power gain. Adjusting the phase of the outgoing signals at the transmitter side is known as transmit beamforming.

A key question is how the transmitter should adjust the phases. Mathematically, the situation can be described as follows. Assume that there is one antenna at the receiver and  $N_{\text{tx}}$  antennas at the transmitter. Ignoring noise, the received signal is given by

$$y = \mathbf{h}\mathbf{w}_t x, \quad (2.1)$$

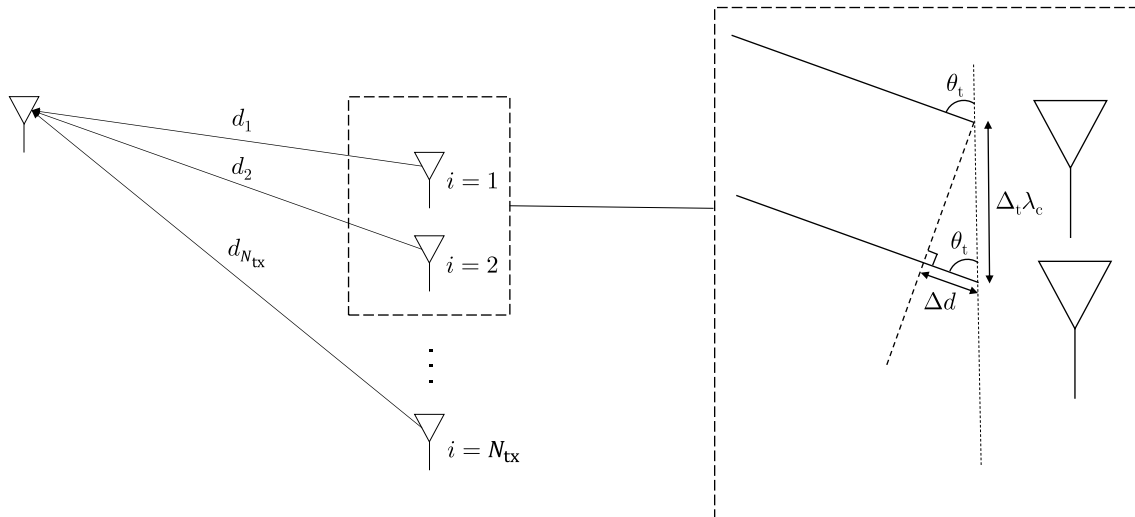
where  $y$  is the received signal,  $\mathbf{h}$  a  $1 \times N_{\text{tx}}$  vector describing the channel,  $\mathbf{w}_t$  a  $N_{\text{tx}} \times 1$  vector containing the phases for each antenna, and  $x$  the transmitted symbol. The distance between each transmit antenna is  $\Delta_t \lambda_c$ , where  $\lambda_c$  is the carrier wavelength and  $\Delta_t$  the distance between the transmit antennas, normalized to the carrier wavelength. Each signal travels a path with length  $d_i$ , where  $i = 1, 2, \dots, N_{\text{tx}}$  is the path from antenna  $i$ . Assuming the transmitted signals are parallel to each other, which is a good approximation since the path distance is typically much larger than the antenna array length, the system can be modeled as illustrated in Figure 2.1. The additional distance,  $\Delta d$ , for path  $i$  is then [15]

$$(i - 1) \Delta d = (i - 1) \Delta_t \lambda_c \cos \theta_t, \quad (2.2)$$

where  $\theta_t$  is the angle of departure.

The shortest path between the transmitter antennas and the receiver antenna has distance  $d$  and the attenuation  $a$  is the same for all rays. The channel can be then be modelled as [15]

$$\mathbf{h} = a \exp\left(-\frac{j2\pi d}{\lambda_c}\right) \begin{bmatrix} 1 \\ \exp(-j2\pi \Delta_t \Omega_t) \\ \exp(-j2\pi 2\Delta_t \Omega_t) \\ \vdots \\ \exp(-j2\pi (N_{\text{tx}} - 1) \Delta_t \Omega_t) \end{bmatrix}^T, \quad (2.3)$$



**Figure 2.1:** A narrowband LOS MISO system. Under the assumption that  $d_1, \dots, d_{N_{\text{tx}}} \gg \Delta_t \lambda_c$ , the outgoing signals can be approximated as parallel.

where  $\Omega_t = \cos \theta_t$  is known as the directional cosine and  $(\cdot)^T$  denotes transpose. To compensate for the phase differences, the transmitter can multiply the input symbol  $x$  with the vector

$$\mathbf{w}_t = \frac{1}{N_{\text{tx}}} \begin{bmatrix} 1 \\ \exp(j2\pi \Delta_t \Omega_t) \\ \exp(j2\pi 2\Delta_t \Omega_t) \\ \vdots \\ \exp(j2\pi (N_{\text{tx}} - 1) \Delta_t \Omega_t) \end{bmatrix}, \quad (2.4)$$

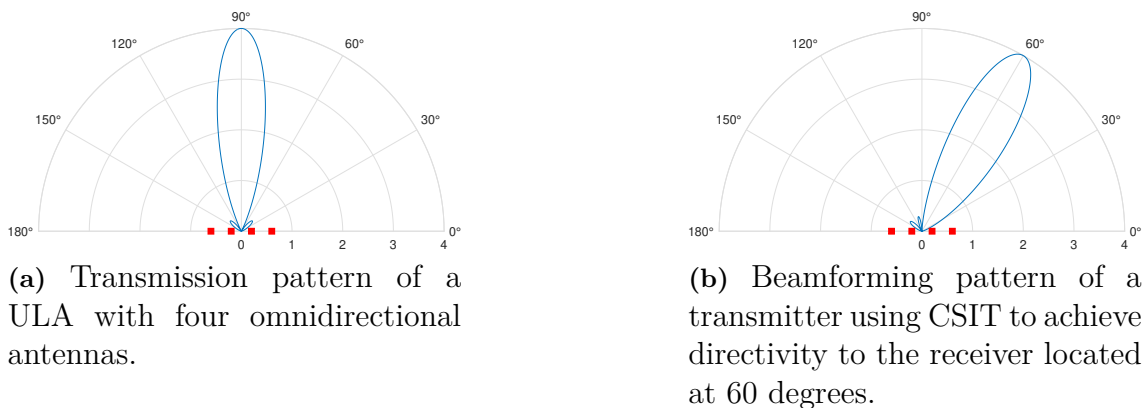
which effectively undoes the spatial effects of the channel. The factor  $\frac{1}{N_{\text{tx}}}$  ensures that the power is divided equally between all antennas at the transmitter. The resulting received scalar is

$$y = \mathbf{h} \mathbf{w}_t x = a \exp\left(-\frac{j2\pi d}{\lambda_c}\right) \sqrt{N_{\text{tx}}} x, \quad (2.5)$$

which has an  $N_{\text{tx}}$  power gain compared to the case when only one antenna at the transmitter is used.

The vector  $\mathbf{w}_t$  derived here is also known as a maximum ratio transmission (MRT) precoder [15]. Constructing the MRT precoder requires full information about the channel. The retrieval of channel information is presented in further detail in section 2.4 and 2.5.

The transmission pattern of a ULA with  $N_{\text{tx}} = 4$  and  $\Delta_t = 0.5$  that does not use a precoder and transmits the same signal on each antenna can be seen in Figure 2.2(a). It shows the gain in different directions, with  $\theta_t$  ranging from 0 to  $\pi$ . If we assume that the receiver antenna is at an angle of 60 degrees from the transmitter,



**Figure 2.2:** Visualization of the transmission and beamforming patterns. The red dots correspond to the antennas. The antennas are not to scale.

the transmitter can adjust the phases for each outgoing signal using  $\mathbf{w}_t$ , which can be seen as steering the beam to that direction, as shown in Figure 2.2(b).

Note that although this section focuses on a MISO system, the principles are also true for a single input multiple output (SIMO) system in which there is only one antenna at the transmitter and multiple antennas at the receiver. In such a system, receive beamforming can be achieved by adjusting the phase of each incoming signal, which would require CSI at the receiver (CSIR).

To summarize, using CSIT allows for beamforming which can be used to steer a beam toward the intended user and provides a power gain. The next section will introduce precoding, which can be seen as a generalization of beamforming.

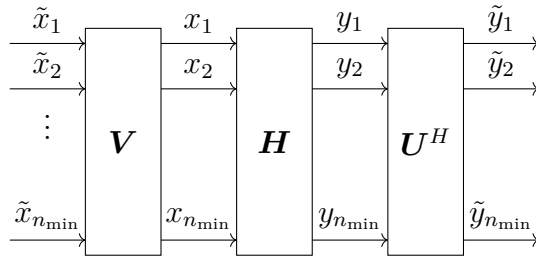
### 2.3 Spatial multiplexing in MIMO

Not only the gNBs are equipped with multiple antennas, many UEs also have multiple antennas. In a multiple input multiple output (MIMO) system, the data rate can be increased by means of spatial multiplexing. That is, transmitting multiple data streams on the same time and frequency resource. To explain spatial multiplexing, the MISO system from section 2.2 will be extended to a MIMO system.

Assume a narrowband MIMO system with  $N_{\text{tx}}$  transmitter antennas and  $N_{\text{rx}}$  receiver antennas. The complex channel matrix  $\mathbf{H}$  has dimensions of  $N_{\text{rx}} \times N_{\text{tx}}$ , with each entry  $h_{i,j}$  corresponding to the channel gain between receiver antenna  $i$  and transmitter antenna  $j$ . Ignoring noise, the received signal can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N_{\text{rx}}} \end{bmatrix} = \begin{bmatrix} h_{1,1} & \cdots & h_{1,N_{\text{tx}}} \\ \vdots & \ddots & \vdots \\ h_{N_{\text{rx}},1} & \cdots & h_{N_{\text{rx}},N_{\text{tx}}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N_{\text{tx}}} \end{bmatrix}, \quad (2.6)$$

where  $y_i$  is the received signal at receiver antenna  $i$ , and  $x_j$  the transmitted signal



**Figure 2.3:** An illustration of how independent channels can be created by the singular value decomposition method.

from transmitter antenna  $j$ .

The channel matrix can be decomposed into three matrices by using the singular value decomposition (SVD)

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \quad (2.7)$$

where  $\mathbf{U}$  is an  $N_{\text{rx}} \times N_{\text{rx}}$  unitary matrix,  $\mathbf{\Sigma}$  an  $N_{\text{rx}} \times N_{\text{tx}}$  matrix with singular values  $\sigma_1, \sigma_2, \dots, \sigma_{n_{\text{min}}} > 0$  on the main diagonal, and  $\mathbf{V}$  is also a unitary matrix, but with dimensions  $N_{\text{tx}} \times N_{\text{tx}}$ . The notation  $(\cdot)^H$  denotes the conjugate transpose. The number of singular values is  $n_{\text{min}}$ , which is given by the rank of  $\mathbf{H}$ .

If  $\mathbf{V}$  is used as a transmit precoder and  $\mathbf{U}^H$  for receiver combining, parallel independent channels are obtained. A property of a unitary matrix  $\mathbf{X}$  is that  $\mathbf{X}\mathbf{X}^H = \mathbf{X}^H\mathbf{X} = \mathbf{I}$ , and following the notation in Figure 2.3, we have that [15]

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{U}^H \mathbf{y} \\ &= \mathbf{U}^H \mathbf{H} \mathbf{x} \\ &= \mathbf{U}^H \mathbf{H} \mathbf{V} \tilde{\mathbf{x}} \\ &= \mathbf{U}^H \mathbf{U} \mathbf{\Sigma} \mathbf{V}^H \mathbf{V} \tilde{\mathbf{x}} \\ &= \mathbf{\Sigma} \tilde{\mathbf{x}}, \end{aligned} \quad (2.8)$$

and since  $\mathbf{\Sigma}$  is diagonal,  $n_{\text{min}}$  parallel independent channels are created. Data can be transmitted on each independent channel, which is known as spatial multiplexing. Each data stream is also known as a layer, and the rate at which information can be transmitted is determined by the singular values in  $\mathbf{\Sigma}$ .

The rank of  $\mathbf{H}$  is determined by the physical properties of the channel. A pure LOS MIMO channel would be fully correlated and have a rank of 1, while the channel matrix for a rich scattering environment would have full rank, which is  $\min(N_{\text{tx}}, N_{\text{rx}})$ . Furthermore, the rank determines how many layers can be used for transmission and is limited by the number of antennas. For example, a gNB with 32 antennas and a UE with two antennas will be limited to rank 2 transmission due to the UE only having two antennas.

An important property of the SVD is that any column of  $\mathbf{U}$  and  $\mathbf{V}$  matrix can be rotated by an arbitrary phase factor  $e^{j\phi}$  [16] and still be a valid solution, given that the corresponding columns from both matrices are rotated by the same factor. For

real valued matrices this behavior is known as sign ambiguity, but it will be referred as phase ambiguity throughout this report as the matrices here are often complex valued.

### 2.4 CSI reports

As mentioned in the previous sections, designing a precoder requires information about the channel. There are two main ways of acquiring CSIT: reciprocity-based and feedback-based CSI acquisition [17]. The reciprocity-based CSI acquisition is useful in time division duplex (TDD) systems where the uplink and downlink share the same frequency band. The UE can send a sounding reference signals (SRS) to the BS and from that uplink signal, the gNB can estimate the required precoder for downlink transmission. One drawback of this method is that the signal strength from the UE is typically low, which lowers the SNR and leads to an inferior channel estimate. Another drawback is that the channel is not necessarily reciprocal if different number of antennas are used for uplink and downlink [18]. For example, to save power, a UE might use more antennas for receiving, but less antennas for transmitting. In such cases, only partial channel reciprocity holds.

In feedback-based CSI acquisition, the gNB sends a downlink reference signal to the UE. The UE estimates the channel and sends back some value corresponding to the parameters of the channel (implicit feedback), which in our case is the PMI values that correspond to a precoder. One advantage of the feedback-based CSI acquisition is that it can be used in both FDD and TDD systems, which can be useful if there is only partial channel reciprocity. However, one disadvantage is that it has larger overhead compared to the reciprocity-based CSI acquisition [17].

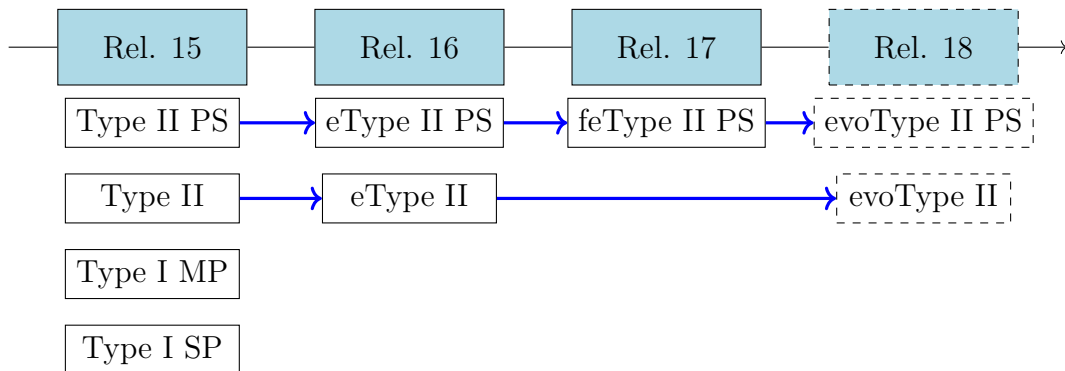
Listed below are some examples of the reporting quantities that a CSI report can contain [3]:

- A PMI, indicating the recommended precoder matrix.
- A CQI, indicating the UEs recommended modulation and coding scheme (MCS).
- An RI, indicating the recommended rank of transmission (same as number of layers).
- A layer indicator (LI), indicating the strongest layer when the rank indicator is greater than one.

Together, these make up a CSI report. Note that not all reporting quantities are listed here. A list and description of all reporting quantities can be found in [3].

### 2.5 Codebook structure

Codebooks determine how the PMI is calculated. The introduction of 5G NR began with Release 15 which supported four codebooks: Type I single-panel (SP), Type



**Figure 2.4:** A timeline of the codebooks from Release 15 to 18 [4, 8, 13, 19].

I multi-panel (MP), Type II, and Type II port-selection (PS) [4]. In Release 16, enhancements related to overhead for the Type II codebooks were made, which resulted in the eType II and eType II PS [13]. Additional enhancements were made in Release 17, resulting in the Further Enhanced Type II (feType II) PS [19]. As of Release 17, a total of seven codebooks are supported in 5G NR<sup>1</sup>. A timeline of the codebooks can be seen in Figure 2.4. Note that the Release 18 codebooks are not yet available, but may include UE-side prediction [8].

In general, the Type II codebooks are more sophisticated compared to the Type I codebooks as they provide richer information of the channel. As a result, the Type II codebooks target multi-user MIMO (MU-MIMO) use cases, while Type I codebooks target single-user MIMO (SU-MIMO) use cases [2]. The Type I codebooks support only one spatial beam per layer<sup>2</sup> and rely on the UE to suppress the inter-layer interference, which is made possible by keeping the number of layers less than the number of receivers at the UE. In MU-MIMO transmission, however, the UE will receive signals that are designated to other UEs from side lobes, causing UE interference. The interference can be canceled at the transmitter side if rich CSI is available, which is supported by the Type II codebooks [20]. However, richer CSI comes at the price of larger feedback overhead. The Type I codebooks typically require some tens of bits, while Type II codebooks require several hundreds of bits per CSI report [2, 16].

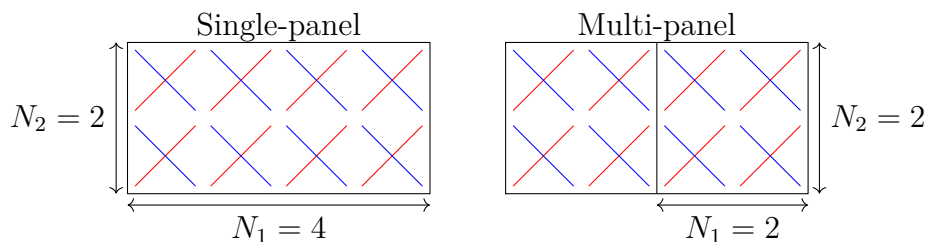
The terms SP and MP refer to the antenna array structure. gNBs are equipped with many antennas with orthogonal polarizations. The antennas can be structured in a single panel or multiple panels, as illustrated in Figure 2.5. The number of antennas in the horizontal and vertical direction is denoted  $N_1$  and  $N_2$ , respectively. The total number of ports<sup>3</sup> is  $P = 2N_1N_2$  for the SP and  $P = 2N_gN_1N_2$  for the MP, where  $N_g$  is the number of panels.

The main difference between PS codebooks and non PS codebooks is that in the former, the computational complexity takes place at the gNB instead of at the UE. Only non PS codebooks are studied in this thesis. Below, the codebooks from

<sup>1</sup>These seven codebooks are for downlink. There are other codebooks for uplink.

<sup>2</sup>In some codebook modes, there could be different beams for each polarization.

<sup>3</sup>These are typically referred to as CSI-RS ports in literature.



**Figure 2.5:** Illustration of a single-panel and multi-panel uniform planar array, also showing the cross-polarized antennas and how  $N_1$  and  $N_2$  are defined.

Release 15 and Release 16 are explained in detail.

### 2.5.1 Release 15 codebooks

Despite the seven codebooks being quite different, they share a common so-called dual-stage structure [4], which means the precoder matrix  $\mathbf{W}$  is given by the product of two matrices,  $\mathbf{W}_1$  and  $\mathbf{W}_2$ , as

$$\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2. \quad (2.9)$$

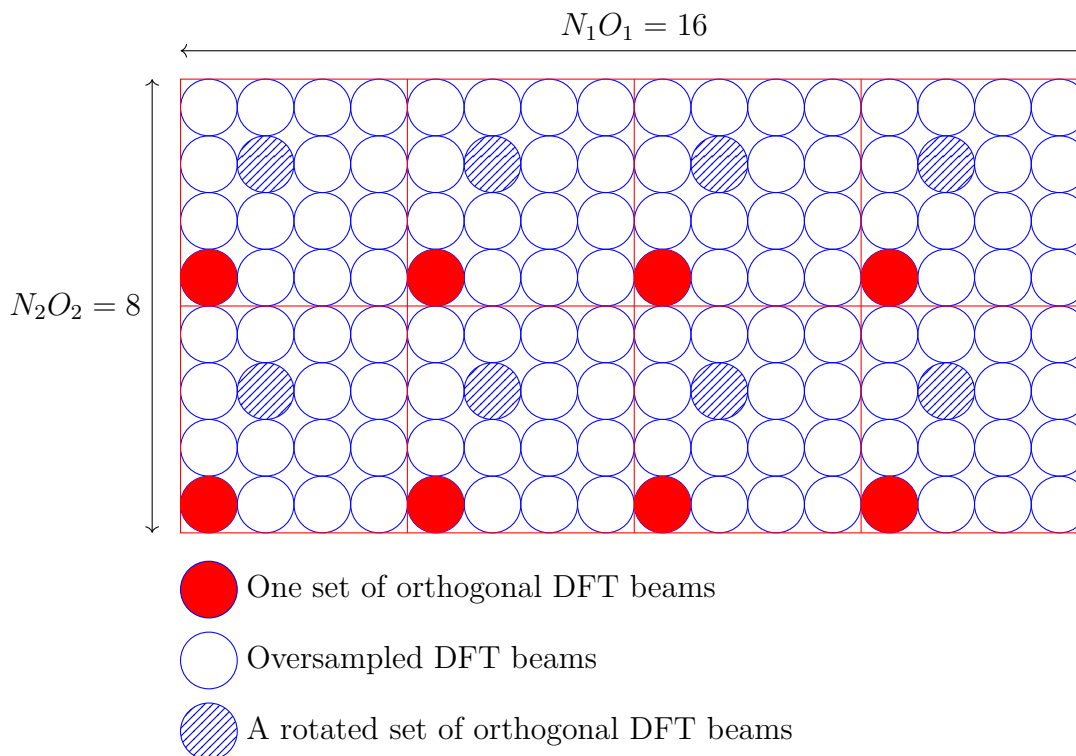
The first stage matrix,  $\mathbf{W}_1$ , contains information on long-term frequency independent characteristics of the channel, while the second stage matrix,  $\mathbf{W}_2$ , contains information on short-term frequency dependent characteristics per subband. Thus,  $\mathbf{W}_1$  can be thought of as capturing the dominant beam direction(s) and  $\mathbf{W}_2$  as capturing the frequency-selective properties [2]. More specifically,  $\mathbf{W}_1$  contains the beamforming vectors and  $\mathbf{W}_2$  applies cophasing for the beams and polarizations so that the signals add up coherently at the receiver. For the Type II codebooks,  $\mathbf{W}_2$  can also adjust the amplitudes for different beams.

In particular,  $\mathbf{W}_1$  has a block diagonal form with entries  $\mathbf{B}$  on the main diagonal and is written as

$$\mathbf{W}_1 = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}. \quad (2.10)$$

Here  $\mathbf{B}$  is a matrix of size  $N_1 N_2 \times L$ , where  $L$  is the number of beams per polarization and per layer. Each column of  $\mathbf{B}$  represents a discrete Fourier transform (DFT) beam taken from an oversampled DFT matrix of size  $N_1 N_2 \times N_1 N_2 O_1 O_2$ , where  $O_1$  and  $O_2$  are the oversampling factors in the horizontal and vertical dimension, respectively.

Each DFT beam in  $\mathbf{B}$  is orthogonal to every other beam. The reason for having two identical matrices in  $\mathbf{W}_1$  is because the same beams are used for both polarizations, but the two polarizations require cophasing, which is handled by the  $\mathbf{W}_2$  matrix. The possible spatial DFT beams are sometimes visualized in a grid [21, 22], as in Figure 2.6. The red circles correspond to one set of orthogonal DFT beams and the white circles correspond to oversampled beams. The figure also gives an example of the values of  $N_1$ ,  $N_2$ ,  $O_1$ , and  $O_2$ .



**Figure 2.6:** Visualization of the possible beams to select from when calculating  $\mathbf{W}_1$ . The figure shows an example where  $N_1 = 4$ ,  $N_2 = 2$ ,  $O_1 = 4$ , and  $O_2 = 4$ .

Once  $\mathbf{W}_1$  has been computed,  $\mathbf{W}_2$  can be computed. In the Type II codebooks, the co-phasing and amplitude to combine the beams from  $\mathbf{W}_1$  are given by the linear combining (LC) coefficients, which are calculated by applying an SVD of the channel associated with the subband. Finally, the first  $\nu$  columns from  $\mathbf{V}$  are used as the LC coefficients, where  $\nu$  is the number of layers used. The columns of  $\mathbf{V}$  and  $\mathbf{U}$  are known as the gNB-side singular vectors and UE-side singular vectors, respectively.

### 2.5.1.1 Example of a PMI calculation

An example of calculating  $\mathbf{W}$  is presented below. For simplicity, assume that a channel matrix  $\mathbf{H}$  with dimensions  $[N_{\text{rx}} \times N_{\text{tx}} \times N_{\text{PRB}}]$  is available, where  $N_{\text{rx}}$  is the number of antennas at the UE,  $N_{\text{tx}}$  is the number of antennas at the gNB, and  $N_{\text{PRB}}$  is the number of PRBs. Additionally, assume that the PRBs can be grouped into  $N_3$  subbands at that cross-polarized antennas are used at both the UE and gNB.

To compute  $\mathbf{W}_1$ , the strongest DFT beams need to be found. First, however, the strongest rotation factor needs to be found. As mentioned earlier, the DFT beams are given by an oversampled DFT matrix, which for the rotation factors  $q_1 = 0, 1, \dots, O_1 - 1$  and  $q_2 = 0, 1, \dots, O_2 - 1$  can be computed by [21]

$$\tilde{\mathbf{B}}(q_1, q_2) = [\mathbf{R}_{N_1}(q_1) \mathbf{D}_{N_1}] \quad [\mathbf{R}_{N_2}(q_2) \mathbf{D}_{N_2}] \quad (2.11)$$

where  $\otimes$  denotes Kronecker product,  $\mathbf{R}_N(q) = \text{diag}\left([1 \ e^{j2\pi\frac{q}{NO}} \ \dots \ e^{j2\pi(N-1)\frac{q}{NO}}]\right)$  is a rotation matrix with oversampling factor  $O$ , and  $\mathbf{D}_N$  is an  $N \times N$  DFT matrix with entries  $[\mathbf{D}_N]_{m,n} = \frac{1}{\sqrt{N}}e^{j2\pi\frac{mn}{N}}$ ,  $m, n = 0, 1, \dots, N-1$ . Note that  $\tilde{\mathbf{B}}$  in (2.11) contains all  $N_1N_2$  DFT beams from one rotation factor, whereas  $\mathbf{B}$  in (2.10) only contains  $L$  out of  $N_1N_2$  beams.

To find the strongest rotation factor, all the beams from each rotation factor can be correlated with the covariance of the channel, averaged over all subbands. This gives the power per beam

$$P_i = |\mathbf{b}_i^H \mathbf{R} \mathbf{b}_i|, \quad (2.12)$$

where  $P_i$  is the power from beam  $i$ ,  $|\cdot|$  denotes absolute value,  $\mathbf{b}_i$  is the  $i$ :th beam from the DFT matrix and  $\mathbf{R}$  is the covariance of the channel matrix averaged over all subbands. For polarization  $r$ , the covariance of the channel can be calculated as

$$\mathbf{R}_r = \frac{1}{N_3} \sum_{f=1}^{N_3} \mathbf{H}_{f,r}^H \mathbf{H}_{f,r}, \quad (2.13)$$

where  $\mathbf{H}_{f,r}$  denotes the  $\frac{N_{\text{rx}}}{2} \times \frac{N_{\text{tx}}}{2}$  channel matrix for subband  $f$  and polarization  $r$ . Finally, the  $L$  strongest beams from the strongest rotation factor are selected and  $\mathbf{W}_1$  is of size  $2N_1N_2 \times 2L$  with entries  $\mathbf{b}_1, \dots, \mathbf{b}_L$ , arranged as

$$\mathbf{W}_1 = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_L & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_L \end{bmatrix}. \quad (2.14)$$

Once  $\mathbf{W}_1$  has been computed,  $\mathbf{W}_2$  can be computed. The LC coefficients are given by calculating the SVD of the channel per subband. In this example, the beamspace channel for subband  $f$  is given by

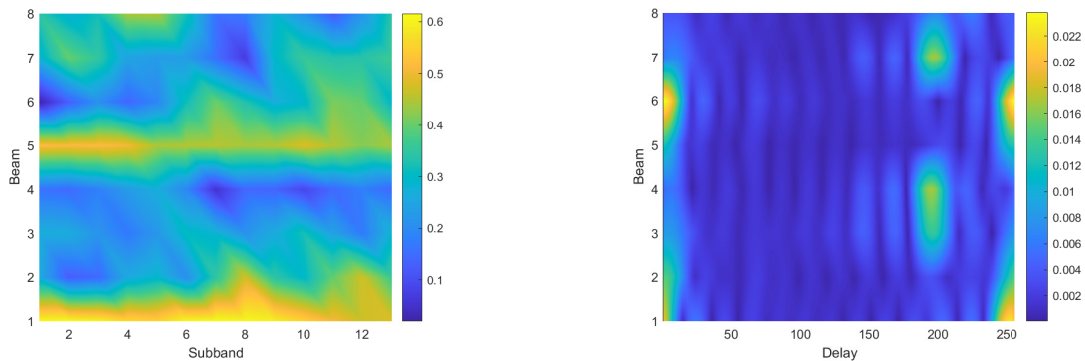
$$\mathbf{H}_{\text{beamspace},f} = \mathbf{H}_f \mathbf{W}_1, \quad (2.15)$$

where  $\mathbf{H}_f$  is the  $N_{\text{rx}} \times N_{\text{tx}}$  channel matrix for subband  $f$ . Subsequently, an SVD is applied on  $\mathbf{H}_{\text{beamspace},f}$ , and the first  $\nu$  columns of  $\mathbf{V}$  are used to construct  $\mathbf{W}_2$ . For one layer  $k$ ,  $\mathbf{W}_2$  is denoted  $\mathbf{W}_{2,k}$  and has dimensions  $2L \times N_3$ . The final precoder  $\mathbf{W}$  for layer  $k$  is then constructed at the gNB as

$$\mathbf{W}_k = \mathbf{W}_1 \mathbf{W}_{2,k}. \quad (2.16)$$

## 2.5.2 Release 16 codebooks

There were two major drawbacks of the Release 15 Type II codebooks: large reporting overhead and high computational complexity [4]. In Release 16, the reporting overhead was addressed, resulting in the eType II codebook and the eType II PS codebook.



(a)  $\mathbf{W}_2$  in the beam-frequency domain for one layer.

(b)  $\mathbf{W}_2$  in the beam-delay domain for one layer.

**Figure 2.7:** (a) Energy of  $\mathbf{W}_2$  in the beam-frequency and (b) beam-delay domain. The beam-delay domain is much sparser and allows for compression.

The  $\mathbf{W}_2$  matrix in the beam-frequency domain is transformed to the beam-delay domain by applying a  $N_3 \times N_3$  DFT matrix over the frequency dimension. Figure 2.7(a) and 2.7(b) shows  $\mathbf{W}_2$  before and after applying the DFT matrix. The reporting overhead is reduced by keeping  $M$  of delay taps, where  $M < N_3$ . Subsequently, the strongest  $K_1$  out of  $K_0 = 2LM$  beam-delay pairs are selected for transmission. The parameter  $\beta$  selects the fraction of beam-delay pairs to be reported, and takes values  $\{1/4, 1/2, 3/4\}$ , thus  $K_1 = \beta 2LM$ . The precoder for layer  $k$  is finally given by

$$\mathbf{W}_k = \mathbf{W}_1 \tilde{\mathbf{W}}_{2,k} \mathbf{W}_{f,k}^H, \quad (2.17)$$

where  $\tilde{\mathbf{W}}_{2,k}$  is of size  $2L \times M$ , and  $\mathbf{W}_{f,k}^H$  is of size  $M \times N_3$ .

In Release 16, using the phase ambiguity property of the SVD, the columns of  $\mathbf{W}_2$  can be rotated in a such a way that the strongest beam over all subbands obtains zero phase. In addition to reducing feedback overhead, normalizing in this way also decreases the effects of phase jumps, which otherwise would create an additional frequency component and make the matrix less sparse in the beam-delay domain [16].

The beams selected in  $\mathbf{W}_1$  are reported based on defined tables in [19] and only the indices of the beams are reported, which limits overhead. The same is true for  $\mathbf{W}_f^H$ . Furthermore,  $\tilde{\mathbf{W}}_2$  is normalized per polarization and is quantized with  $N_{\text{phase}}$  bits for phase and  $N_{\text{ampl}}$  bits for amplitude. The amplitude is quantized on a logarithmic scale, whereas the phase is uniformly quantized using a  $2^{N_{\text{phase}}}$ -phase-shift-keying constellation.

## 2.6 AR modeling

To perform prediction on the PMI, a model for the beam-delay pairs in  $\tilde{\mathbf{W}}_2$  is required. AR modeling is used in many applications, such as in finance and biology, but can also be used for modeling and predicting a wireless channel [9, 10]. An AR model has an output variable that can be estimated by a linear combination of its previous values. For example, a complex channel scalar  $h_t$  at time  $t$  may be modeled as a weighted sum of  $p$  previous values via

$$h_t = -a_1 h_{t-1} - a_2 h_{t-2} - \dots - a_p h_{t-p} + e_t \quad (2.18)$$

where  $a_1, \dots, a_p$  are the model coefficients,  $p$  is the model order, and  $e_t$  is the process noise.

One way of obtaining the model coefficients is by solving the Yule-Walker equations, which are given by [23]

$$\mathbf{a} = \mathbf{R}^{-1} \mathbf{v} \quad (2.19)$$

$$\text{where } \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \mathbf{R} = \begin{bmatrix} R(0) & R(-1) & \dots & R(-p+1) \\ R(1) & R(0) & \dots & R(-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix}, \mathbf{v} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}$$

and  $R(\tau) = \mathbb{E}[\mathbf{h}(t-\tau)\mathbf{h}^*(t)]$  is the autocorrelation of  $\mathbf{h}(t)$  at lag  $\tau$ . The notation  $(\cdot)^*$  refers to complex conjugate. The number of samples in  $\mathbf{h}(t)$  is  $N_{\text{bank}}$  and will be referred to as bank size, while  $\mathbf{h}(t)$  itself will be referred to as training data.

The model order and bank size are two important parameters. A high model order will lead to accurate predictions if the training data has little noise, but will result in increased complexity for the Kalman filter. However, if the training data is noisy, a lower model order can be beneficial [9]. If the model coefficients are updated frequently, a too high model order might not be able to change fast enough if the channel changes. Finally, the model order cannot be larger than the bank size. The bank size should be selected to capture the statistics of the channel. Once the model order and bank size has been selected and the coefficients have been obtained, they can be used with the Kalman filter equations.

## 2.7 The Kalman Filter

The AR model can be used in combination with a Kalman filter, which was first described in a 1960 paper [24] by Rudolf E. Kálmán. It is an optimal filter in terms of minimizing the mean squared error (MSE) between the estimated value and the true value, assuming the error is normally distributed. An advantage of the Kalman filter compared to the Wiener filter, which is also optimal in the same sense, is

that the computational complexity of the Kalman filter does not increase with each measurement [9].

For channel prediction, the Kalman filter is based on the state space model [25]

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{b}e_t \quad (2.20)$$

$$h_t = \mathbf{c}\mathbf{x}_t \quad (2.21)$$

$$y_t = h_t + v_t \quad (2.22)$$

where  $\mathbf{A} = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ ,  $\mathbf{b} = [1 \ 0 \ \cdots \ 0]^T$ , and  $\mathbf{c} = [1 \ 0 \ \cdots \ 0]$  are

the state transition matrices. These may be time-varying, but are in this thesis stationary as we assume the channel to be wide sense stationary (WSS). The state  $\mathbf{x}_t = [h_t \ h_{t-1} \ \cdots \ h_{t-p+1}]^T$  is of dimension  $p \times 1$ ,  $y_t$  is the measured channel, and  $v_t \sim \mathcal{CN}(0, \sigma_o^2)$  is the measurement noise. The term  $e_t$  is the process noise. Note that (2.20) corresponds to the model in (2.18).

The true state  $\mathbf{x}_t$  is in practice not known. Therefore, the Kalman filter needs to create an estimate of the state, denoted  $\hat{\mathbf{x}}_t$ . The Kalman filter does this in two steps: a prediction step and an update step.

In the prediction step, the Kalman filter estimates the state and the error covariance by

$$\hat{\mathbf{x}}_{(t+1/t)} = \mathbf{A}\hat{\mathbf{x}}_{(t/t)} \quad (2.23)$$

$$\mathbf{P}_{(t+1/t)} = \mathbf{A}\mathbf{P}_{(t/t)}\mathbf{A}^H + \mathbf{Q}_p, \quad (2.24)$$

where  $\mathbf{P}$  is the error covariance matrix and  $\mathbf{Q}_p$  is the process noise covariance matrix. The notation  $(\cdot)_{(t+1/t)}$  denotes a prediction for time  $t + 1$ , using measurements up until time  $t$ .

Whenever a measurement is received, the Kalman filter progresses to the update step. It first calculates the Kalman gain,  $\mathbf{K}_t$ , and then the updated state and error covariance matrix using

$$\mathbf{K}_t = \mathbf{P}_{(t/t-1)}\mathbf{c}^H \left( \mathbf{c}\mathbf{P}_{(t/t-1)}\mathbf{c}^H + \sigma_o^2 \right)^{-1} \quad (2.25)$$

$$\hat{\mathbf{x}}_{(t/t)} = \hat{\mathbf{x}}_{(t/t-1)} + \mathbf{K}_t \left( y_t - \mathbf{c}\hat{\mathbf{x}}_{(t/t-1)} \right) \quad (2.26)$$

$$\mathbf{P}_{(t/t)} = (\mathbf{I} - \mathbf{K}_t\mathbf{c})\mathbf{P}_{(t/t-1)}. \quad (2.27)$$

If there are time slots with missing measurements, the Kalman filter will continue to predict without performing the update step.



# 3

## Methods

To investigate the predictability of codebooks, a simulation framework that allows comparing different configurations is needed. One option would be to use a system-level simulator, which replicates the behavior between the UE and gNB. Although this option would produce accurate results, it would extend the scope of the thesis and have slow experimentation cycles.

Another option, which is the one used for this thesis, relies on using simulated channel samples from a given model. For each simulation, the actual transmission is substituted by calculating the coefficients that would be transmitted in the CSI report. Prediction is performed only using what is available at the gNB and then precoders can be calculated, whose performances can be compared using different metrics.

### 3.1 Channel model

Similar to close studies [26], simulations were based on a clustered delay line (CDL) channel model, as specified in [27]. The CDL models a 3D system using multiple clusters with pre-defined angles of arrival and angles of departure, making it well suited for MIMO link-level simulations.

More specifically, the chosen CDL model profile was the CDL-C, which defines a non line of sight (NLOS) environment. Note that the codebook design is independent of the channel model.

### 3.2 Codebook design

As a baseline for comparison, the Release 16 eType II codebook was implemented in MATLAB. The implementation is similar to the one presented in Section 2.5.1 and 2.5.2. The implementation is compliant with the standard in [19] except for the quantization. While the standard uses a logarithmic scale for quantizing the amplitude, this thesis used linear quantization for simplicity.

Furthermore, only SP ULAs are used in this thesis, but the ideas presented can be generalized to include other antenna configurations.

To narrow the scope of project, the channel is considered to be perfectly estimated

at the UE. This assumption can be removed by tweaking the process noise in the Kalman equations. Also, the precoder assumes that the power is equally distributed between the layers, a behavior that may be different between different gNB vendors.

### 3.3 Kalman filter and AR model implementation

The Kalman filter was implemented in MATLAB based on the equations in 2.7 to enable prediction, a reporting periodicity of  $N$  was assumed, which meant that only every  $N$ -th measurement was available to the gNB. Moreover, a feedback delay of  $\delta$  time slots was assumed. This is to simulate an environment where the received CSI reports are outdated. In real applications, the delay is mainly due to computational and scheduling delays. The Kalman filter was designed to compensate for the delay by predicting  $\delta$  time slots into the future. Moreover, considering that the channel is perfectly estimated at UE, the measurement noise is set to zero.

For the AR model, the model coefficients given by the Yule-Walker equations in (2.19) could be updated at each time slot. However, as the channel is assumed to be WSS for a number of time slots, the model coefficients were only calculated once based on the training data.

### 3.4 Performance evaluation of precoders

To compare the different prediction options and their precoders, the throughput was calculated assuming a linear minimum mean squared error (LMMSE) equalizer, which optimally compromises capturing the energy of from one layer and canceling inter-layer interference [15]. To calculate the throughput, the signal-to-interference-plus-noise ratio (SINR) is needed, in which the interference refers to inter-layer interference. For an LMMSE equalizer, it is given by [28]

$$\text{SINR}_k = \frac{1}{\text{MMSE}_k} - 1 = \frac{1}{\left[ \left( \mathbf{I}_\nu + \frac{\text{SNR}}{N_{\text{rx}}} \mathbf{H}_{\text{eff}}^H \mathbf{H}_{\text{eff}} \right)^{-1} \right]_{kk}} - 1 \quad (3.1)$$

where  $\mathbf{I}_\nu$  is a  $\nu \times \nu$  identity matrix and  $\mathbf{H}_{\text{eff}} = \mathbf{H}_f \mathbf{W}$  is the effective channel. Subsequently, the total achievable rate is given by adding the achievable rate per layer as

$$R_{\text{tot}} = \sum_{k=1}^{\nu} \log_2 (1 + \text{SINR}_k). \quad (3.2)$$

The training data was not included in the throughput calculations.

# 4

## Results

Several simulations using the the aforementioned methods were performed to study predictability. The results are presented in this chapter.

### 4.1 Simulation setup

Samples from a CDL-C channel model were created with a UE speed of 30 km/h. The channel matrix had four dimensions and were of size  $N_{\text{rx}} \times N_{\text{tx}} \times N_{\text{PRB}} \times N_{\text{slots}}$ , where  $N_{\text{rx}}$  is the number of receiver ports,  $N_{\text{tx}}$  the number of ports at the transmitter,  $N_{\text{PRB}}$  the number of PRBs, and  $N_{\text{slots}}$  the number of time slots. The bandwidth of the channel was 10 MHz and used a subcarrier spacing of 15 kHz, resulting in 52 PRBs. Moreover, there were 4 PRBs per subband, resulting in a total of  $N_3 = 13$  subbands.

The number of receiver and transmitter ports in the simulations are  $N_{\text{rx}} = 2$  and  $N_{\text{tx}} = 16$ , respectively. Orthogonal polarizations are used. An oversampling factor of  $O_1 = 4$  was used and the number of selected beams was  $L = 4$ . For the compression of  $\mathbf{W}_2$ ,  $M = 6$  out of  $N_3 = 13$  delays were used, and a fraction  $\beta = 1/4$  of the beam-delay pairs were reported. The beam-delay pairs were quantized with  $N_{\text{ampl}} = 3$  bits for the amplitude and  $N_{\text{phase}} = 4$  bits for the phase.

For the Kalman filter, the reporting periodicity was set to  $N = 5$  ms, and the reporting delay  $\delta = 4$  ms. A bank size of  $N_{\text{bank}} = 150$  was used, which calculated  $p = 4$  model coefficients. The channel, codebook, and Kalman filter parameters are summarized in Table 4.1

**Table 4.1:** Simulation parameters for the channel, codebook, Kalman filter and AR model.

	Parameter	Value	Description
Channel	Bandwidth	10 MHz	
	Subcarrier spacing	15 kHz	
	$N_{\text{PRB}}$	52	Number of PRBs
	PRBs per subband	4	
	$N_3$	13	Number of subbands
	$N_{\text{slots}}$	500	Number of time slots
	Slot interval	1 ms	Time between channel samples
	UE speed	30 km/h	
	$f_c$	3.5 GHz	Carrier frequency
Codebook	$N_1$	8	Antennas per polarization in horizontal direction
	$N_2$	1	Antennas per polarization in vertical direction
	$O_1$	4	Oversampling factor in horizontal direction
	$O_2$	1	Oversampling factor in vertical direction
	$L$	4	Number of beams in $\mathbf{W}_1$
	$M$	6	Number of delays in $\mathbf{W}_f$
	$\beta$	1/4	Fraction of beam-delay pairs
	$\nu$	2	Number of layers
	$N_{\text{phase}}$	4	Number of bits for phase quantization
	$N_{\text{ampl}}$	3	Number of bits for amplitude quantization
Kalman filter	$N$	5 ms	Reporting periodicity
	$\delta$	4 ms	CSI delay
	$\sigma_o$	0	Standard deviation of measurement noise
	$N_{\text{bank}}$	150	Amount of training data for AR model
	$p$	4	Model order

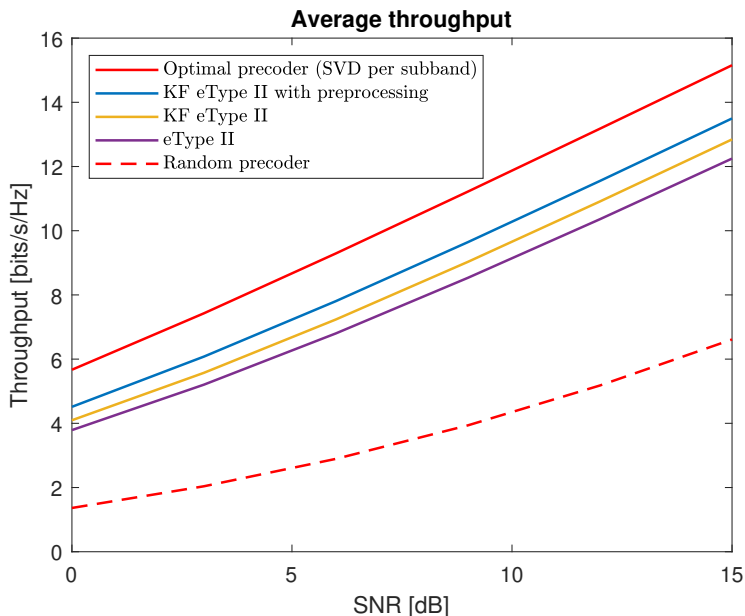
## 4.2 Prediction-based on the eTypeII codebook

Figure 4.1 shows the average throughput as a function of SNR for prediction based on the eType II codebook. The throughput is averaged over the first 100 time slots after the training data.

There are five curves. The solid red curve corresponds to an optimal precoder with full CSI that calculates the SVD per subband of the beam-space channel and uses  $\mathbf{V}$  as a precoder. It can be seen as an upper bound. The red dotted curve corresponds to a random precoder where the precoder is a random vector and can be seen as a lower bound.

The purple curve corresponds to a codebook that does not perform any prediction. Instead, it calculates a new precoder every time it has access to a measurement, which is every 5:th time slot. When it does not have access to a measurement, it reuses the previously calculated precoder. This is closest to the existing eType II codebook.

The yellow curve corresponds to the eType II codebook where a Kalman filter is implemented. To enable prediction,  $\mathbf{W}_1$ ,  $\mathbf{W}_f$ , and the beam-delay pairs are calculated



**Figure 4.1:** Throughput versus SNR for prediction methods based on the eType II codebook. The KF eType II with preprocessing and the KF eType II lines correspond to methods with prediction, while eType II is the baseline model where no prediction is performed.

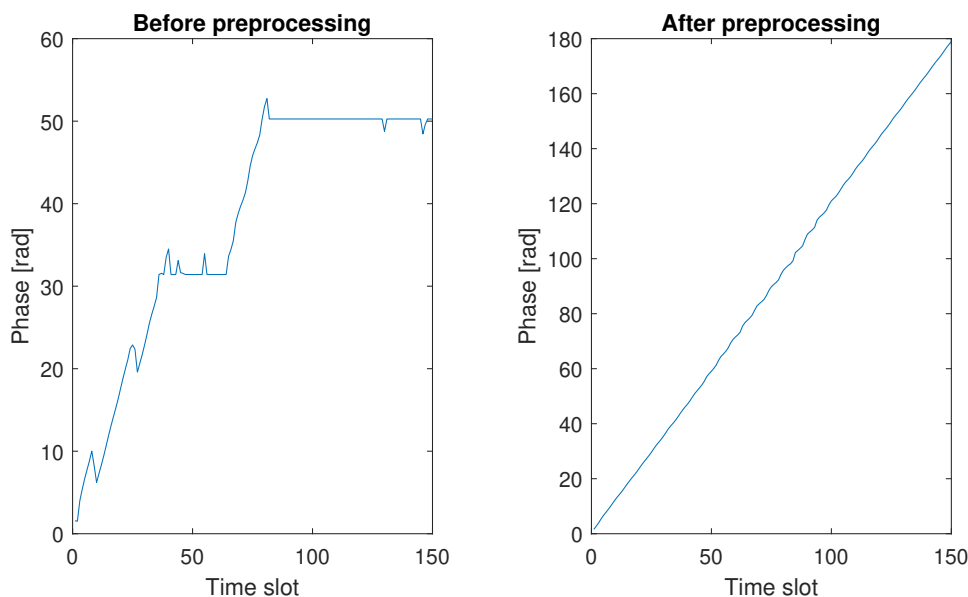
for the first time slot and reused for all subsequent time slots<sup>1</sup>. The  $\tilde{\mathbf{W}}_2$  coefficients are then predicted using the previous values for the respective beam-delay pairs over time. The Kalman filter provides a gain of about 1 dB compared to the eType II.

The blue curve is an enhanced version of the yellow curve. It obtains a higher throughput by performing a preprocessing step in  $\tilde{\mathbf{W}}_2$  at the gNB, which consists of normalizing the beam-delay pairs by the same beam for all time slots. The gain is about 1.9 dB compared to the eType II.

The eType II codebooks with a Kalman filter outperforms the existing eType II codebook and applying the preprocessing step increases the throughput further. Figure 4.2 shows the phase over time for the same beam delay pair. On the left hand side, no preprocessing is done, while on the right hand side the preprocessing step is performed. Applying the preprocessing step reduces phase jumps caused by the phase ambiguity of the SVD, thus increasing the predictability.

More insight can be gained by viewing throughput over time for a fixed SNR value, which can be seen in Figure 4.3. The throughput over time is shown for the first 125 slots after the training data. The spiky behavior from the eType II codebook (purple curve) can be attributed to using outdated precoders. Whenever a measurement is received, the precoder is updated and the throughput increases. The KF eType II with preprocessing maintains a high throughput until the 250:th time slot, after which the channel decorrelates and the negative effect of keeping  $\mathbf{W}_1$ ,  $\mathbf{W}_f$ , and the

<sup>1</sup>As  $\mathbf{W}_1$  models the strongest directions from the gNB to the UE, these do not change over a time frame of less than one second for most scenarios.



**Figure 4.2:** Phase over time for a beam-delay pair, with and without the preprocessing step. Applying the preprocessing step decreases the phase jumps, making the phase more continuous and increasing the predictability.

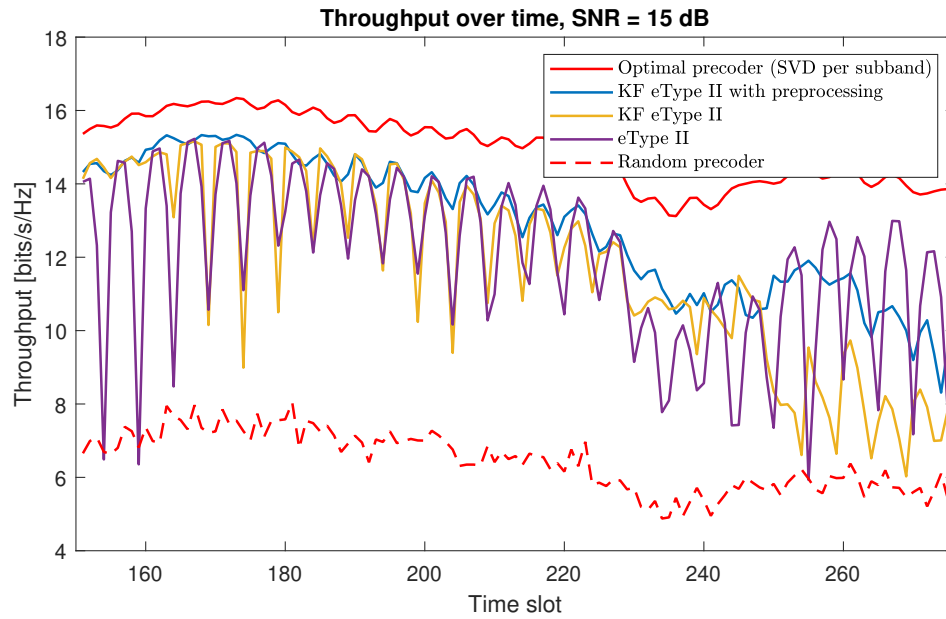
beam-delay pairs fixed becomes apparent.

### 4.3 Prediction based on augmented reporting scheme

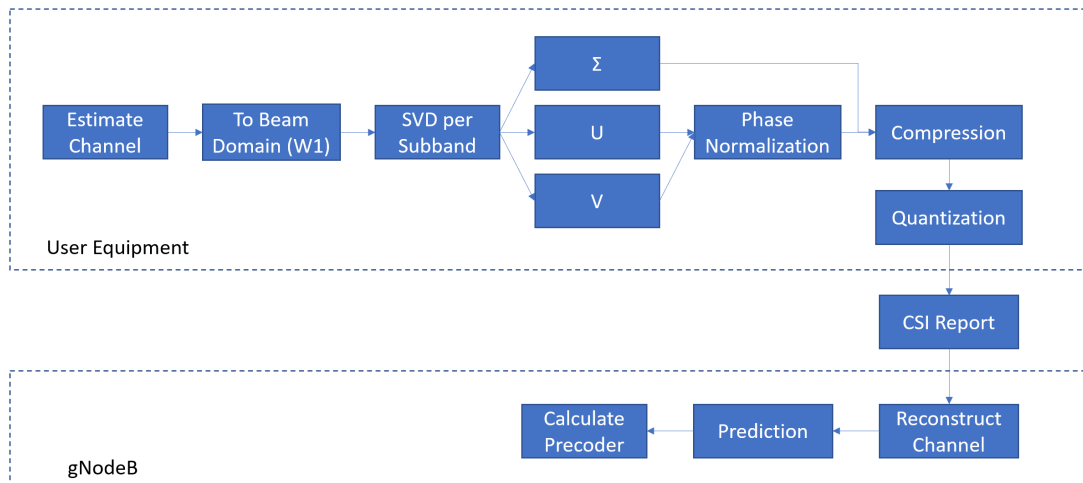
While the eType II codebook relies on only transmitting a compressed version of the gNB-side singular values of the beamspace channel ( $\mathbf{V}$ ), one idea could be to also transmit the UE-side singular vectors and singular values ( $\mathbf{U}$  and  $\mathbf{\Sigma}$ ). This would enable the gNB to reconstruct the channel, which is known to be predictable [9]. However, this would add a lot of overhead to the PMI report. Without compression,  $N_{\text{rx}} \times \nu \times N_3$  coefficients are needed to transmit  $\mathbf{U}$  while  $\nu \times N_3$  are needed for  $\mathbf{\Sigma}$ . Below, we present a method for compressing the information, as well as a preprocessing step that aids the compression.

A block diagram of the procedure is shown in Figure 4.4. Similar to the eType II codebook, an SVD is applied on the beamspace channel  $\mathbf{H}_{\text{beamspace}} = \mathbf{H}_f \mathbf{W}_1$ . After the SVD, the phase in the columns of  $\mathbf{U}$  and  $\mathbf{V}$  are normalized by the strongest row in  $\mathbf{U}$ . Any row of  $\mathbf{U}$  or  $\mathbf{V}$  could be used for normalization, but normalizing by  $\mathbf{U}$  yielded higher throughputs. Figure 4.5 visualizes how  $\mathbf{U}$  and  $\mathbf{V}$  are normalized.

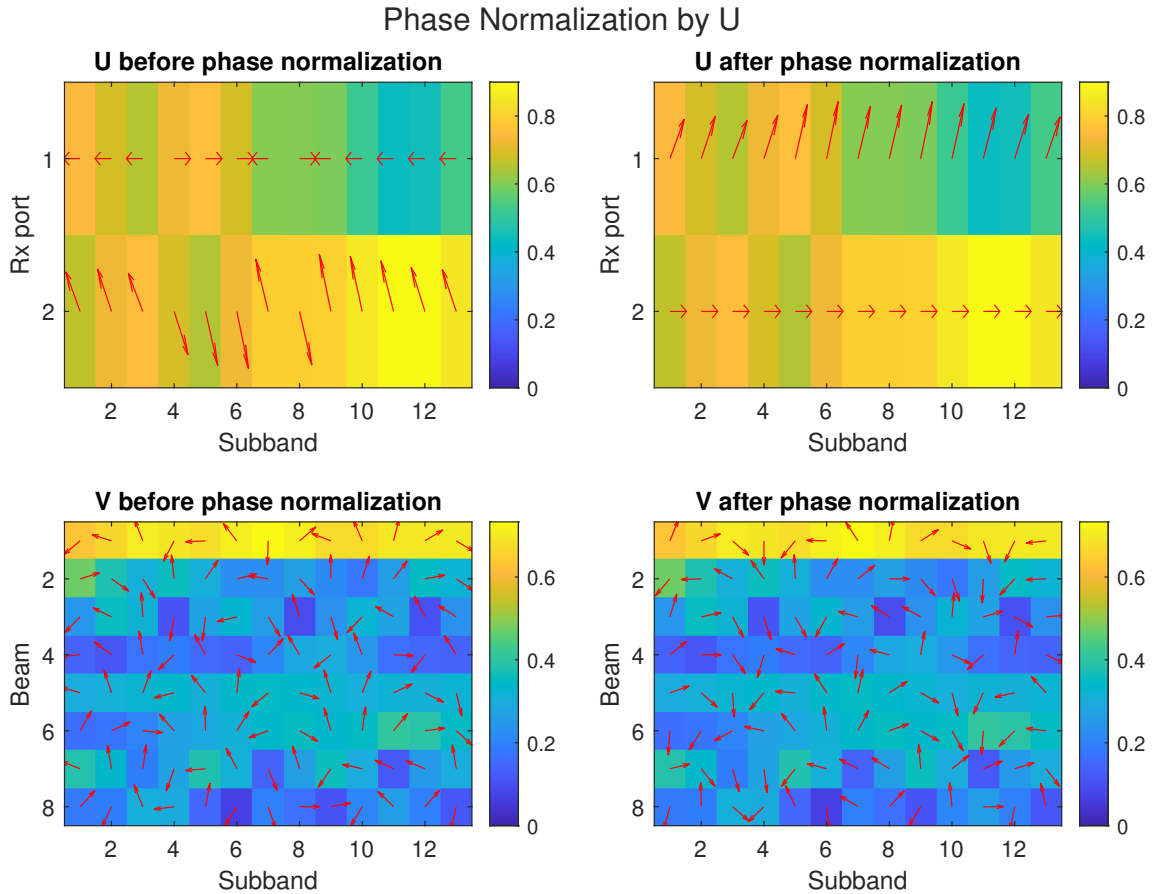
After normalizing, the matrices  $\mathbf{U}$ ,  $\mathbf{\Sigma}$  and  $\mathbf{V}$  are compressed. The matrix  $\mathbf{V}$  is compressed in the same way as in the eType II codebook. The compression of  $\mathbf{U}$  is as follows. By applying a DFT matrix on  $\mathbf{U}$  over the subband domain, it was observed that most of the energy is concentrated on the first delay, meaning that the gNB side singular vectors take care of most of the frequency selectivity of the channel. By transmitting only the coefficients from the first delay, only  $N_{\text{rx}} \times \nu$



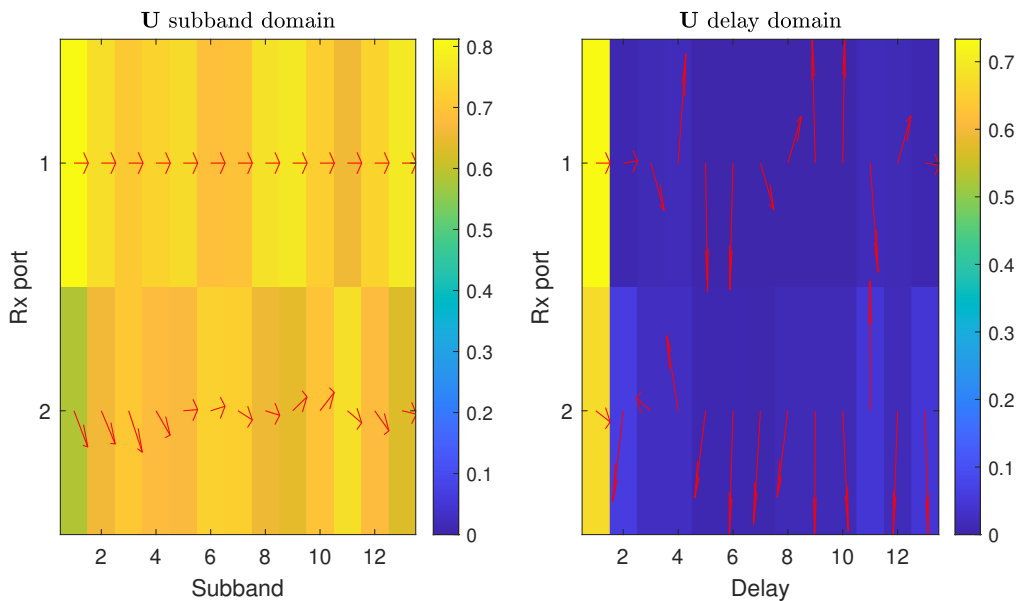
**Figure 4.3:** Throughput over time for prediction methods based on the eType II codebook.



**Figure 4.4:** Block Diagram of the augmented method.



**Figure 4.5:** Phase normalization of  $\mathbf{U}$  and  $\mathbf{V}$  with the objective of making the second receiver port coefficients with phase zero. Here, only the first layer is presented, other layers must be normalized independently. In these plots, the color represent the amplitude of the coefficient while the red arrow represent the phase.



**Figure 4.6:** Applying a DFT matrix on  $\mathbf{U}$  shows that the energy of is concentrated in the first delay, allowing for compression.

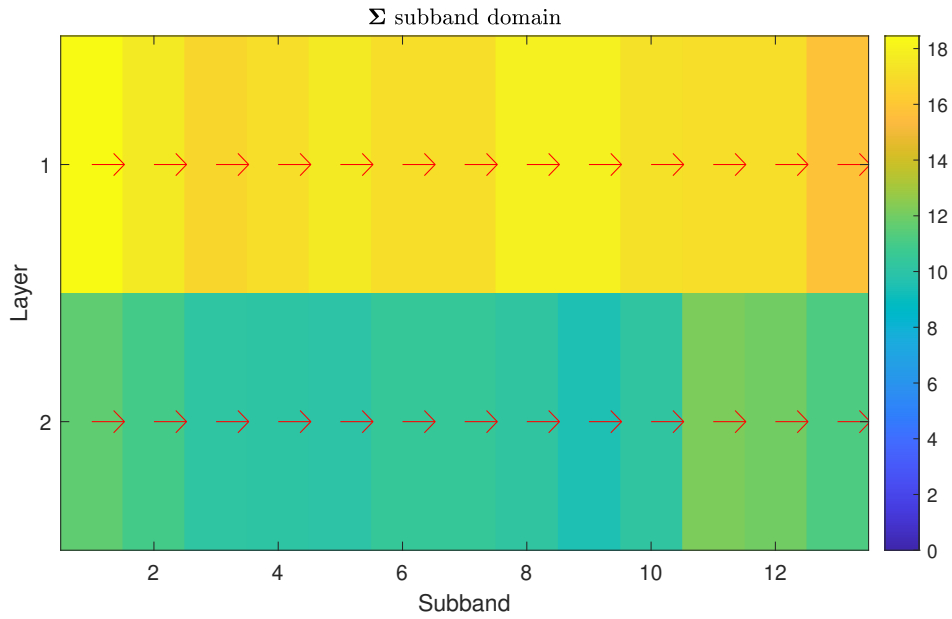
coefficients are required<sup>2</sup>.  $\mathbf{U}$  can be compressed further by only transmitting the phase. Results indicate that this compression did not considerably affect the quality of the precoders. Figure 4.6 shows  $\mathbf{U}$  before and after applying the DFT matrix. Note that the energy is concentrated on the first delay, which allows reporting  $\mathbf{U}$  in a wideband manner.

The matrix  $\mathbf{\Sigma}$  can either be reported in a wideband manner similar to  $\mathbf{U}$ , or not transmitted at all. As can be seen from Figure 4.7, the singular values in  $\mathbf{\Sigma}$  for both layers is almost the same for all subbands, which means that it can be reported in a wideband manner. However, the quality of the precoders did not degrade when assuming  $\mathbf{\Sigma}$  was an identity matrix or a diagonal matrix with values  $2^{-(k-1)}$  where  $k$  is the layer index, meaning that  $\mathbf{\Sigma}$  might not have to be reported at all. Note that this might be an artifact since only SU-MIMO performance is evaluated. For MU-MIMO, there could be benefits of reporting  $\mathbf{\Sigma}$ , which allows better scheduling decisions.

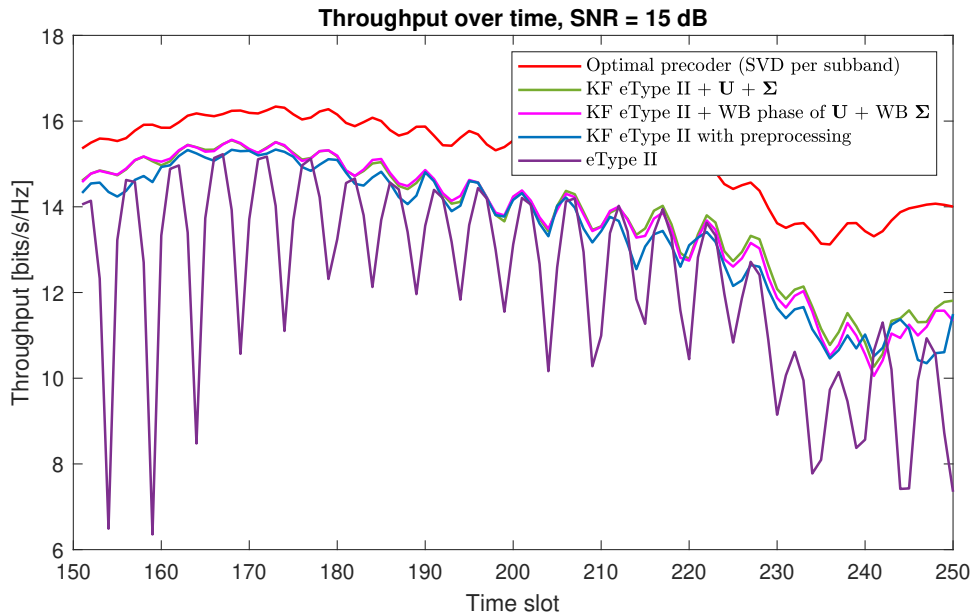
Figure 4.8 compares the throughput over time for the enhanced version and the eType II with preprocessing. It can be seen that transmitting only wideband information of  $\mathbf{U}$  and  $\mathbf{\Sigma}$  has a similar performance to transmitting full information. As can be seen from Figure 4.9, the gain is approximately 2.2 dB when compressing the information, and about 2.3 dB without compression.

Figure 4.10 shows the throughput of different CSI reporting schemes as a function of the number of bits transmitted. Note that the compression of  $\mathbf{U}$  and  $\mathbf{\Sigma}$  significantly reduces the number of bits per PMI report, but has little effect on the throughput.

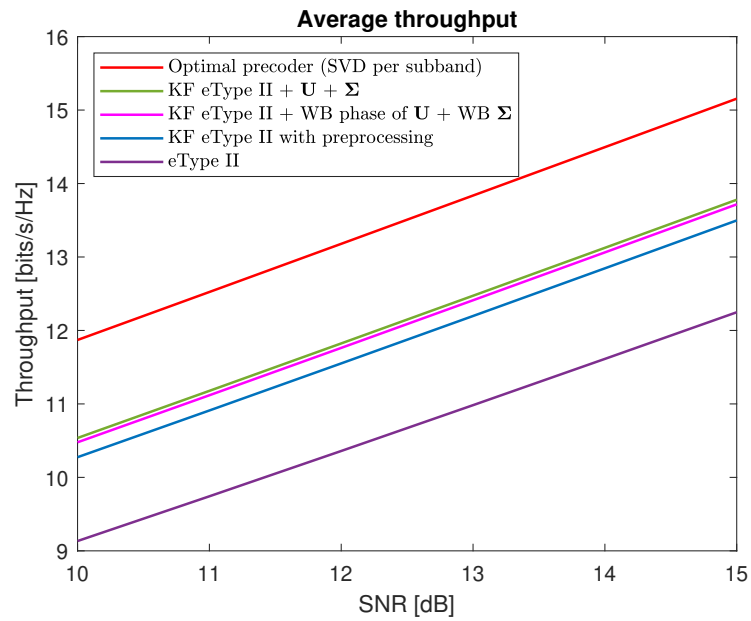
<sup>2</sup>This number could be further reduced considering that the layers must be orthogonal, therefore some coefficients can be perfectly derived from others.



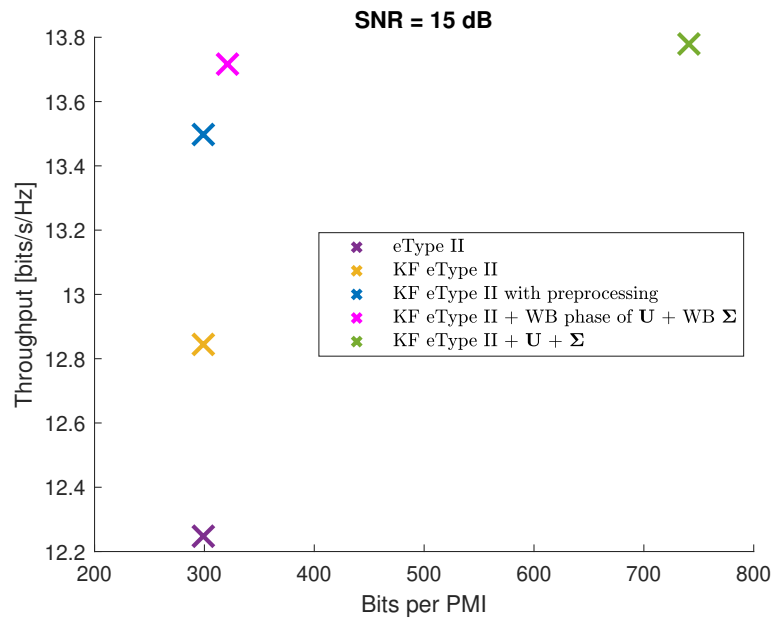
**Figure 4.7:** The singular values in  $\Sigma$  are almost the same for all subbands, allowing them to be reported in a wideband manner.



**Figure 4.8:** Comparison between the regular KF eType II with preprocessing against the augmented method, with and without compression.



**Figure 4.9:** Throughput as a function of SNR. Note that the gain of the augmented method does not decrease significantly by compressing  $\mathbf{U}$  and  $\mathbf{\Sigma}$ .



**Figure 4.10:** Throughput as a function of bits per PMI report.



# 5

## Discussion

This section discusses the obtained results and comments on their validity and significance. It also provides an outlook.

### 5.1 Key findings

The key takeaway from Section 4.2 is that predicting based on the current eType II codebook is feasible, given that the preprocessing step is performed. The physical interpretation of the preprocessing step is that one beam is used as a reference point for  $\tilde{\mathbf{W}}_2$  over time, allowing the phase of the other beams to change in a more continuous and thus predictable manner.

The results could be further improved if  $\mathbf{W}_1$  and  $\mathbf{W}_f$  were updated more frequently since it was seen that fixing them for more than 100 time slots eventually decreased performance. This is left as a potential for future work, where investigating how often  $\mathbf{W}_1$  needs to be updated as a function of UE mobility could also be interesting.

The key takeaway from Section 4.3 is that predicting based on a reconstructed beamspace channel yielded higher throughputs, even when some of the information was compressed. One possible reason for this behavior is that by transmitting  $\mathbf{U}$ , additional information on how the layers are combined and the Doppler information at the receiver side is available at the gNB, which increases the prediction quality.

To reconstruct the channel at the gNB, compressing  $\mathbf{U}$  and  $\Sigma$  was crucial to limit overhead. Applying a DFT on  $\mathbf{U}$  resulted in a sparse matrix and allowed for compression. The reason for the sparse matrix could potentially be due to the low amount of receiver antennas, or that the columns of  $\mathbf{V}$  somehow precompensates for the phase differences. It is possible that  $\mathbf{U}$  is not as sparse if the receiver has more antennas.

### 5.2 Simulation parameters

For the simulations, a bank size of 150 was used. In practice, training data of this size might be unfeasible since the channel might change a lot during the time it takes for that training data to be obtained. A possible solution would be to use a burst CSI reporting scheme, as presented in [4].

The simulations used a model order of 4. This is in line with previous works [9, 10] that considered model orders of 2-6, depending on the UE speed. Furthermore, the model coefficients could be updated at every time step, but since the channel is assumed to be WSS, the model coefficients should hold for a relatively long time, which was confirmed by for example Figure 4.3.

The thesis used a standardized CDL-C channel model for simulations. Although such a channel model has been used in previous works [26], the result's validity could be further increased by using a system-level simulator. Moreover, additional channel samples could strengthen the statistical validity of the results.

### 5.3 Outlook

Being able to predict the precoder might not only increase the throughput but may also have a positive environmental impact. By the use of a predictor, less frequent RS are necessary, which decreases the energy footprint of the communications system. However, the energy costs of performing the prediction algorithms would also have to be considered.

Having a close estimate of the full CSI at the gNB side might also open the path to other possibilities, as most of the research currently done assumes that the full CSI is known. Commercially, vendors may be able to apply this research on their equipment, improving their quality of service against competitors in other areas than prediction.

# 6

## Conclusion

This thesis studied CSI prediction at the gNB, which is a potential solution to the channel aging problem. The current Release 16 eType II codebook and a proposed new augmented method were compared and analyzed. The study used an AR model combined with the Kalman filter to perform prediction.

It was shown that prediction based on the current Release 16 eType II CSI report performed decently, given that an initial pre-processing step is performed at the gNB. This implies that current vendors can implement prediction algorithms using existing codebooks, which potentially could increase system performance.

The throughput could be increased further by including the additional UE-side singular vectors information into the CSI report, even when compressing the information. This behavior can be explained by the fact that transmitting the extra coefficients allows the gNB to calculate an approximate full CSI from the precoder, which in turn is known to be predictable.

Incorporating prediction methods could potentially reduce overhead as less frequent CSI reports are required to be transmitted, though the costs of running the prediction algorithms need to be estimated.

### 6.1 Future work

A limitation of the study is that only channels generated by channel models were evaluated, further analysis using system-level simulations or even real-life experimentation could increase the confidence of the results.

Another limitation is the lack of a deeper mathematical modeling of the solution. Although the results conform to the high-level concepts of the problem, an in-depth mathematical analysis could provide additional insights and possibly improvements.

Finally, this thesis aimed to show if prediction is possible with AR models, but did not aim to find the best prediction algorithm. More accurate prediction models could be designed for both Release 16 eType II CSI and the augmented method using different techniques, such as machine learning.



# Bibliography

- [1] Ericsson AB, “Ericsson Mobility Report,” Ericsson, Tech. Rep., November 2022.
- [2] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020, p. 231.
- [3] H. Asplund, D. Astely, P. von Butovitsch, T. Chapman, M. Frenne, F. Ghasemzadeh, M. Hagström, B. Hogan, G. Jongren, J. Karlsson *et al.*, *Advanced antenna systems for 5G network deployments: bridging the gap between theory and practice*. Academic Press, 2020.
- [4] V. Ramireddy, M. Grossmann, M. Landmann, and G. D. Galdo, “Enhancements on Type-II 5G New Radio Codebooks for UE Mobility Scenarios,” *IEEE Communications Standards Magazine*, vol. 6, no. 1, pp. 35–40, 2022.
- [5] H. Guo, B. Makki, D.-T. Phan-Huy, E. Dahlman, M.-S. Alouini, and T. Svensson, “Predictor antenna: A technique to boost the performance of moving relays,” *IEEE Communications Magazine*, vol. 59, no. 7, pp. 80–86, 2021.
- [6] 3GPP RP-213598, “MIMO Evolution for Downlink and Uplink,” 3GPP, Tech. Rep., December 2021.
- [7] 3GPP TSG-RAN WG1 #111, “On CSI enhancements for Rel-18 NR MIMO evolution,” Ericsson, Tech. Rep. R1-2212174, November 2022.
- [8] H. Jin, K. Liu, M. Zhang, L. Zhang, G. Lee, E. N. Farag, D. Zhu, E. Onggosanusi, M. Shafi, and H. Tataria, “Massive MIMO evolution towards 3GPP Release 18,” *IEEE Journal on Selected Areas in Communications*, 2023.
- [9] R. Apelfröjd, “Channel estimation and prediction for 5G applications.” Ph.D. dissertation, Uppsala University, 2018.
- [10] D. Aronsson, “Channel estimation and prediction for MIMO OFDM systems: Key design and performance aspects of Kalman-based algorithms.” Ph.D. dissertation, Uppsala University, 2011.
- [11] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, “Channel state information prediction for 5G wireless communications: A deep learning approach,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 227–236, 2018.
- [12] 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, 2021.
- [13] 3GPP, *NR: Physical layer procedures for data (Release 16)*, 2023, Technical Report TR 38.214.
- [14] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.
- [15] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
- [16] R. Ahmed, F. Tosato, and M. Maso, “Overhead reduction of NR type II CSI for NR release 16,” in *WSA 2019; 23rd International ITG Workshop on Smart Antennas*. VDE, 2019, pp. 1–5.
- [17] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, “Chapter 7 - Multiantenna Techniques,” in *5G Physical Layer*, A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, Eds. Academic Press, 2018, pp. 199–252. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128145784000126>
- [18] J. Tan and L. Dai, “Channel feedback in TDD massive MIMO systems with partial reciprocity,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, pp. 12 960–12 974, 2021.
- [19] 3GPP, *NR: Physical layer procedures for data (Release 17)*, 2023, Technical Report TR 38.214.
- [20] E. Onggosanusi, M. S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant *et al.*, “Modular and high-resolution channel state information and beam management for 5G new radio,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 48–55, 2018.
- [21] 3GPP TSG-RAN WG1 #87, “Advanced CSI Codebook Structure,” 3GPP, Tech. Rep. R1-1612661, November 2016.
- [22] D. A. Urquiza Villalonga, H. OdetAlla, M. J. Fernández-Getino García, and A. Flizikowski, “Spectral efficiency of precoded 5G-NR in single and multi-user scenarios under imperfect channel knowledge: A comprehensive guide for implementation,” *Electronics*, vol. 11, no. 24, p. 4237, December 2022.
- [23] K. E. Baddour and N. C. Beaulieu, “Autoregressive modeling for fading channel simulation,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1650–1662, 2005.
- [24] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, 1960.
- [25] A. J. Tenenbaum, R. S. Adve, and Y.-S. Yuk, “Channel prediction and feedback in multiuser broadcast channels,” in *2009 11th Canadian Workshop on Information Theory*. Ottawa, ON, Canada: IEEE, 2009, pp. 67–70.

- [26] Z. Qin, H. Yin, Y. Cao, W. Li, and D. Gesbert, “A partial reciprocity-based channel prediction framework for FDD massive MIMO with high mobility,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9638–9652, 2022.
- [27] 3GPP, *5G; Study on channel model for frequencies from 0.5 to 100 GHz*, 2023, Technical Report TR 38.901.
- [28] P. Li, D. Paul, R. Narasimhan, and J. Cioffi, “On the distribution of SINR for the MMSE MIMO receiver and performance analysis,” *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 271–286, 2005.



DEPARTMENT OF ELECTRICAL ENGINEERING  
CHALMERS UNIVERSITY OF TECHNOLOGY  
Gothenburg, Sweden  
[www.chalmers.se](http://www.chalmers.se)



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY