Technische Universität München

TUM School of Computation, Information and Technology

Generative Modeling-based Feedback Schemes for Precoder and Pilot Design in FDD Systems

Nurettin Turan

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:		Prof. Dr.-Ing. Antonia Wachter-Zeh
Prüfende der Dissertation:	1.	Prof. Dr.-Ing. Wolfgang Utschick
	2.	Prof. Dr.-Ing. Stephan ten Brink

Die Dissertation wurde am 13.02.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 12.08.2025 angenommen.

Abstract

In this dissertation, we propose and analyze generative modeling-based feedback schemes for precoder and pilot design in single- and multi-user multiple-input multiple-output (MIMO) frequency division duplex (FDD) systems. In particular, a versatile Gaussian mixture model (GMM)-based feedback scheme for precoder design is introduced, enabling low computational complexity, parallelization at the mobile terminals (MTs), and adaptability to varying signal-to-noise ratio (SNR) levels, numbers of MTs, pilot configurations, and transmission modes without requiring retraining. To further enhance the proposed precoder design scheme's versatility, methods for adapting a single learned GMM to variable feedback bit budgets are also introduced. Moreover, to reduce the memory and model transfer overhead while maintaining performance structured GMM variants are investigated. Additionally, a graph neural network (GNN)-based framework for statistical precoder design is proposed and is further combined with the GMM-based feedback scheme. Furthermore, a vector quantized-variational autoencoder (VQ-VAE)-based scheme is introduced for precoder design in systems with higher feedback bit budgets. For pilot design, the sum conditional mutual information (CMI) pilot optimization framework, originally designed for multi-user systems with single antenna MTs, is extended to multi-user MIMO (MU-MIMO) systems. Finally, a GMM-based pilot design scheme is proposed, offering versatility without requiring perfect knowledge of channel statistics. Extensive simulations demonstrate the effectiveness of these schemes in systems with reduced pilot overhead. The performance gains can be exploited, e.g., to deploy systems with fewer pilots or feedback bits.

Acknowledgements

First and foremost, I would like to thank Prof. Dr.-Ing. Wolfgang Utschick for giving me the opportunity to work at the Chair of Signal Processing at the Technical University of Munich (TUM). I am sincerely grateful for your continuous support and guidance throughout this journey. I would also like to express my heartfelt thanks to Dr.-Ing. Michael Joham for his continuous support and collaboration.

I am grateful to Prof. Dr.-Ing. Stephan ten Brink for reviewing my dissertation and to Prof. Dr.-Ing. Antonia Wachter-Zeh for chairing my doctoral examination.

I would like to express my heartfelt gratitude to my colleagues and friends at the Chair of Signal Processing, TUM, for sharing this journey with me. I am especially thankful to Dr.-Ing. Andreas Barthelme, Michael Baur, Dr.-Ing. Donia Ben Amor, Benedikt Böck, Dr.-Ing. Benedikt Fesl, Dr.-Ing. Oliver De Candido, Amar Kasibovic, Dr.-Ing. Michael Koller, Dr.-Ing. Valentina Rizzello, Sadaf Syed, Dominik Semmler, Julia Sistermanns, Florian Strasser, Franz Weißer, and Michael Würth for the collaborative and supportive atmosphere we built together. Furthermore, I would like to thank all the students I had the pleasure of supervising.

Finally, I would like to thank my family and friends for their continuous support throughout this journey. This work is dedicated to you.

Contents

Acronyms

3GPP	3rd Generation Partnership Project
AE	autoencoder
AWGN	additive white Gaussian noise
BD	block diagonalization
BS	base station
cCDF	complementary cumulative distribution function
CME	conditional mean estimator
CMI	conditional mutual information
CSI	channel state information
DFT	discrete Fourier transform
DL	downlink
DNN	deep neural network
EM	expectation maximization
FDD	frequency division duplex
GMM	Gaussian mixture model
GNN	graph neural network
KKT	Karush-Kuhn-Tucker
LMMSE	linear minimum mean square error
LS	least squares

MAP	maximum a posteriori
MIMO	multiple-input multiple-output
MISO	multiple-input single-output
ML	machine learning
MSE	mean squared error
MT	mobile terminal
MU-MIMO	multi-user MIMO
MU-MISO	multi-user MISO
NMSE	normalized MSE
NSE	normalized spectral efficiency
OMP	orthogonal matching pursuit
PDF	probability density function
PGA	projected gradient ascent
RBD	regularized block diagonalization
RCI	regularized channel inversion
RMP	range matching pursuit
SNR	signal-to-noise ratio
SR	sum-rate
SVD	singular value decomposition
SWMMSE	stochastic WMMSE
UL	uplink
ULA	uniform linear array
UMa	urban macrocell
UMi	urban microcell
URA	uniform rectangular array
VAE	variational autoencoder
VQ-VAE	vector quantized-variational autoencoder
WMMSE	weighted minimum mean squared error

1 Introduction

In multiple-input multiple-output (MIMO) communication systems, acquiring channel state information (CSI) at the base station (BS) at regular time intervals is crucial for efficient operation. In frequency division duplex (FDD) systems, the BS and mobile terminal (MT) transmit simultaneously but on different frequencies, breaking the reciprocity between instantaneous uplink (UL) and downlink (DL) CSI. This makes DL CSI acquisition at the BS particularly challenging [1]. Solutions include transferring the DL CSI estimated at the MT or corresponding pilot observations received at the MTs to the BS, either directly or in a compressed form [2–4]. Alternatively, limited feedback systems, which utilize a relatively small number of feedback bits to quantize the DL CSI, are commonly employed [5–7]. These approaches emphasize two fundamental tasks considered in this dissertation: designing precoders based on feedback from multiple MTs and optimizing downlink pilot matrices to improve DL channel estimation.

1.1 Precoder Design in FDD Systems

In conventional methods, the DL CSI is estimated at the MT, and then the feedback information is determined. This feedback information can either represent an index for selecting a precoder from a predefined codebook or provide quantized information (channel directions) about the DL CSI [5–7]. However, in massive MIMO systems which typically feature many antenna elements at the BS, the pilot overhead required to fully illuminate the DL channel becomes prohibitively high [8]. Therefore, there is significant interest in developing algorithms that can infer feedback in low pilot overhead systems, where the number of pilots is fewer than the number of transmit antennas.

Recent works [9–15] have introduced various end-to-end deep neural network (DNN)-based techniques that process pilot observations through neural network modules to infer feedback information. These approaches include single-user MIMO systems for single-stream transmission [9], multi-stream transmissions supporting various signal-to-noise ratio (SNR) levels [10], and multi-user systems for single-

antenna MTs [11, 12]. Extensions to multi-user MIMO (MU-MIMO) systems with multiple antennas have also been explored [14, 15]. While these methods outperform conventional techniques in specific settings, they face practical challenges. In particular, the DNN-based approaches are inflexible, as they only support single- or multi-user transmission modes, requiring distinct task-specific DNNs for each mode. This includes the adaptation to different SNR levels, number of pilots, number of feedback bits, and number of MTs, which usually requires additional and specifically trained DNNs. Furthermore, scalability is a significant concern, as the number of DNN parameters increases quadratically with the product of transmit and receive antennas, leading to convergence issues during training. The associated model transfer overhead, i.e., communicating DNN parameters for different system configurations from the BS to the MTs, results in an impractical signaling burden.

Generally, machine learning (ML) techniques typically require a representative dataset of channels from the BS communication environment for training. In FDD mode, the MTs would need to collect DL CSI data and either perform training locally or share the collected data with the BS. This results in significant computation and signaling overhead. However, recent work in [16] showed that DL CSI training data can be replaced with UL CSI data in FDD systems with commonly employed frequency gaps, even for designing DL functionalities. This approach eliminates the overhead and significantly improves the practicality of ML-based methods. Its effectiveness has been further validated across various DL functionalities [10, 17–20], including our work on codebook construction in [10, 20]. Building on this, the UL CSI can be gathered at the BS during regular UL transmissions. In this dissertation, we adopt the approach of centrally learning DL-related functionalities at the BS using UL training data, necessitating model transfer to the MTs. It is worth noting that model transfer is currently being explored in a recent 3rd Generation Partnership Project (3GPP) technical report [21].

In recent years, generative models have attracted considerable attention in wireless communications. These models aim to learn the underlying probability density function (PDF) of a wireless communication environment, given a training dataset and provide a generative prior for various applications. Generative techniques such as generative adversarial networks [22], variational autoencoders (VAEs) [23], and diffusion models [24] have gained significant interest. In wireless communications, generative models have been applied to channel estimation [25–27], precoding [12], and as channel modeling frameworks, e.g., [28, 29]. In this dissertation, the main focus is on Gaussian mixture models (GMMs), which are widely used in wireless communications, e.g., for channel state prediction [30], multi-path clustering [31],

and pilot optimization [32]. More recently, we adopted GMMs for channel estimation in [33] and for channel prediction in scenarios with high mobility in [34]. The universal approximation capability of GMMs, see [35], underpins their strong performance in these applications.

In this dissertation, we introduce a versatile precoder design scheme for point-to-point MIMO and MU-MIMO FDD systems. In particular, we propose an innovative use of GMMs for feedback inference and codebook design, leveraging GMMs centrally fitted at the BS using UL training data. The GMM components are used for clustering channel data, allowing the design of environment-specific codebooks for single-user systems. For multi-user systems, we propose two approaches: one utilizes directional information for joint precoder design, while the other exploits the GMM's sample generation ability combined with a state-of-the-art stochastic iterative algorithm for precoder design. A major advantage of the proposed GMM-based feedback scheme is its versatility, enabling adaptation to different SNR levels, pilot configurations, numbers of served MTs, and transmission modes without requiring retraining. Furthermore, the online computational complexity is independent of the number of transmit antennas, and parallelization is supported. Additionally, two methods to support variable feedback bit lengths via the GMM are proposed, enabling flexibility in the feedback process, reducing the model-transfer overhead, and further enhancing the versatility. The GMM-based feedback scheme also benefits from model-based insights, which further significantly reduce the model transfer overhead.

To alleviate the scaling issues of DNNs, graph neural network (GNN)-based approaches have recently been employed for sum-rate maximization with perfect CSI in [36,37] and energy efficiency optimization under perfect statistical knowledge in [38]. Building on [37], the work in [39] extends GNN architectures to systems with imperfect CSI. These approaches exploit the property that reordering MT indices permutes the precoders without impacting the sum-rate, effectively acting as regularization [36]. A significant advantage of GNN-based techniques is their scalability to varying numbers of MTs without requiring retraining. We propose a GNN-based precoder design framework for massive MIMO systems utilizing statistical channel knowledge, ensuring efficient, compact network architectures that only scale linearly with the number of BS antennas. By integrating the proposed GMM-based feedback approach, the framework enables a parallelizable scheme for enhanced precoder design.

For systems with larger feedback bit budgets ranging from tens to hundreds of bits, this dissertation employs vector quantized-variational autoencoders (VQ-VAEs). Originally developed for image processing [40], VQ-VAEs have also been widely adopted in wireless communications for tasks such as channel compression in [41, 42]

and channel prediction/interpolation [43]. VQ-VAEs are well-suited for feedback applications due to their discrete latent space, enabled by an embedding, a codebook of finite size, that quantizes the latent space, compactly representing the feedback information. We propose to use VQ-VAEs for feedback inference and precoder design, combined with a state-of-the-art stochastic iterative algorithm for massive MIMO systems with low pilot overhead. Thereby, we propose adapting the loss function and the network architecture based on model-based insights for enhanced training.

1.2 Pilot Design in FDD Systems

The quality of channel estimation heavily depends on the choice of pilots, highlighting the need for advanced pilot design strategies in FDD systems. In massive MIMO systems with a large number of BS antennas, conventional approaches require transmitting a number of pilots equal to the number of antennas to fully illuminate the channel and avoid systematic errors with least squares (LS) estimation at the MT. As highlighted earlier, this leads to substantial pilot overhead, which can become prohibitive [8]. In environments with spatial correlation at the BS and MTs, the DL training overhead can be significantly reduced by exploiting the statistical properties of the channel and noise [32, 44–50], e.g., by using Bayesian estimation approaches combined with designed pilot matrices based on specific optimization criteria.

In particular, the single-user case is examined in [44–47], with [44] focusing on a multiple-input single-output (MISO) system, while [45–47] investigate MIMO setups utilizing a Kronecker correlation model, cf. [51]. These studies use the minimization of the mean squared error (MSE) as the optimization criterion for DL channel estimation. A common strategy involves transmitting the eigenvectors of the transmit-side covariance matrix associated with the strongest eigenvalues as pilots. Additionally, adaptive scaling of these pilot vectors can be achieved by employing a water-filling procedure, though unitary pilots (equal power allocation across pilot vectors) are often used for simplicity [44]. For multi-user MISO (MU-MISO) systems, works such as [32, 48–50] analyze different optimization approaches. The work [48] focuses on minimizing the sum of MSEs for all MTs, while [50] minimizes a weighted sum MSE. Leveraging the relationship between the mutual information and MSE [52], the work [49] aims to maximize the sum conditional mutual information (CMI) to design the pilot matrix. In [32], the channel distribution is assumed to be a Gaussian mixture distribution and mutual information maximization is adopted. The studies [32, 49, 50] apply iterative algorithms for pilot matrix design. However, all these works rely on perfect or estimated statistical knowledge at the BS and/or the

MTs, which can be challenging to obtain in practice. Nonetheless, assuming perfect statistical knowledge for pilot design offers valuable insights and serves as a foundation for algorithm development, even when such knowledge is unavailable, as demonstrated in this dissertation.

Recent works in [53, 54] propose ML-based approaches that combine pilot matrix design with DL channel estimation. Specifically, end-to-end trained DNNs are used to jointly learn a pilot matrix and a channel estimation module. This approach optimizes the pilot design for the entire BS environment, meaning a global pilot matrix is learned offline and remains fixed during the online channel estimation phase. However, these end-to-end trained DNNs cannot provide MT-adaptive pilot matrices, whether for single-user or multi-user scenarios, as they cannot leverage MT-specific statistics that could be inferred in the online phase. Additionally, end-to-end DNNs are inflexible with respect to the number of pilots, as varying pilot counts require different pilot matrix dimensions and dedicated channel estimation modules, necessitating multiple end-to-end DNNs. The requirement for SNR-specific training may further hinder the practicality of these solutions.

In this dissertation, we leverage GMMs to propose a versatile pilot design scheme for point-to-point MIMO and MU-MIMO FDD systems. In single-user systems, the proposed GMM-based pilot design scheme eliminates the need for perfect knowledge of the channel's statistics at the BS or the MT. Instead, the statistical prior information captured by the GMM during an initial offline phase is used to determine a feedback index at the MT, encoded using a few feedback bits. This feedback index is sufficient to establish common knowledge of the pilot matrix in single-user systems. A significant advantage of this approach is the ability to pre-construct a set of pilot matrices, which removes the need for online pilot optimization. We further extend the sum CMI-based multi-user pilot optimization framework from [49], which was originally designed for single-antenna MTs, to MU-MIMO systems, where both the BS and MTs are equipped with multiple antennas. We further establish a lower bound on the sum CMI in MU-MIMO systems and identify conditions under which the bound is tight or exhibits the largest gap. This lower bound serves as a proxy objective, providing complexity savings in the online pilot matrix optimization. Building on this, we utilize the GMM-based scheme for pilot matrix design in MU-MIMO systems. Each MT sends its feedback information to the BS, which then broadcasts the feedback indices to all MTs via a feedforward link to enable common pilot knowledge. The proposed approach improves upon the state-of-the-art signaling scheme from [50] that quantizes the pilot matrix after optimization.

1.3 Outline and Contributions

The remainder of this dissertation is organized as follows:

- Chapter 2 presents the system models under consideration, which are divided into a pilot transmission phase and a data transmission phase. We then provide a brief discussion of the channel data sources that are used to evaluate the proposed methods. The chapter concludes with a brief overview of GMMs and VQ-VAEs, introducing the generative concepts with discrete latent spaces employed in this dissertation.
- Chapter 3 focuses on the precoder design task in point-to-point and multi-user MIMO FDD systems, see Section 1.1. A detailed outline of this chapter is provided at the beginning. The contributions presented in this chapter have been published in [55–60].
- Chapter 4 addresses the pilot matrix design in point-to-point and multi-user MIMO FDD systems, see Section 1.2. The chapter begins with a detailed outline. The contributions discussed have been published in [61, 62].
- Chapter 5 concludes this dissertation with a summary of the findings and an outlook on potential future research directions.

While not central to the primary contributions of this dissertation, several works published during the course of this dissertation explore various applications that utilize similar concepts and methodologies, as outlined below. These include [10, 20], which demonstrated the effectiveness of utilizing UL data for codebook construction, forming a basis for this work. Additionally, ML-based techniques, with a focus on generative models, were investigated for various purposes: channel estimation in high-resolution systems [18, 33, 63–70], reconfigurable intelligent surface-aided systems [71], systems that incorporate payload data for enhanced estimation [72, 73], and systems leveraging direction of arrival information [74]. Channel estimation in low-resolution systems was analyzed in [75, 76] and channel prediction was explored in [34, 77]. Lastly, [78] derived regularization techniques for generative models tailored to wireless communications, and [79] introduced evaluation metrics and methods for generative models in wireless communication applications.

1.4 Notation

Matrices and vectors are represented using bold uppercase and bold lowercase letters, respectively. The transpose, conjugate, and conjugate transpose of a matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^{\mathrm{T}}$, $\boldsymbol{A}^{*}$, and $\boldsymbol{A}^{\mathrm{H}}$, respectively. The all-zeros vector and the identity matrix of appropriate dimensions are represented by $\mathbf{0}$ or $\mathbf{I}$, respectively. The canonical unit vector, with a one at the i-th position and zero elsewhere, is denoted by $\mathbf{e}_i$. The Euclidean norm of a vector $\boldsymbol{a} \in \mathbb{C}^N$ is written as $\|\boldsymbol{a}\|$. The cardinality of a set $\mathcal{V}$ is given by $|\mathcal{V}|$. A complex-valued normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$ is expressed as $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{C})$ and $\sim$ represents "distributed as." The determinant, the rank, and the trace of matrix $\boldsymbol{A}$ are denoted by $\det(\boldsymbol{A})$, $\mathrm{rk}(\boldsymbol{A})$, and $\mathrm{tr}(\boldsymbol{A})$, respectively. The vectorization (column-wise stacking) of a matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ is denoted by $\boldsymbol{a} = \mathrm{vec}(\boldsymbol{A}) \in \mathbb{C}^{mN}$, and the reverse operation is given by $\boldsymbol{A} = \mathrm{unvec}_{m,N}(\boldsymbol{a})$. The vector containing the diagonal elements of a diagonal matrix $\boldsymbol{A}$ is given as $\mathrm{diag}(\boldsymbol{A})$, and the diagonal matrix with vector $\boldsymbol{a}$ as its diagonal is denoted by $\mathrm{diag}(\boldsymbol{a})$. The Kronecker product of two matrices $\boldsymbol{A} \in \mathbb{C}^{m_1 \times N_1}$ and $\boldsymbol{B} \in \mathbb{C}^{m_2 \times N_2}$ is denoted by $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{C}^{m_1 m_2 \times N_1 N_2}$.

2 Preliminaries

2.1 System Model

2.1.1 Pilot Transmission Phase

To enable feedback inference at the MTs, the BS broadcasts pilot signals to all MTs. The system comprises a BS with N_{tx} antenna elements and J MTs, where each MT $j \in \mathcal{J} = \{1, 2, \ldots, J\}$ is equipped with N_{rx} antennas. A block-fading model, cf., e.g., [80], is assumed, where the DL signal received at MT $j \in \mathcal{J}$ for block t is

$$\boldsymbol{Y}_{j,t} = \boldsymbol{H}_{j,t}\boldsymbol{P}_t^{\mathrm{T}} + \boldsymbol{N}_{j,t}, \tag{2.1}$$

where $t = 0, \ldots, T$, $\boldsymbol{H}_{j,t} \in \mathbb{C}^{N_{\mathrm{rx}} \times N_{\mathrm{tx}}}$ is the MIMO channel matrix, $\boldsymbol{P}_t \in \mathbb{C}^{n_{\mathrm{p}} \times N_{\mathrm{tx}}}$ is the pilot matrix, and $\boldsymbol{N}_{j,t} = [\boldsymbol{n}'_{j,t,1}, \ldots, \boldsymbol{n}'_{j,t,n_{\mathrm{p}}}] \in \mathbb{C}^{N_{\mathrm{rx}} \times n_{\mathrm{p}}}$ is the additive white Gaussian noise (AWGN) with $\boldsymbol{n}'_{j,t,p} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \sigma_n^2 \mathbf{I}_{N_{\mathrm{rx}}})$ for all $p \in \{1, 2, \ldots, n_{\mathrm{p}}\}$, and n_{p} is the number of pilots. In this work, we focus on system setups where the number of pilots is fewer than the number of transmit antennas, i.e., $n_{\mathrm{p}} < N_{\mathrm{tx}}$, as such configurations are of crucial importance for substantially reducing the DL training overhead [8]. For the subsequent analysis, it is convenient to vectorize (2.1):

$$\boldsymbol{y}_{j,t} = (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})\boldsymbol{h}_{j,t} + \boldsymbol{n}_{j,t} \tag{2.2}$$

where $\boldsymbol{h}_{j,t} = \mathrm{vec}(\boldsymbol{H}_{j,t})$, $\boldsymbol{y}_{j,t} = \mathrm{vec}(\boldsymbol{Y}_{j,t})$, $\boldsymbol{n}_{j,t} = \mathrm{vec}(\boldsymbol{N}_{j,t})$, and $\boldsymbol{n}_{j,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \sigma_n^2 \mathbf{I}_{N_{\mathrm{rx}} n_{\mathrm{p}}}$.

In Chapter 3, we focus on precoder design for the data transmission phase that follows the pilot transmission phase of a specific block. Thus, we omit the block index t in the descriptions of the data transmission phase (Sections 2.1.2.1 and 2.1.2.2) and throughout Chapter 3 for notational simplicity. In Chapter 3, discrete Fourier transform (DFT)-based submatrices tailored to the BS antenna geometry, cf., [81], are used as the pilot matrix, which is assumed to be commonly known at both the BS and the MTs. In contrast, Chapter 4 addresses the design of pilot matrices for blocks $t > 0$ to enhance DL channel estimation performance by utilizing feedback information from the previous block $t - 1$.

2.1.2 Data Transmission Phase

2.1.2.1 Point-to-Point MIMO System

In the point-to-point MIMO system, the MT index j is omitted for notational convenience. The DL received signal is given by:

$$\boldsymbol{y}' = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}', \tag{2.3}$$

where $\boldsymbol{y}' \in \mathbb{C}^{N_{\mathrm{rx}}}$ is the receive vector, $\boldsymbol{x} \in \mathbb{C}^{N_{\mathrm{tx}}}$ is the precoded data vector sent over the channel $\boldsymbol{H} \in \mathbb{C}^{N_{\mathrm{rx}} \times N_{\mathrm{tx}}}$, and $\boldsymbol{n}' \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_{\mathrm{rx}}})$ denotes the AWGN. We consider system configurations with $N_{\mathrm{rx}} < N_{\mathrm{tx}}$. When perfect CSI is available at both the transmitter and receiver and assuming the transmit data follows a zero-mean Gaussian distribution, the channel capacity is given by [82, p. 326]

$$C = \max_{\boldsymbol{Q} \succeq \mathbf{0}, \mathrm{tr}(\boldsymbol{Q}) \leq \rho} \log_2 \det \left(\mathbf{I} + \frac{1}{\sigma_n^2} \boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}} \right), \tag{2.4}$$

where $\boldsymbol{Q} \in \mathbb{C}^{N_{\mathrm{tx}} \times N_{\mathrm{tx}}}$ is the transmit covariance matrix, ρ is the transmit power, and the transmit vector is $\boldsymbol{x} = \boldsymbol{Q}^{1/2}\boldsymbol{s}$ with $\mathrm{E}[\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}] = \mathbf{I}_{N_{\mathrm{tx}}}$ [5]. The capacity-achieving transmit covariance matrix $\boldsymbol{Q}^{\star}$ is determined by decomposing the channel into N_{rx} parallel streams and applying the water-filling algorithm [83]. Since channel reciprocity does not hold in FDD systems, the optimal transmit covariance matrix $\boldsymbol{Q}^{\star}$ can only be computed if the DL CSI is known perfectly. Consequently, some form of feedback from the MT to the BS is required. Ideally, the MT would transmit the complete DL CSI to the BS, but this is infeasible in general. Instead, a small number of B bits is sent back to the BS. These B feedback bits typically encode an index $k^{\star} \in \{1, 2, \cdots, 2^B\}$ that identifies an element from a set of 2^B pre-computed transmit covariance matrices $\mathcal{Q} = \{\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_{2^B}\}$. The BS then uses the transmit covariance matrix $\boldsymbol{Q}_{k^{\star}}$ for data transmission [5].

2.1.2.2 Multi-user MIMO System

We consider linear precoding for the DL in a single-cell MU-MIMO system. The transmit signal vector for MT j is denoted as $\boldsymbol{s}_j \in \mathbb{C}^{d_j}$, where d_j represents the number of data streams. We assume that $\mathrm{E}[\boldsymbol{s}_j] = \mathbf{0}$ and $\mathrm{E}[\boldsymbol{s}_j \boldsymbol{s}_j^{\mathrm{H}}] = \mathbf{I}_{d_j}$, with the signals intended for different MTs being mutually independent. The overall precoded DL data vector is expressed as $\boldsymbol{x} = \sum_{j=1}^{J} \boldsymbol{M}_j \boldsymbol{s}_j$ where $\boldsymbol{M}_j \in \mathbb{C}^{N_{\mathrm{tx}} \times d_j}$ is the precoding matrix applied at the BS to process the transmit signal of MT j. Unless stated otherwise, we set $d_j = N_{\mathrm{rx}}, \forall j$ without loss of generality. The precoders are normalized to satisfy

the transmit power constraint $\mathrm{tr}(\sum_{j=1}^{J} \boldsymbol{M}_j \boldsymbol{M}_j^{\mathrm{H}}) = \rho$. Thus, the signal received at MT j is given by (see, e.g., [84]):

$$\boldsymbol{y}_j' = \boldsymbol{H}_j \boldsymbol{M}_j \boldsymbol{s}_j + \sum_{n=1, n\neq j}^{J} \boldsymbol{H}_j \boldsymbol{M}_n \boldsymbol{s}_n + \boldsymbol{n}_j', \ \forall j \in \mathcal{J} \tag{2.5}$$

where $\boldsymbol{H}_j \in \mathbb{C}^{N_{\mathrm{rx}} \times N_{\mathrm{tx}}}$ is the MIMO channel from the BS to MT j and $\boldsymbol{n}_j' \in \mathbb{C}^{N_{\mathrm{rx}}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_{\mathrm{rx}}})$ denotes the AWGN at MT j. The instantaneous rate of MT j is then expressed as (see, e.g., [84]):

$$R_j^{\mathrm{inst}} \ = \ \log_2 \det \Big(\mathbf{I} + \boldsymbol{H}_j \boldsymbol{M}_j \boldsymbol{M}_j^{\mathrm{H}} \boldsymbol{H}_j^{\mathrm{H}} \big(\sum_{n \neq j} \boldsymbol{H}_j \boldsymbol{M}_n \boldsymbol{M}_n^{\mathrm{H}} \boldsymbol{H}_j^{\mathrm{H}} + \sigma_n^2 \mathbf{I} \big)^{-1} \Big). \tag{2.6}$$

If the BS had perfect knowledge of the DL CSI for all MTs, it could employ various methods to jointly optimize the precoders $\boldsymbol{M}_j$ for all MTs $j \in \mathcal{J}$. These methods include non-iterative approaches such as block diagonalization (BD) [85], regularized block diagonalization (RBD) [86], and regularized channel inversion (RCI) [87, 88], as well as iterative techniques like the iterative weighted minimum mean squared error (WMMSE) algorithm [84, 89, 90]. In the FDD systems considered in this work, however, each MT j encodes feedback information using only B feedback bits, cf. [7]. This feedback information about the DL CSI is transmitted to the BS, which then designs the precoders.

2.2 Channel Data

For a comprehensive evaluation of the discussed methods, different datasets are considered. Each training dataset, comprising M representative channels (see Sections 2.2.1, 2.2.2, and 2.2.3), is defined as

$$\mathcal{H} = \{\boldsymbol{h}^{(m)}\}_{m=1}^{M}, \tag{2.7}$$

and characterizes a wireless communication environment with an unknown PDF $p_{\boldsymbol{h}}$. Any channel $\boldsymbol{h}$ within the BS's coverage area can be viewed as a realization of a random variable following this PDF $p_{\boldsymbol{h}}$, though this PDF is not analytically available.

2.2.1 Spatial Channel Model

We adopt the spatial channel model outlined in [80, 91], where the channels of each MT j at each block t are modeled conditionally Gaussian, i.e., $\boldsymbol{h}_{j,t} | \boldsymbol{\delta}_j \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\delta}_j})$.

It is assumed that the covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}_j}$ of each MT remains constant over $T+1$ blocks. The random vectors $\boldsymbol{\delta}_j$ contain the main angles of arrival and departure of the multipath propagation clusters between the BS and the MTs. These angles are drawn independently and are uniformly distributed within the range $[\frac{-\pi}{2}, \frac{\pi}{2}]$. The BS and the MTs are equipped with uniform linear arrays (ULAs) such that the transmit- and receive-side spatial channel covariance matrices are expressed as

$$\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\{\mathrm{rx,tx}\}} = \int_{-\pi}^{\pi} g_{\mathrm{p}}^{\{\mathrm{rx,tx}\}}(\theta; \boldsymbol{\delta}_j) \boldsymbol{\omega}^{\{\mathrm{rx,tx}\}}(\theta) \boldsymbol{\omega}^{\{\mathrm{rx,tx}\},\mathrm{H}}(\theta) \mathrm{d}\theta, \tag{2.8}$$

where $\boldsymbol{\omega}^{\{\mathrm{rx,tx}\}}(\theta) = [1, \mathrm{e}^{\mathrm{j}\pi\sin(\theta)}, \ldots, \mathrm{e}^{\mathrm{j}\pi(N_{\{\mathrm{rx,tx}\}}-1)\sin(\theta)}]^{\mathrm{T}}$ denotes the array steering vector for an angle of arrival/departure θ. The function $g_{\mathrm{p}}^{\{\mathrm{rx,tx}\}}$ represents a Laplacian power density, with its standard deviation describing the angular spread $\sigma_{\mathrm{AS}}^{\{\mathrm{rx,tx}\}}$ of the multipath propagation cluster at the BS ($\sigma_{\mathrm{AS}}^{\mathrm{tx}} = 2^\circ$) and MT j ($\sigma_{\mathrm{AS}}^{\mathrm{rx}} = 35^\circ$) side [91]. Assuming independent scattering in the vicinity of both the transmitter and receiver, the overall channel covariance matrix for MT j is expressed as $\boldsymbol{C}_{\boldsymbol{\delta}_j} = \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \otimes \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}$ (see, e.g., [51]). For single antenna MTs, $\boldsymbol{C}_{\boldsymbol{\delta}_j}$ simplifies to the transmit-side covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}$. Due to the low angular spread at the BS, the covariance matrices are characterized by low numerical rank. This channel model has also been utilized similarly in studies such as [49,50], where $\boldsymbol{\delta}_j$ was assumed to be given for all $j \in \mathcal{J}$. In simulations for the evaluation of genie-aided benchmarks, perfect statistical knowledge is assumed to be available.

Using the procedure for modeling the channel $\boldsymbol{h}_{j,t}$ of a specific MT j for block t, the training dataset $\mathcal{H}$, see (2.7), is constructed from randomly drawn channels $\boldsymbol{h}^{(m)}$. In particular, for each sample $\boldsymbol{h}^{(m)}$, with $m = 1, \ldots, M$, random angles are first generated and collected in $\boldsymbol{\delta}^{(m)}$, after which the sample is drawn as $\boldsymbol{h}^{(m)} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\delta}^{(m)}})$. In this way, the dataset $\mathcal{H}$ captures variations in the MT locations within the BS's coverage area.

2.2.2 QuaDRiGa Channel Generator

The QuaDRiGa channel simulator [92, 93] is utilized to generate channels in an urban macrocell (UMa) scenario. With this simulator, UL and DL domain channels can be generated, which enables learning DL functionalities solely from UL training data, following [16], see Section 1.1. The employed carrier frequency for the UL and DL domain is $2.53\,\mathrm{GHz}$ and $2.73\,\mathrm{GHz}$, respectively, corresponding to a typical frequency gap of $200\,\mathrm{MHz}$. The BS is equipped with a uniform rectangular array (URA) with "3GPP-3D" antennas, while the MTs are equipped with ULAs with "omni-directional" antennas. The BS is placed at a height of $25\,\mathrm{m}$ and covers a 120° sector. The distance

between the MTs and the BS ranges from $35\,\mathrm{m}$ to $500\,\mathrm{m}$. Approximately 80% of the MTs are located indoors across various floor levels, while outdoor MTs are positioned at a height of $1.5\,\mathrm{m}$. The QuaDRiGa channel simulator models MIMO channels as

$$\boldsymbol{H} = \sum_{\ell=1}^{L_{\mathrm{M}}} \boldsymbol{G}_\ell e^{-2\pi j f_{\mathrm{c}} \tau_\ell} \tag{2.9}$$

where ℓ denotes the path index, L_{M} represents the number of multi-path components, f_{c} is the carrier frequency, and τ_ℓ is the delay associated with the ℓ-th path. The value of L_{M} varies depending on the propagation type, i.e., whether there is a line of sight, non-line of sight, or outdoor-to-indoor condition. The coefficients matrix $\boldsymbol{G}_\ell$ contains one complex entry for each antenna pair and accounts for path attenuation, antenna radiation pattern weighting, and polarization. The generated channels are post-processed to remove the path gain, as described in the manual [93].

Using the QuaDRiGa channel simulator, the training dataset $\mathcal{H}$, see (2.7), comprising M channels (DL or UL domain) characterizing the UMa scenario is constructed by generating samples $\boldsymbol{H}^{(m)}$, for $m = 1, \ldots, M$, following (2.9) and subsequently vectorizing each as $\boldsymbol{h}^{(m)} = \mathrm{vec}\,(\boldsymbol{H}^{(m)})$.

2.2.3 Measured Channel Data

Lastly, we utilize real-world measurement data from a massive MIMO urban microcell (UMi) scenario, collected during a measurement campaign conducted at the Nokia campus in Stuttgart, Germany, in October/November 2017 [94]. The BS antenna was placed at a rooftop at a height of $20\,\mathrm{m}$ with a down-tilt of 10°. The BS comprised a URA with $N = 64$ single-polarized patch antennas, arranged with $N_{\mathrm{v}} = 4$ vertical elements with a spacing of $\lambda/2$ and $N_{\mathrm{h}} = 16$ horizontal elements with a spacing of λ, where λ denotes the wavelength. The carrier frequency used was $f_{\mathrm{c}} = 2.18\,\mathrm{GHz}$. The receive antenna, a single monopole representing the MTs, was installed on a vehicle at a height of $1.5\,\mathrm{m}$. During the campaign, the vehicle traveled at a maximum velocity of $25\,\mathrm{km/h}$. Further details about the measurement campaign can be found in [94].

In this case, the training dataset $\mathcal{H}$, see (2.7), consists of M measured MISO channels $\boldsymbol{h}^{(m)}$, with $m = 1, \ldots, M$.

2.3 Generative Models with a Discrete Latent Space

To capture the underlying PDF $p_{\boldsymbol{h}}$ characterizing the channels $\boldsymbol{h}$ within the wireless communication environment, generative models that comprise a discrete latent space

are utilized. In particular, we consider GMMs and VQ-VAEs, which are briefly outlined below.

2.3.1 Gaussian Mixture Model

GMMs assume that all channels $\boldsymbol{h}$ within the environment stem from a mixture of K Gaussian components [95, Sec. 2.3.9], expressed as:

$$p_{\text{GMM}}(\boldsymbol{h}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}; \boldsymbol{\mu}_k, \boldsymbol{C}_k), \tag{2.10}$$

where each component is parametrized by a mixing coefficient π_k, a mean vector $\boldsymbol{\mu}_k$, and a covariance matrix $\boldsymbol{C}_k$. Using the training dataset $\mathcal{H}$, see (2.7), maximum likelihood estimates of the GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{C}_k\}_{k=1}^{K}$ can be computed with an expectation maximization (EM) algorithm, see [95, Sec. 9.2.2]. This process comprises four main steps, summarized below:

- Initialization: The parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{C}_k\}_{k=1}^{K}$ are initialized by, e.g., the K-means algorithm, and the initial value of the log-likelihood is calculated.
- E-step: The posterior probabilities for each component k given the current parameters are computed:

$$p(k \mid \boldsymbol{h}^{(m)}) = \frac{\pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}^{(m)}; \boldsymbol{\mu}_k, \boldsymbol{C}_k)}{\sum_{i=1}^{K} p(i) \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}^{(m)}; \boldsymbol{\mu}_i, \boldsymbol{C}_i)}. \tag{2.11}$$

- M-step: The parameters are updated using the posterior probabilities:

$$\boldsymbol{\mu}_k = \frac{1}{M_k} \sum_{m=1}^{M} p(k \mid \boldsymbol{h}^{(m)}) \boldsymbol{h}^{(m)}, \tag{2.12}$$

$$\boldsymbol{C}_k = \frac{1}{M_k} \sum_{m=1}^{M} p(k \mid \boldsymbol{h}^{(m)}) (\boldsymbol{h}^{(m)} - \boldsymbol{\mu}_k)(\boldsymbol{h}^{(m)} - \boldsymbol{\mu}_k)^{\text{H}}, \tag{2.13}$$

$$\pi_k = \frac{M_k}{M}, \tag{2.14}$$

  where $M_k = \sum_{m=1}^{M} p(k \mid \boldsymbol{h}^{(m)})$.

- Log-likelihood evaluation: The log-likelihood with the updated parameters is evaluated as

$$\mathcal{L}_{\text{GMM}} = \sum_{m=1}^{M} \log \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}^{(m)}; \boldsymbol{\mu}_k, \boldsymbol{C}_k). \tag{2.15}$$

The E-step and the M-step are repeated until the log-likelihood converges.

After a GMM is fitted, the probability that a given channel $\boldsymbol{h}$ stems from one of the components can be computed via [95, Sec. 9.2]:

$$p(k \mid \boldsymbol{h}) = \frac{\pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}; \boldsymbol{\mu}_k, \boldsymbol{C}_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}_{\mathbb{C}}(\boldsymbol{h}; \boldsymbol{\mu}_i, \boldsymbol{C}_i)}. \tag{2.16}$$

These posterior probabilities are commonly referred to as *responsibilities*.

2.3.2 Vector Quantized-Variational Autoencoder

Similar to VAEs [23], VQ-VAEs consist of an encoder, which maps the input to a latent vector, and a decoder, both implemented using neural networks. However, unlike VAEs, VQ-VAEs comprise a discrete latent space [40]. The overall distribution parametrized by the VQ-VAE can be expressed as [40]:

$$p_{\text{VQ-VAE}}(\boldsymbol{h}) = \sum_{k=1}^{K} p(\boldsymbol{h} \mid \boldsymbol{f}_k) p(\boldsymbol{f}_k), \tag{2.17}$$

where the vectors $\boldsymbol{f}_k$ denote the discrete latent vectors obtained after quantizing the encoder output $\boldsymbol{z}_e$. The conditional decoder distributions $p(\boldsymbol{h}_k \mid \boldsymbol{f}_k)$ can be modeled to be Gaussian, i.e., $p(\boldsymbol{h}_k \mid \boldsymbol{f}_k) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_k, \boldsymbol{C}_k)$, wile the priors $p(\boldsymbol{f}_k)$ are kept uniform during training [40].

An embedding $\mathcal{E}$, consisting of C elements, i.e., $\mathcal{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_C\}$, is used to map the unquantized encoder output $\boldsymbol{z}_e$ to the quantized latent vector $\boldsymbol{f}_k$ through the following process. Let each embedding vector $\boldsymbol{e}_c \in \mathcal{E}$ have a dimension of N_{E}, while the latent space dimension is N_{L}. The unquantized latent vector $\boldsymbol{z}_e$ can then be divided into $\frac{N_{\mathrm{L}}}{N_{\mathrm{E}}}$ sub-vectors of dimension N_{E}, i.e., $\boldsymbol{z}_e = [\boldsymbol{z}_{e,1}, \ldots, \boldsymbol{z}_{e,N_{\mathrm{L}}/N_{\mathrm{E}}}]^{\mathrm{T}}$. The i-th sub-vector $\boldsymbol{f}_{k,i}$ of the quantized latent vector $\boldsymbol{f}_k$ is determined as follows [40]:

$$\boldsymbol{f}_{k,i} = \underset{\boldsymbol{e}_c \in \mathcal{E}}{\arg\min} \, \|\boldsymbol{e}_c - \boldsymbol{z}_{e,i}\|_2, \tag{2.18}$$

where $\boldsymbol{z}_{e,i}$ is the i-th sub-vector of the unquantized latent vector $\boldsymbol{z}_e$, and $i = 1, \ldots, \frac{N_{\mathrm{L}}}{N_{\mathrm{E}}}$. Thus, the quantized latent vector $\boldsymbol{f}_k$ can be fully represented using $B = \frac{N_{\mathrm{L}}}{N_{\mathrm{E}}} \log_2 C$ bits.

Using the training dataset $\mathcal{H}$, see (2.7), the VQ-VAE is trained with a loss function comprising three components [40],

$$\mathcal{L}_{\mathrm{VQ-VAE}} = \mathcal{L}_{\mathrm{rec}} + \|\mathrm{sg}(\boldsymbol{z}_e^{(m)}) - \boldsymbol{f}_k^{(m)}\|_2^2 + \beta \|\boldsymbol{z}_e^{(m)} - \mathrm{sg}(\boldsymbol{f}_k^{(m)})\|_2^2, \tag{2.19}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss and $\text{sg}(\cdot)$ represents the stop gradient operator, treating its argument as a constant. To ensure a continuous gradient flow between the encoder and decoder, despite the discontinuity introduced by quantization via (2.18), the gradients from the decoder input to the encoder output are copied using $\boldsymbol{f}_k = \boldsymbol{z}_e^{(m)} + \text{sg}(\boldsymbol{f}_k^{(m)} - \boldsymbol{z}_e^{(m)})$ [40]. The three components of the loss are briefly outlined below:

- Reconstruction loss: The reconstruction loss $\mathcal{L}_{\text{rec}} = -\log(p(\boldsymbol{h}^{(m)} \mid \boldsymbol{f}_k^{(m)}))$ varies depending on the design of the decoder output:

  1. If the decoder outputs a statistical description of the channel $\boldsymbol{h}^{(m)}$ with a mean vector $\boldsymbol{\mu}_k^{(m)}$ and a covariance matrix $\boldsymbol{C}_k^{(m)}$, the reconstruction loss is:

  $$\mathcal{L}_{\text{rec}} = \log\det(\pi \boldsymbol{C}_k^{(m)}) + (\boldsymbol{h}^{(m)} - \boldsymbol{\mu}_k^{(m)})^{\text{H}} (\boldsymbol{C}_k^{(m)})^{-1} (\boldsymbol{h}^{(m)} - \boldsymbol{\mu}_k^{(m)}). \tag{2.20}$$

  The inference of a statistical description of a channel was similarly achieved with a VAE for channel estimation in [26].

  2. If an instantaneous reconstruction of the channel $\boldsymbol{h}^{(m)}$ is desired, the decoder's covariance matrix is fixed to be the identity, and only a mean vector denoted by $\bar{\boldsymbol{h}}^{(m)}$, is inferred. Accordingly, the reconstruction loss in this case is the MSE:

  $$\mathcal{L}_{\text{rec}} = \|\boldsymbol{h}^{(m)} - \bar{\boldsymbol{h}}^{(m)}\|_2^2. \tag{2.21}$$

  This approach was similarly employed in VQ-VAE-based channel compression and decompression in, e.g., [41, 42].

- Dictionary learning term: The second term ensures that the embedding $\mathcal{E}$ is learned such that the quantized latent vector $\boldsymbol{f}_k^{(m)}$ is as close as possible to the unquantized latent vector $\boldsymbol{z}_e^{(m)}$.

- Commitment loss: The third term encourages the encoder to commit to the embedding $\mathcal{E}$ and prevents the encoder output from growing unbounded. The hyperparameter $\beta = 0.25$ balances the commitment loss and enhances the focus on the dictionary learning for improved reconstruction [40, 41].

In summary, the encoder is optimized using the first and third terms of the loss (2.19), the embedding $\mathcal{E}$ is updated through the second term, and the decoder is optimized using the first term.

Generative Modeling-based Feedback Schemes for Precoder Design

3

This chapter focuses on the data transmission phase following the pilot transmission phase of one block of the block fading setup, see Section 2.1. For simplicity, we omit the block index t throughout. The goal is to design precoders to achieve a high throughput, measured as spectral efficiency in single-user systems and sum rate in multi-user systems, similar to, e.g., [11]. In the considered FDD system, the BS is unaware of the perfect CSI of the MTs and relies on feedback information provided by the MTs after pilot-based feedback inference, subject to a budget of B bits.

The main contributions discussed in this chapter are organized as follows:

- First, we review conventional codebook-based methods for designing transmit covariance matrices in point-to-point MIMO systems. Inspired by the conventional methods, we propose leveraging GMMs for feedback inference and codebook design in point-to-point MIMO systems. Parts of Sections 3.1.1 and 3.1.2 and the accompanying simulations have been published in [55] (©2022 IEEE) and [56].

- We demonstrate how model-based insights can be utilized to impose structural constraints on GMMs, thereby reducing the model transfer overhead. Additionally, we introduce two novel approaches that enable variable feedback bit lengths with GMMs, further reducing the model transfer overhead. Parts of Sections 3.1.3 and 3.1.4 and the accompanying simulations have been published in [57] (©2023 IEEE).

- Similar to the single-user case, we provide a brief overview of conventional methods for precoder design in MU-MIMO systems. We propose utilizing GMMs for providing feedback information of multiple MTs for precoder design in MU-MIMO systems. This includes introducing a subspace-based approach and an alternative method that utilizes the sample generation capabilities of GMMs combined with a state-of-the-art stochastic iterative algorithm. Structurally

constrained GMMs and variable bit-length feedback strategies similarly transfer to the multi-user context. Parts of Sections 3.2.1 and 3.2.2 and the accompanying simulations have been published in [56] and [57] (©2023 IEEE).

- A GNN-based framework for sum-rate maximization in massive MIMO systems utilizing statistical channel knowledge is proposed. The framework is adapted to incorporate the GMM-based feedback scheme, benefiting from structural constraints imposed on GMM covariance matrices. Parts of Section 3.2.3 and the accompanying simulations have been published in [58] and [59] (©2024 IEEE).

- For systems with higher bit budgets, VQ-VAEs are proposed for feedback inference and precoder design combined with a state-of-the-art stochastic iterative algorithm in massive MIMO systems. The loss function and the architecture of the VQ-VAE are tailored using model-based insights. Parts of Section 3.2.4 and the accompanying simulations have been published in [60].

- The versatility of the proposed GMM-based feedback approach is emphasized, showcasing its adaptability to single- or multi-user scenarios and its flexibility in adapting to any SNR level and pilot configuration without retraining. Moreover, a complexity analysis is provided. Parts of Section 3.3 have been published in [56] and [57] (©2023 IEEE).

- We conclude this chapter with simulations showcasing the superior performance of the proposed methods for precoder design, particularly in systems characterized by low pilot overhead.

3.1 Single-user Systems

3.1.1 Conventional Codebooks and Limited Feedback Encoding

In an offline training phase, one can construct a codebook $\mathcal{Q}$ with $K = 2^B$ elements. A standard codebook construction approach uses Lloyd's algorithm [6, 96]. Given the training dataset $\mathcal{H}$, see (2.7), the iterative Lloyd clustering algorithm alternates between two stages until a convergence criterion is met. Note that we use the channel matrix $\boldsymbol{H}^{(m)}$ and its vectorized expression $\boldsymbol{h}^{(m)} \in \mathcal{H}$ interchangeably in the following for ease of notation. We write $\{\boldsymbol{Q}_k^{(i)}\}_{k=1}^K$ for the codebook in iteration i. The two stages in iteration i are:

1. Divide the training dataset $\mathcal{H}$ into K clusters $\mathcal{V}_k^{(i)}$, with $k = 1, \ldots, K$:

$$\mathcal{V}_k^{(i)} = \{\boldsymbol{h}^{(m)} \in \mathcal{H} \mid r(\boldsymbol{H}^{(m)}, \boldsymbol{Q}_k^{(i)}) \geq r(\boldsymbol{H}^{(m)}, \boldsymbol{Q}_j^{(i)}) \text{ for } j \neq k\}. \tag{3.1}$$

2. Update the codebook:

$$\boldsymbol{Q}_k^{(i+1)} = \underset{\boldsymbol{Q} \succeq \mathbf{0}}{\arg\max} \frac{1}{|\mathcal{V}_k^{(i)}|} \sum_{\mathrm{vec}(\boldsymbol{H}^{(m)}) \in \mathcal{V}_k^{(i)}} r(\boldsymbol{H}^{(m)}, \boldsymbol{Q}) \quad \text{s.t.} \quad \mathrm{tr}(\boldsymbol{Q}) \leq \rho \tag{3.2}$$

where for a channel matrix $\boldsymbol{H}^{(m)}$ and a covariance matrix $\boldsymbol{Q}$,

$$r(\boldsymbol{H}^{(m)}, \boldsymbol{Q}) = \log_2 \det \left(\mathbf{I} + \frac{1}{\sigma_n^2} \boldsymbol{H}^{(m)} \boldsymbol{Q} \boldsymbol{H}^{(m),\mathrm{H}} \right) \tag{3.3}$$

is the spectral efficiency. The optimization problem in stage 2 is solved via projected gradient ascent (PGA) [10, 97]. To initialize the algorithm, stage 1 is replaced with a random partition of $\mathcal{H}$ in the first iteration.

Lau's heuristic [6]: To avoid solving the costly optimization problem in stage 2 in every iteration, a heuristic for the codebook update is given in [6]. In particular, a representative matrix $\boldsymbol{S}_k^{(i)} = \frac{1}{|\mathcal{V}_k^{(i)}|} \sum_{\mathrm{vec}(\boldsymbol{H}^{(m)}) \in \mathcal{V}_k^{(i)}} \boldsymbol{H}^{(m),\mathrm{H}} \boldsymbol{H}^{(m)}$ is calculated for every cluster $\mathcal{V}_k^{(i)}$, and then the matrices $\boldsymbol{S}_k^{(i)}$ are decomposed into N_{tx} parallel streams and water-filling is employed, yielding the updated codebook entries $\boldsymbol{Q}_k^{(i+1)}$. However, as attested by the simulation results in [55], the heuristic approach leads to a performance loss as compared to the PGA approach. Thus, we restrict our analysis to the PGA approach in this dissertation.

In the online phase, following the pilot transmission phase, the MT is assumed to estimate the DL channel, denoted as $\hat{\boldsymbol{H}}$, based on the pilot observation $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, see (2.2). Using this estimate, the MT determines the best codebook entry $\boldsymbol{Q}_{k^\star}$ of the commonly shared codebook $\mathcal{Q}$ via:

$$k^\star = \underset{k \in \{1, \ldots, 2^B\}}{\arg\max} \log_2 \det \left(\mathbf{I} + \frac{1}{\sigma_n^2} \hat{\boldsymbol{H}} \boldsymbol{Q}_k \hat{\boldsymbol{H}}^{\mathrm{H}} \right). \tag{3.4}$$

The feedback consists of the index $k^\star$ encoded using B bits, which the BS then uses to select the transmit covariance matrix $\boldsymbol{Q}_{k^\star}$ for data transmission. Several channel estimators, discussed in Section 3.4.1, are considered for this baseline.

3.1.2 GMM-based Feedback Encoding and Codebook Design

Once a GMM with $K = 2^B$ components is fitted using the training dataset $\mathcal{H}$, see (2.7), the Gaussian nature of each GMM component, see (2.10), combined with the AWGN, see (2.2), allows for simple computation of the GMM of the observations based on the GMM from (2.10) with $\boldsymbol{A} = (\boldsymbol{P} \otimes \mathbf{I}_{N_{\mathrm{rx}}})$ as

$$p_{\mathrm{GMM}}(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}_k, \boldsymbol{A}\boldsymbol{C}_k\boldsymbol{A}^{\mathrm{H}} + \boldsymbol{\Sigma}). \tag{3.5}$$

Accordingly, the MT can compute the responsibilities based on the observation $\boldsymbol{y}$ as

$$p(k \mid \boldsymbol{y}) = \frac{\pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}_k, \boldsymbol{A}\boldsymbol{C}_k\boldsymbol{A}^{\mathrm{H}} + \boldsymbol{\Sigma})}{\sum_{i=1}^{K} \pi_i \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}_i, \boldsymbol{A}\boldsymbol{C}_i\boldsymbol{A}^{\mathrm{H}} + \boldsymbol{\Sigma})}. \tag{3.6}$$

The idea of the proposed method is to compute a transmit covariance matrix $\boldsymbol{Q}_k$ for every component of the GMM and to use the responsibilities $p(k \mid \boldsymbol{y})$ to determine the feedback index. In detail, in an offline training phase, we take $K = 2^B$ as the number of GMM components, and after fitting the GMM we compute a codebook $\mathcal{Q} = \{\boldsymbol{Q}_k\}_{k=1}^{K}$ of transmit covariance matrices—one matrix for every GMM component—exclusively at the BS. We explain the codebook construction in another paragraph below. In the online phase, instead of performing explicit channel estimation, the feedback index is directly determined using the responsibilities derived from the maximum a posteriori (MAP) estimation based on the pilot observation $\boldsymbol{y}$:

$$k^{\star} = \arg\max_{k} \; p(k \mid \boldsymbol{y}), \tag{3.7}$$

i.e., the highest responsibility of the observed pilot signal of a MT serves as the feedback information. Consequently, the feedback index $k^{\star}$ is determined without requiring (estimated) CSI. Furthermore, the knowledge of the codebook at the MT is not required. It only requires the parameters of the GMM to compute (3.7).

The term $p(k \mid \boldsymbol{y})$ can be interpreted as an approximation of $p(k \mid \boldsymbol{h})$ from (2.16), due to the fixed noise covariance for each component. Specifically, given the true underlying channel $\boldsymbol{h}$ that leads to the pilot observation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{h} + \boldsymbol{n}$, the responsibility $p(k \mid \boldsymbol{y})$ approximates the probability $p(k \mid \boldsymbol{h})$, which represents the likelihood that the channel $\boldsymbol{h}$ originates from the k-th GMM component. To investigate the influence of using $p(k \mid \boldsymbol{y})$ instead of $p(k \mid \boldsymbol{h})$, it is interesting to look at the performance of the feedback information calculated as

$$k^{\star} = \arg\max_{k} \; p(k \mid \boldsymbol{h}). \tag{3.8}$$

While this approach is impractical, as it requires the channel $\boldsymbol{h}$ to be known during the online phase, it provides a baseline for assessing performance.

Codebook construction: Once the training dataset $\mathcal{H}$ has been used to fit a K-components GMM, we cluster the training data according to their GMM responsibilities, i.e., channels exhibiting high similarities measured in terms of the responsibilities are assigned to the same component and thereby form a cluster. That is, we partition $\mathcal{H}$ into K disjoint sets denoted by

$$\mathcal{V}_k = \{\boldsymbol{h}^{(m)} \in \mathcal{H} \mid p(k \mid \boldsymbol{h}^{(m)}) \geq p(j \mid \boldsymbol{h}^{(m)}) \text{ for } k \neq j\} \tag{3.9}$$

for $k = 1, \ldots, K$. The codebook $\mathcal{Q} = \{\boldsymbol{Q}_k\}_{k=1}^{K}$ is then constructed by computing every transmit covariance matrix $\boldsymbol{Q}_k$ such that it maximizes the average rate in $\mathcal{V}_k$:

$$\boldsymbol{Q}_k = \underset{\boldsymbol{Q} \succeq \boldsymbol{0}}{\arg\max} \frac{1}{|\mathcal{V}_k|} \sum_{\mathrm{vec}(\boldsymbol{H}^{(m)}) \in \mathcal{V}_k} r(\boldsymbol{H}^{(m)}, \boldsymbol{Q}) \quad \text{s.t.} \quad \mathrm{tr}(\boldsymbol{Q}) \leq \rho \tag{3.10}$$

where $r(\boldsymbol{H}^{(m)}, \boldsymbol{Q})$ is the spectral efficiency defined in (3.3). This optimization problem is solved via a PGA algorithm similar as in Section 3.1.1 [10, 97]. Analogously, we can replace the optimization problem in (3.10) with Lau's heuristic from [6], see Section 3.1.1, to compute a transmit covariance matrix for every GMM component, which degrades the performance [55]. Thus, we again restrict our analysis to the PGA approach in this dissertation.

In summary, the GMM is utilized in two stages, i.e., for offline codebook construction, and later for determining the feedback index during the online phase. For the latter, it is not necessary to estimate the channel, and evaluating (3.4) is avoided. This significantly reduces the online computational complexity, which is discussed in Section 3.3. Moreover, the GMM of the observations, see (3.5), can be straightforwardly adapted at the MT to any SNR and pilot configuration by simply updating the means and covariance matrices, see (3.5), without retraining.

3.1.3 Reducing the Model Transfer Overhead of GMMs

To enable a MT to compute feedback indices via (3.7), the parameters of the GMM need to be offloaded to the MT upon entering the BS' coverage area. To reduce the memory and model transfer overhead, model-based insights can be utilized to obtain GMMs with fewer parameters. For example, one can constrain the GMM covariance matrices to a Kronecker factorization of the form

$$\boldsymbol{C}_k = \boldsymbol{C}_k^{\mathrm{tx}} \otimes \boldsymbol{C}_k^{\mathrm{rx}}. \tag{3.11}$$

This is motivated by the Kronecker channel model, which assumes independent scattering in the vicinity of transmitter and receiver [51]. It was observed that this constraint has no notable impact on the performance of the GMM applied for channel estimation [33, 66]. In this context, instead of fitting an unconstrained GMM with $N \times N$-dimensional covariance matrices (where $N = N_{\mathrm{tx}}N_{\mathrm{rx}}$), a GMM specific to the transmit-side and another for the receive-side is fitted. These transmit- and receive-side GMMs have $N_{\mathrm{tx}} \times N_{\mathrm{tx}}$ and $N_{\mathrm{rx}} \times N_{\mathrm{rx}}$-dimensional covariance matrices with K_{tx} and K_{rx} components, respectively. Subsequently, by computing all Kronecker products of the corresponding transmit- and receive-side covariance matrices, a $K = K_{\mathrm{tx}}K_{\mathrm{rx}}$-components GMM with $N \times N$-dimensional covariance matrices is obtained.

Additional structural characteristics can be integrated into the transmit- and receive-side GMM covariance matrices, depending on the antenna array geometry deployed at the transmitter or receiver. In the case of a ULA, it is common to assume a Toeplitz covariance matrix, which, for a large number of antenna elements, is well approximated by a circulant matrix, see, e.g., [65, 66]. If a URA is deployed, the structural assumptions result in block-Toeplitz matrices with Toeplitz blocks, or block-circulant matrices with circulant blocks, see [65, 66]. For example, with a URA deployed at the transmitter with $N_{\mathrm{tx,v}}$ vertical and $N_{\mathrm{tx,h}}$ horizontal ($N_{\mathrm{tx}} = N_{\mathrm{tx,v}}N_{\mathrm{tx,h}}$) elements, the structured covariance matrices can be expressed as

$$\boldsymbol{C}_k^{\mathrm{tx}} = \boldsymbol{D}^{\mathrm{H}} \operatorname{diag}(\boldsymbol{c}_k^{\mathrm{tx}}) \boldsymbol{D}, \tag{3.12}$$

where on the one hand, when assuming a Toeplitz structure, $\boldsymbol{D} = \boldsymbol{D}_{N_{\mathrm{tx,v}}} \otimes \boldsymbol{D}_{N_{\mathrm{tx,h}}}$, where $\boldsymbol{D}_Q$ (with $Q \in \{N_{\mathrm{tx,v}}, N_{\mathrm{tx,h}}\}$) contains the first Q columns of a $2Q \times 2Q$ DFT matrix, and $\boldsymbol{c}_k^{\mathrm{tx}} \in \mathbb{R}_+^{4N_{\mathrm{tx}}}$, see [65, 66]. On the other hand, when assuming circulant structure, we have $\boldsymbol{D} = \boldsymbol{F}_{N_{\mathrm{tx,v}}} \otimes \boldsymbol{F}_{N_{\mathrm{tx,h}}}$, where $\boldsymbol{F}_Q$ (with $Q \in \{N_{\mathrm{tx,v}}, N_{\mathrm{tx,h}}\}$) is the $Q \times Q$ DFT-matrix, and $\boldsymbol{c}_k^{\mathrm{tx}} \in \mathbb{R}_+^{N_{\mathrm{tx}}}$, see [65, 66]. As another example, with a ULA deployed at the receiver, the structured covariance matrices can be expressed as

$$\boldsymbol{C}_k^{\mathrm{rx}} = \boldsymbol{D}^{\mathrm{H}} \operatorname{diag}(\boldsymbol{c}_k^{\mathrm{rx}}) \boldsymbol{D}, \tag{3.13}$$

where in this case $\boldsymbol{D}$ contains either the first N_{rx} columns of a $2N_{\mathrm{rx}} \times 2N_{\mathrm{rx}}$ DFT matrix and correspondingly $\boldsymbol{c}_k^{\mathrm{rx}} \in \mathbb{R}_+^{2N_{\mathrm{rx}}}$ (Toeplitz), or $\boldsymbol{D} = \boldsymbol{F}_{N_{\mathrm{rx}}}$ and $\boldsymbol{c}_k^{\mathrm{rx}} \in \mathbb{R}_+^{N_{\mathrm{rx}}}$ (circulant), see [65, 66]. Enforcing these structural properties due to the antenna array geometries on the GMM covariance matrices requires adapting the M-step of the EM algorithm outlined in Section 2.3.1 following [66].

Overall, these insights significantly reduce the model transfer overhead, enable a reduced complexity in offline training, facilitate parallelization of the fitting process, and

Name	Covariance Parameters (real-valued)	Example
Full	KN^2	$1.68 \cdot 10^7$
Kronecker	$K_{\mathrm{rx}}N_{\mathrm{rx}}^2 + K_{\mathrm{tx}}N_{\mathrm{tx}}^2$	$1.74 \cdot 10^4$
Toeplitz	$2K_{\mathrm{rx}}N_{\mathrm{rx}} + 4K_{\mathrm{tx}}N_{\mathrm{tx}}$	$2.18 \cdot 10^3$
Circulant	$K_{\mathrm{rx}}N_{\mathrm{rx}} + K_{\mathrm{tx}}N_{\mathrm{tx}}$	$5.76 \cdot 10^2$

Table 3.1: Analysis of the number of covariance parameters of the GMMs variants. Table adapted from [57] (©2023 IEEE).

require fewer training samples since fewer parameters need to be learned, cf. [33,65,66]. A theoretical justification for the Toeplitz structure regularization is provided in [78], which requires the far-field approximation to hold true. Furthermore, it is shown that the means of the GMM components can be set to zero, i.e., $\boldsymbol{\mu}_k = \mathbf{0}$, acting as another form of regularization.

Table 3.1 depicts the number of (structured) GMM covariance parameters (accounting for symmetries). There, "Full" corresponds to unconstrained covariance matrices. "Kronecker" enforces a Kronecker factorization on the covariance matrices. "Toeplitz" further imposes a block-Toeplitz structure on the transmit-side (reflecting the URA at the BS) and a Toeplitz structure on the receive-side covariance matrices (due to the ULA at the MT). Similarly, "Circulant" enforces block-circulant and circulant structures on the transmit- and receive-side covariance matrices, respectively. We use exemplarily the simulation parameters of a configuration with $B = 6$, $(N_{\mathrm{tx}}, N_{\mathrm{rx}}) = (32, 16)$, and $(K_{\mathrm{tx}}, K_{\mathrm{rx}}) = (16, 4)$, which we consider in Section 3.4. We can observe that structural constraints significantly reduce the model transfer overhead.

Although the discussed model-based insights serve as regularization and affect the model quality and decrease the model transfer overhead, they do not affect the algorithmic procedures for feedback inference and precoder design discussed throughout this chapter.

3.1.4 Enabling Variable Feedback Bit Lengths via GMMs

The GMM-based feedback scheme is enhanced to support variable feedback bit lengths by leveraging a single learned GMM and performing merging or pruning of mixture components. In particular, we investigate mixture reduction techniques, where given a GMM with $K = 2^B$ components, another GMM with fewer components $K_S = 2^{B_S}$, with $K_S < K$, can be obtained. Interestingly, simulation results, see

Algorithm 1 Variable Bit Lengths Enabled Through Merging. Adapted from [57, Algorithm 1] (©2023 IEEE).

Require: GMM with $K = 2^B$ components. Desired number of components $K_S = 2^{B_S}$ with $K_S < K$.

$i = K$ {track number of GMM components}
repeat
  Compute moment preserving merge of all pairs of components using (3.14).
  Find pair with smallest dissimilarity $d_{a,b}$, via (3.15).
  Replace this pair by their moment preserving merge.
  $i \leftarrow i - 1$
until $i = K_S$
Partition $\mathcal{H}$ into K_S disjoint sets $\mathcal{V}_k$ with $k = 1, \ldots, K_S$ according to the GMM responsibilities, cf. (3.9).
Construct codebook $\mathcal{Q} = \{\boldsymbol{Q}_k\}_{k=1}^{K_S}$ by solving (3.10) for each entry.

Sections 3.4.3 and 3.4.2, show that reducing the number of components using these techniques can sometimes improve performance compared to directly fitting a GMM with a smaller component count. While GMMs possess the universal approximation property [35], their components may not be distinct enough from each other to be interpreted as clusters [98]. This impacts the GMM-based feedback scheme discussed in Section 3.1.2, where the number of mixture components is predetermined by the feedback bit budget. Consequently, the GMM should exhibit both a strong clustering ability for selecting the feedback index and an accurate likelihood model of the underlying channel distribution of the communication scenario while maintaining a fixed number of components. To address this, mixture reduction techniques that have been proposed in the literature can be used [98–100]. Additionally, these mixture reductions may also act as regularization, helping to prevent overfitting.

In this dissertation, we analyze two types of mixture reduction techniques. In particular, we investigate one merging strategy, which successively combines pairs of components with high similarity [99] and one pruning strategy, which simply discards components with a low contribution to the overall mixture [100].

The merging technique from [99] replaces two mixture components with the smallest dissimilarity denoted by $\{\pi_a, \boldsymbol{\mu}_a, \boldsymbol{C}_a\}$ and $\{\pi_b, \boldsymbol{\mu}_b, \boldsymbol{C}_b\}$, by their moment

Algorithm 2 Variable Bit Lengths Enabled Through Pruning. Adapted from [57, Algorithm 2] (©2023 IEEE).

Require: GMM with $K = 2^B$ components. Desired number of components $K_S = 2^{B_S}$ with $K_S < K$.

1: Remove the $K - K_S$ components with the smallest mixing coefficients π_k.
2: Re-normalize the remaining mixing coefficients to one.
3: Discard codebook entries correspondingly and obtain $\mathcal{Q} = \{\boldsymbol{Q}_k\}_{k=1}^{K_S}$.

preserving merge $\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{C}_m\}$ [99]:

$$\begin{aligned} \pi_m &= \pi_a + \pi_b, \\ \boldsymbol{\mu}_m &= \tfrac{1}{\pi_m}(\pi_a \boldsymbol{\mu}_a + \pi_b \boldsymbol{\mu}_b), \\ \boldsymbol{C}_m &= (\tfrac{\pi_a}{\pi_m}\boldsymbol{C}_a + \tfrac{\pi_b}{\pi_m}\boldsymbol{C}_b + \tfrac{\pi_a \pi_b}{\pi_m^2}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^{\mathrm{H}}). \end{aligned} \tag{3.14}$$

The dissimilarity between two components is measured by

$$d_{a,b} = \pi_m \log\det(\boldsymbol{C}_m) - \pi_a \log\det(\boldsymbol{C}_a) - \pi_b \log\det(\boldsymbol{C}_b), \tag{3.15}$$

which follows from an upper bound on the Kullback-Leibler divergence between the GMMs before and after the merge, cf. [99]. The merging procedure is successively repeated until a GMM of the desired size is obtained, i.e., when the number of GMM components is reduced to 2^{B_S}. After a GMM of the desired size is obtained, the corresponding codebook needs to be constructed at the BS in the same way as described in Section 3.1.2. Algorithm 1 summarizes the merging procedure and the accompanying codebook construction.

The pruning strategy simply removes $K - K_S$ components with the smallest mixing coefficient π_k, see (2.10), and the mixing coefficients of the remaining components are re-normalized to one [100]. A main advantage of the pruning strategy is its low computational complexity. In particular, it only requires traversing a list of sorted mixing coefficients. In this case, an updated codebook $\mathcal{Q} = \{\boldsymbol{Q}_k\}_{k=1}^{K_S}$ can simply be obtained by discarding the respective codebook entries which correspond to the removed GMM components. Algorithm 2 summarizes the pruning procedure.

By using these mixture reduction techniques, only a single GMM needs to be transferred to the MT, which can then support various feedback bit budgets $(B, B-1, B-2, \dots)$, thereby reducing the overall model transfer overhead.

3.2 Multi-user Systems

3.2.1 Conventional Limited Feedback Encoding and Precoder Design

In the seminal work [7], limited feedback in MU-MIMO systems was investigated, where quantized CSI information for each MT is fed back to the BS after determining the best entry of a randomly generated MT-specific codebook. The random quantization codebook of each MT is assumed to be perfectly known to the BS [7]. The feedback conveys only the spatial direction of each MT's channel, omitting magnitude information, and BD, see [85], was employed at the BS to jointly design precoders with a uniform power allocation policy, avoiding water-filling across streams. Multi-user systems with J MTs were considered, ensuring that the total number of receive antennas JN_{rx} equals the number of transmit antennas N_{tx}, thereby eliminating the need to select a subset of MTs for transmission [7]. We also consider such setups in this dissertation.

Consider the singular value decomposition (SVD) of the (estimated) channel $\hat{\boldsymbol{H}}_j = \boldsymbol{U}_j \boldsymbol{S}_j \boldsymbol{V}_j^{\mathrm{H}}$ of MT j, where $\boldsymbol{S}_j$ contains the singular values in descending order on its diagonal, and let $\bar{\boldsymbol{V}}_j^{\mathrm{H}}$ contain the first N_{rx} rows of $\boldsymbol{V}_j^{\mathrm{H}}$, cf. e.g., [101]. Given the matrix $\bar{\boldsymbol{V}}_j^{\mathrm{H}}$, the idea from [7] is to feed back the information regarding $\bar{\boldsymbol{V}}_j^{\mathrm{H}}$ to the BS by using a random quantization codebook [102]. In particular, each MT-specific random quantization codebook is fixed beforehand and is known to the BS. That is, the codebook $\mathcal{Q} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_K\}$ of a particular MT consists of $K = 2^B$ sub-unitary matrices (i.e., $\boldsymbol{W}_k^{\mathrm{H}} \boldsymbol{W}_k = \mathbf{I}_{N_{\mathrm{rx}}}$, for $k = 1, \ldots, K$) which are chosen independently and are uniformly distributed over the Grassmann manifold [7, 103]. The elements of the random quantization codebook are thus constructed by generating an $N_{\mathrm{tx}} \times N_{\mathrm{rx}}$ dimensional matrix with i.i.d. complex standard Gaussian entries and then computing an $N_{\mathrm{tx}} \times N_{\mathrm{rx}}$ dimensional subspace spanned by the matrix using the procedure described in [104]. The selection method considered in [7] was the chordal distance metric, i.e., $k_j^{\star} = \arg\min_k \left(N_{\mathrm{rx}} - \mathrm{tr}(\bar{\boldsymbol{V}}_j^{\mathrm{H}} \boldsymbol{W}_k \boldsymbol{W}_k^{\mathrm{H}} \bar{\boldsymbol{V}}_j) \right)$, and other metrics were not investigated in [7]. We also considered the chordal distance in our simulations but found that using a capacity-inspired selection metric (like in [101]) performed consistently better. Thus, we use the same evaluation principle per MT as in (3.4), but replace the transmit covariance matrices by $\frac{\rho}{N_{\mathrm{rx}}} \boldsymbol{W}_k \boldsymbol{W}_k^{\mathrm{H}}$:

$$k_j^{\star} = \arg\max_k \ \log_2 \det \left(\mathbf{I} + \frac{\rho}{\sigma_n^2 N_{\mathrm{rx}}} \hat{\boldsymbol{H}}_j \boldsymbol{W}_k \boldsymbol{W}_k^{\mathrm{H}} \hat{\boldsymbol{H}}_j^{\mathrm{H}} \right). \tag{3.16}$$

In this way, we also account for the SNR compared to the chordal distance metric, which does not depend on the SNR. For the sake of brevity, we do not show the results

obtained by using the chordal distance metric in Section 3.4.3, since we compare to the consistently better baseline.

Each MT reports $k_j^\star$ to the BS which then represents each MT's channel by $\widetilde{\boldsymbol{H}}_j = \boldsymbol{W}_{k_j^\star}^{\mathrm{H}}$. Given the quantized directional information regarding each MT's channel, the BS can employ a common precoding algorithm in order to jointly design the precoders. The authors of [7] focused their analysis on BD and adopted the uniform power allocation policy. In contrast, we consider a broad range of precoding algorithms, including non-iterative algorithms, RBD with the uniform power allocation policy [86] (an extension of BD), RCI [88], and the iterative WMMSE algorithm [90, Algorithm 1], in order to jointly design the precoders $\boldsymbol{M}_j$ of all MTs $j \in \mathcal{J}$.

3.2.2 GMM-based Feedback Encoding and Precoder Design

3.2.2.1 Subspace-based Method

Inspired by the approach from [7], we propose to use the following procedure to obtain quantized feedback information regarding each MT's channel. Each transmit covariance matrix of the Lloyd codebook, see Section 3.1.1, or the GMM codebook, see Section 3.1.2, contains quantized information regarding the spatial directions of each MT's channel and power loadings per stream. We can extract the directional information of each codebook entry by simply performing an SVD of each transmit covariance matrix, that is,

$$\boldsymbol{Q}_k = \boldsymbol{X}_k \boldsymbol{T}_k \boldsymbol{X}_k^{\mathrm{H}}, \tag{3.17}$$

where $\boldsymbol{T}_k$ contains the singular values in descending order, and $\bar{\boldsymbol{X}}_k$ collects the first N_{rx} vectors of $\boldsymbol{X}_k$, as the respective directional information. Accordingly, we have $\mathcal{Q} = \{\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2, \cdots, \bar{\boldsymbol{X}}_K\}$, which thus constitutes a directional codebook. Note that, we have to ensure to take a codebook at sufficiently large SNR in order to guarantee that the matrices $\boldsymbol{Q}_k$ all have a rank larger than or equal to N_{rx}. In our simulations this was the case with the codebooks constructed for an SNR of $25\,\mathrm{dB}$.

The described subspace-based method can be used in combination with both, the Lloyd codebook, see Section 3.1.1, and the proposed GMM-based codebook, see Section 3.1.2. In case of the subspaces extracted from the Lloyd codebook, the feedback per MT is calculated after channel estimation by evaluating

$$k_j^\star = \arg\max_k \; \log_2 \det \left(\mathbf{I} + \frac{\rho}{\sigma_n^2 N_{\mathrm{rx}}} \hat{\boldsymbol{H}}_j \bar{\boldsymbol{X}}_k \bar{\boldsymbol{X}}_k^{\mathrm{H}} \hat{\boldsymbol{H}}_j^{\mathrm{H}} \right). \tag{3.18}$$

In the multi-user case, the same GMM utilized in the single-user case can be leveraged without requiring additional training or modifications. We propose that each

MT j determines its feedback information through MAP estimation, similarly to the single-user case, see (3.7), using its pilot observations $\boldsymbol{y}_j$ as:

$$k_j^\star = \arg\max_k \; p(k \mid \boldsymbol{y}_j). \tag{3.19}$$

In both cases, each MT reports $k_j^\star$ to the BS which represents each MT's channel using the subspace information associated with the respective codebook entry $\widetilde{\boldsymbol{H}}_j = \bar{\boldsymbol{X}}_{k_j^\star}^{\mathrm{H}}$. The BS can then employ RBD with the uniform power allocation policy, RCI, or the iterative WMMSE algorithm [90, Algorithm 1], to jointly design the precoders $\boldsymbol{M}_j$ of all MTs $j \in \mathcal{J}$.

For performance reference, perfect CSI can be assumed at each MT j to determine the feedback index, cf. (3.8):

$$k_j^\star = \arg\max_k \; p(k \mid \boldsymbol{h}_j) \tag{3.20}$$

In the case of GMMs which are obtained after applying a mixture reduction technique, see Section 3.1.4, the above procedure is similarly applied utilizing the codebooks obtained via Algorithm 1 or Algorithm 2, accordingly.

3.2.2.2 Generative Modeling-based Method

As an alternative to the aforementioned subspace-based method, the channel matrix of each MT may be modeled as a random variable, similar to [105, 106], and the ergodic rate $\mathrm{E}[R_j^{\text{inst}}]$ of each MT $j \in \mathcal{J}$ is considered (note that the expectation is taken with respect to the channel distribution). Then, the ergodic sum-rate maximization problem can be written as [105, 106]

$$\max_{\{\boldsymbol{M}_j\}_{j=1}^J} \sum_{j=1}^J \mathrm{E}[R_j^{\text{inst}}] \;\text{ s.t. }\; \mathrm{tr}\left(\sum\nolimits_{j=1}^J \boldsymbol{M}_j \boldsymbol{M}_j^{\mathrm{H}}\right) = \rho. \tag{3.21}$$

In [105, 106], the authors proposed the stochastic WMMSE (SWMMSE) algorithm. In each iteration step, the SWMMSE algorithm requires channel samples that represent statistical information about each MT. It was shown that the algorithm is guaranteed to converge to the set of stationary points of the stochastic sum-rate maximization problem almost surely as the number of iterations approaches infinity [105, 106].

The discussed conventional methods are unable to perform the SWMMSE iterations since they are lacking of a generative model, i.e., a model that learns the underlying PDF of the channels and allows to generate new samples that resemble the original channel distribution. In contrast, the proposed GMM approach is able to generate

Algorithm 3 GMM-based Feedback Scheme for Precoder Design in MU-MIMO Systems via the SWMMSE. Adapted from [56, Algorithm 1].

Require: Offline trained GMM. The GMM is available at both the MTs and the BS.

GMM-based Feedback Inference at the MTs

1: Infer feedback index $k_j^\star$ from observation $\boldsymbol{y}_j$ at each MT j
$k_j^\star = \arg\max_k \; p(k \mid \boldsymbol{y}_j) \; \forall j$

2: Each MT j feeds back $k_j^\star$ encoded by $B = 2^K$ bits

SWMMSE-based Precoder Design at the BS

3: Provide statistics of each MT j to the SWMMSE and obtain precoders
$\{\boldsymbol{M}_j\}_{j=1}^J \leftarrow \text{SWMMSE}(\{\tilde{\boldsymbol{h}}_j \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_{k_j^\star}, \boldsymbol{C}_{k_j^\star})\}_{j=1}^J)$

samples following the channel's distribution due to the GMM's sample generation ability. This allows to jointly design the precoders via the SWMMSE algorithm by exploiting statistical information about the MTs given their feedback information. In particular, given the feedback index $k_j^\star$ of each MT, see (3.19), one can draw samples from the respective GMM components via

$$\tilde{\boldsymbol{h}}_j \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_{k_j^\star}, \boldsymbol{C}_{k_j^\star}), \tag{3.22}$$

which represent statistical information about the channel of MT j. The BS can subsequently design the precoders exploiting the SWMMSE algorithm based on the generated samples utilizing the GMM. In order to feed the channel samples to the SWMMSE algorithm, the channels have to be reshaped as $\tilde{\boldsymbol{H}}_j = \text{unvec}_{N_{\text{rx}},N_{\text{tx}}}(\tilde{\boldsymbol{h}}_j)$. Please see [105, 106] for more details on the SWMMSE. A summary of the proposed generative modeling-based precoder design method is given in Algorithm 3.

For GMMs derived through a mixture reduction technique as discussed in Section 3.1.4, the same procedure described above is applied.

3.2.3 Statistical Precoder Design via GNNs

In the remainder of this chapter, we focus on massive MIMO systems, i.e., we have a MISO channel $\boldsymbol{h}_j$ per MT j, resulting in $N = N_{\text{tx}}$. Consequently, the precoders are denoted as $\boldsymbol{m}_j$ in the remainder.

3.2.3.1 Precoder Design with Perfect Statistical Knowledge via GNNs

To obtain the precoders $\{\boldsymbol{m}_j\}_{j=1}^J$ solely utilizing statistical knowledge about the channel $\boldsymbol{h}_j$ of each MT, the key adaptation of the GNN compared to the existing

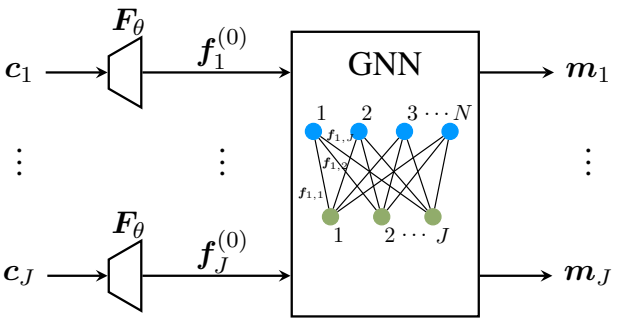


Figure 3.1: Block diagram of the proposed GNN-based precoder design framework. Figure adapted from [58].

works [37, 39], is given at the input. With perfect statistical knowledge, i.e, assuming the knowledge of $\{\boldsymbol{C}_{\boldsymbol{\delta}_j}\}_{j=1}^J$, see (2.2.1), model-based insights can be exploited to present the statistical knowledge to the GNN. Specifically, for a ULA, the covariance matrix is Toeplitz-structured [80], and for a URA, it is block-Toeplitz with Toeplitz blocks [107]. In both cases, the covariance matrix per MT j is fully parametrized by its first row (or column). Thus, for each MT j, the first row of the respective covariance matrix is extracted

$$\boldsymbol{c}_j = \boldsymbol{C}_{\boldsymbol{\delta}_j}[1, :]. \tag{3.23}$$

Subsequently, the real and imaginary parts of $\boldsymbol{c}_j$ are stacked and passed through a feature extractor $\boldsymbol{F}_\theta$ to obtain the input feature vectors $\boldsymbol{f}_j^{(0)}$ which are relevant for the GNN processing. This has the advantage of a reduced input dimension of $2N$ per MT, which consequently does not scale with the covariance matrix dimension of N^2 and additionally inherently reduces the computational complexity. To facilitate scalability, the feature extractor $\boldsymbol{F}_\theta$ consists of a single fully connected layer followed by a PReLU activation function, and parameter sharing is considered such that the same feature extractor $\boldsymbol{F}_\theta$ is used for all MTs, similar to [39].

As GNN architecture, we adopt the Edge-GAT model from [39], originally termed A2D-GNN in [37]. The GNN is structured as a neural network with multiple layers. This model, designed with an attention mechanism in (3.25) that differently weights interfering edges connected to the n-th antenna, demonstrates superior generalization over other GNN architectures, e.g., the Edge-GNN model from [36], see [37, 39]. Also, in our simulations, the Edge-GAT model consistently outperformed the Edge-GNN model from [36]. Thus, we solely focus on Edge-GAT throughout, referring to it as the GNN in the remainder. The GNN's associated graph is bipartite, where one set of vertices corresponds to the N antennas at the BS and another set to the J

single-antenna MTs, and the features are solely associated with the edges of the graph. For the GNN processing, each feature vector $\boldsymbol{f}_j^{(0)} = [\boldsymbol{f}_{1,j}^{(0),\mathrm{T}}, \ldots, \boldsymbol{f}_{N,j}^{(0),\mathrm{T}}]^{\mathrm{T}}$ must be at least of dimension N such that it contains all initial features for the edges (n, j) for $n = 1, \ldots, N$ [39]. Similar to [39], we set the output dimension of the feature extractor $\boldsymbol{F}_\theta$ such that $\boldsymbol{f}_{n,j}^{(0)} \in \mathbb{R}^2$ for $n = 1, \ldots, N$ and $j = 1, \ldots, J$. Let $\boldsymbol{f}_{n,j}^{(\ell-1)}$ denote the feature vector associated with the edge (n, j) at the input of the ℓ-th layer, with $\ell = 1, \ldots, L$. The update rule for the edge (n, j) is then [37, 39]

$$\boldsymbol{f}_{n,j}^{(\ell)} = f_{\text{act}} \left(\boldsymbol{S}^{(\ell)} \boldsymbol{f}_{n,j}^{(\ell-1)} + \alpha \sum_{\substack{i=1 \\ i \neq n}}^{N} \boldsymbol{T}^{(\ell)} \boldsymbol{f}_{i,j}^{(\ell-1)} + \beta \sum_{\substack{k=1 \\ k \neq j}}^{J} \boldsymbol{a}_{j,k}^{(\ell)} \odot \boldsymbol{u}_{n,k}^{(\ell)} \right) \tag{3.24}$$

with $\boldsymbol{a}_{jk}^{(\ell)}$ obtained through a linear attention mechanism as

$$\begin{aligned} \boldsymbol{q}_{n,j}^{(\ell)} = \boldsymbol{Q}^{(\ell)} \boldsymbol{f}_{n,j}^{(\ell-1)} \quad \boldsymbol{k}_{n,j}^{(\ell)} = \boldsymbol{K}^{(\ell)} \boldsymbol{f}_{n,j}^{(\ell-1)} \quad \boldsymbol{u}_{n,j}^{(\ell)} = \boldsymbol{U}^{(\ell)} \boldsymbol{f}_{n,j}^{(\ell-1)} \\ \boldsymbol{a}_{j,k}^{(\ell)} = \sum\nolimits_{n=1}^{N} \boldsymbol{q}_{n,j}^{(\ell)} \odot \boldsymbol{k}_{n,k}^{(\ell)} / N. \end{aligned} \tag{3.25}$$

The matrices $\boldsymbol{S}^{(\ell)}$, $\boldsymbol{T}^{(\ell)}$, $\boldsymbol{Q}^{(\ell)}$, $\boldsymbol{K}^{(\ell)}$, $\boldsymbol{U}^{(\ell)} \in \mathbb{R}^{M_\ell \times M_{\ell-1}}$ denote the shared trainable parameters of layer ℓ and are used to update the features associated with all edges (n, j) for $n = 1, \ldots, N$ and $j = 1, \ldots, J$. Accordingly, precoder design for different numbers of MTs is facilitated. The hyperparameters α and β are scalar coefficients, f_{act} denotes the activation function, and $\odot$ denotes the Hadamard product. The features at the last layer $\boldsymbol{f}_{n,j}^{(L)}$ are enforced to be in $\mathbb{R}^2$ such that the elements correspond to the real and imaginary parts of the entry $m'_{n,j}$ in the n-th row and j-th column of the matrix containing the non-normalized precoders $\boldsymbol{M'} = [\boldsymbol{m}'_1, \ldots, \boldsymbol{m}'_J]$, i.e., $\boldsymbol{f}_{n,j}^{(L)} = [\Re(m'_{n,j}), \Im(m'_{n,j})]^{\mathrm{T}}$. After normalizing the matrix $\boldsymbol{M'}$ to satisfy the power constraint, the precoders $\{\boldsymbol{m}_j\}_{j=1}^{J}$ are obtained. In Figure 3.1, the block diagram of the proposed GNN-based precoder design approach utilizing statistical knowledge is depicted.

During the training stage, the sum-rate expression, cf. (2.6), is used as the negative loss function, and we utilize the training set $\mathcal{D}_{\text{genie}} = \{\{(\boldsymbol{c}_j^{(d)}, \boldsymbol{h}_j^{(d)})\}_{j=1}^{J}\}_{d=1}^{D}$, i.e., the training set consists of in total D multi-user scenarios of J MTs, where for each MT the perfect statistical knowledge $\boldsymbol{C}_{\boldsymbol{\delta}_j}^{(d)}$ and the corresponding channel $\boldsymbol{h}_j^{(d)}$ is available. The GNN generalizes to different numbers of MTs by construction. To enable operation at different SNR levels, we train the GNN for an SNR range from $0\,\mathrm{dB}$ to $20\,\mathrm{dB}$, which is also the evaluation range. Thus, only one GNN model is required to design the precoders for a varying number of MTs and different SNR levels, which is beneficial in

terms of memory. Since the GNN consists of a few layers only, and the feature updates in each layer can be processed in parallel, the inference time can be accelerated, where the computational complexity of the GNN module is $\mathcal{O}(\sum_{\ell=1}^{L} M_\ell M_{\ell-1})$ [39].

However, assuming perfect statistical knowledge might not be practical. Inspired by the above procedure, in the following, we adapt the GNN-based precoder design framework to systems with approximate statistical knowledge obtained via the GMM-based feedback scheme from Section 3.2.2.2.

3.2.3.2 Combining GNNs with GMMs for Precoder Design

In this section, we set the means of all GMM components to zero, i.e., $\boldsymbol{\mu}_k = \mathbf{0}$ for all $k \in \{1, \ldots, K\}$. Furthermore, depending on the geometry of the BS antenna array, the covariance matrices of the GMM are constrained to a Toeplitz structure for a ULA at the BS, or to a block-Toeplitz structure for a URA at the BS, as described in Section 3.1.3.

Beyond the regularization benefits and model transfer overhead reduction offered by the structured GMM variants, see Section 3.1.3, the approximate statistics of each MT j, inferred via (3.19) or (3.20) (for performance reference), i.e., assuming $\boldsymbol{h}_j$ is distributed as $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \boldsymbol{C}_{k_j^\star})$, can be compactly represented by just the first row of the respective GMM component's covariance matrix. Thus, for each MT, the first row of the respective GMM component's covariance matrix,

$$\boldsymbol{c}_j = \boldsymbol{C}_{k_j^\star}[1, :], \tag{3.26}$$

is extracted, of which the real and imaginary parts are stacked, and passed through the feature extractor $\boldsymbol{F}_\theta$ to obtain the input feature vectors for the GNN. The GNN output then provides the precoders $\{\boldsymbol{m}_j\}_{j=1}^{J}$ after normalization, cf. Figure 3.1 and Section 3.2.3.1.

To account for the GMM-based feedback for precoder design using the GNN, the training set $\mathcal{D} = \{\{(\boldsymbol{c}_j^{(d)}, \boldsymbol{h}_j^{(d)})\}_{j=1}^{J}\}_{d=1}^{D}$ again includes D multi-user scenarios with J MTs. For each MT, statistical information $\boldsymbol{c}_j^{(d)} = \boldsymbol{C}_{k_j^\star}^{(d)}[1, :]$ is obtained through a pilot observation $\boldsymbol{y}_j^{(d)}$ using the GMM via (3.19), and the corresponding channel $\boldsymbol{h}_j^{(d)}$ is available in the offline training phase. When assuming perfect CSI at each MT, $\boldsymbol{c}_j^{(d)}$ is obtained utilizing (3.20). The GNN is trained for an SNR range from $0\,\mathrm{dB}$ to $20\,\mathrm{dB}$ and enables precoder design for a varying number of MTs. The GMM-based feedback scheme allows multi-user operation for a varying number of MTs J and, due to the GMM's analytic representation, can adapt to any SNR level without requiring retraining, cf. Section 3.1.2. This combination of the GMM-based feedback scheme

Algorithm 4 Combined GMM- and GNN-based Precoder Design Scheme. Adapted from [58, Algorithm 1].

Require: Offline trained GMM and GNN, where the GMM is available at both the MTs and the BS, and the GNN at the BS.

GMM-based Feedback Inference at the MTs

1: Infer feedback index $k_j^\star$ from observation $\boldsymbol{y}_j$ at each MT j
$k_j^\star = \arg\max_k \; p(k \mid \boldsymbol{y}_j) \; \forall j$

2: Each MT j sends back $k_j^\star$ encoded by B bits

GNN-based Precoder Design at the BS

3: Extract the first row of the respective GMM component's covariance matrix for each MT j
$\boldsymbol{c}_j = \boldsymbol{C}_{k_j^\star}[1, :] \; \forall j$

4: Pass $\boldsymbol{c}_{k_j^\star}$ through feature extractor $\boldsymbol{F}_\theta$ for each MT j
$\boldsymbol{f}_j^{(0)} = \boldsymbol{F}_\theta(\boldsymbol{c}_{k_j^\star}) \; \forall j$

5: Provide feature vector of each MT j to the GNN and obtain precoders
$\{\boldsymbol{m}_j\}_{j=1}^J \leftarrow \text{GNN}(\{\boldsymbol{f}_j^{(0)}\}_{j=1}^J)$

and the GNN-based precoder design approach yields a fully flexible approach for multi-user systems. While the GMM knowledge is required at both the MTs and the BS, the GNN knowledge is only required at the BS. An overview of the combined GMM- and GNN-based precoder design scheme is provided in Algorithm 4.

3.2.4 VQ-VAE-based Feedback Encoding and Precoder Design

In this section, we focus on massive MIMO systems with a higher feedback bit budget than before, ranging from tens to hundreds of bits B. Recall that in the GMM-based approach, the number of components is defined as 2^B. However, when the feedback bit budget increases substantially, the number of components, and consequently, the number of parameters to estimate, becomes a critical limiting factor. This not only makes it challenging to fit the GMM adequately but also introduces significant memory and model transfer overhead, restricting its feasibility in such high-budget scenarios. To address these limitations, we employ the VQ-VAE, which is better suited for providing approximate statistics in scenarios with larger feedback bit budgets. By leveraging the VQ-VAE, we can efficiently capture and represent the channel statistics for each MT j, ensuring scalability and practicality for high-bit-budget feedback systems.

The VQ-VAE can be tailored to the BS environment characteristics to enable

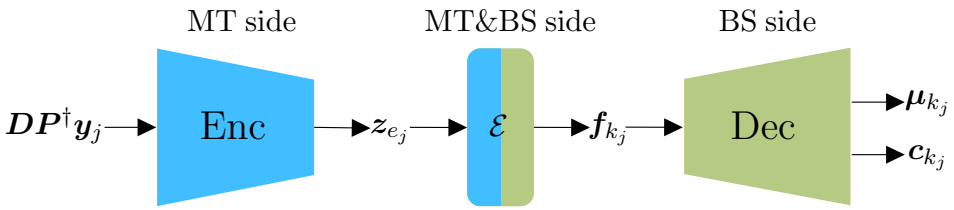


Figure 3.2: Structure of the proposed VQ-VAE for statistical precoder design. Figure adapted from [60].

precoder design utilizing statistical information as follows. We require the VQ-VAE to output a mean vector $\boldsymbol{\mu}_{k_j}$, and another vector $\boldsymbol{c}_{k_j}$ parametrizing a covariance matrix as approximate statistical information about the channel $\boldsymbol{h}_j$ (see Figure 3.2). Depending on the antenna array geometry at the BS, the vector $\boldsymbol{c}_{k_j}$ is post-processed to obtain a covariance matrix

$$\boldsymbol{C}_{k_j} = \boldsymbol{D}^{\mathrm{H}} \operatorname{diag}(\boldsymbol{c}_{k_j}) \boldsymbol{D}, \tag{3.27}$$

cf. Section 3.1.3 and [69]. This structural constraint acts as a regularization and helps to decrease the model parameters since the decoder output vector $\boldsymbol{c}_{k_j}$ fully determines the covariance matrix $\boldsymbol{C}_{k_j}$. Moreover, each MT j first applies the LS estimator to obtain a coarse estimate. Subsequently, we apply $\boldsymbol{D}$ (see Section 3.1.3) as a transformation, which is inspired by [80] and can be interpreted as an angular domain transformation, cf., e.g., [3]. Accordingly, the input to the VQ-VAE is given by

$$\boldsymbol{D}\boldsymbol{P}^{\dagger}\boldsymbol{y}_j, \tag{3.28}$$

with $\boldsymbol{P}^{\dagger} = \boldsymbol{P}^{\mathrm{H}}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{H}})^{-1}$.

To enable operation at different SNR levels, we train the VQ-VAE for an SNR range from 0 to $20\,\mathrm{dB}$ (which is also the evaluation range). In particular, for the offline training, we utilize the training set $\{\boldsymbol{h}^{(m)}\}_{m=1}^{M}$ and generate correspondingly encoder inputs $\boldsymbol{D}\boldsymbol{P}^{\dagger}\boldsymbol{y}^{(m)}$ following (2.1) where we randomly draw a noise variance to obtain an SNR from the range specified above. Taking into account the adaptions to tailor the VQ-VAE for robust precoder design for a specific BS environment, the overall loss function is the loss from (2.19) using the embedding $\mathcal{E}$ and incorporating $\boldsymbol{C}_k^{(m)} = \boldsymbol{D}^{\mathrm{H}} \operatorname{diag}(\boldsymbol{c}_k^{(m)}) \boldsymbol{D}$ to the reconstruction loss from (2.20).

In Figure 3.2, the VQ-VAE with the proposed adaptions is depicted. At the MTs, the knowledge of the encoder and the embedding is required for feedback inference. In the online phase, the observation $\boldsymbol{y}_j$ at each MT j is preprocessed, passed through the encoder to obtain the unquantized latent variable $\boldsymbol{z}_{e_j}$, and finally quantized using the embedding $\mathcal{E}$ to obtain the feedback information vector $\boldsymbol{f}_{k_j}$. Each MT j sends the

Algorithm 5 VQ-VAE-based Feedback Scheme for Precoder Design via the SWMMSE. Adapted from [60, Algorithm 1].

Require: Offline trained VQ-VAE where encoder (Enc) and embedding $\mathcal{E}$ are available at the MTs and decoder (Dec) and embedding $\mathcal{E}$ at the BS.

VQ-VAE-based Feedback Inference at the MTs

1: Infer unquantized latent $\boldsymbol{z}_{e_j}$ from preprocessed observation at each MT j
$\boldsymbol{z}_{e_j} = \mathrm{Enc}(\boldsymbol{D}\boldsymbol{P}^{\dagger}\boldsymbol{y}_j)\ \forall j$

2: Obtain quantized latent representation $\boldsymbol{f}_{k_j}$ at each MT j
$\boldsymbol{f}_{k_j,i} = \arg\min_{\boldsymbol{e}_c \in \mathcal{E}} \|\boldsymbol{e}_c - \boldsymbol{z}_{e_j,i}\|_2$ with $i = 1, \ldots, \frac{N_{\mathrm{L}}}{N_{\mathrm{E}}}\ \forall j$

3: Each MT j sends back $\boldsymbol{f}_{k_j}$ encoded by $B = \frac{N_{\mathrm{L}}}{N_{\mathrm{E}}} \log_2 C$ bits

SWMMSE-based Precoder Design at the BS

4: Obtain statistical description of each MT j
$(\boldsymbol{\mu}_{k_j}, \boldsymbol{c}_{k_j}) = \mathrm{Dec}(\boldsymbol{f}_{k_j})\ \forall j$

5: Compute covariance matrix $\boldsymbol{C}_{k_j}$ for each MT j
$\boldsymbol{C}_{k_j} = \boldsymbol{D}^{\mathrm{H}} \mathrm{diag}(\boldsymbol{c}_{k_j})\boldsymbol{D}\ \forall j$

6: Provide statistics of each MT j to the SWMMSE and obtain precoders
$\{\boldsymbol{m}_j\}_{j=1}^{J} \leftarrow \mathrm{SWMMSE}(\{\tilde{\boldsymbol{h}}_j \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_{k_j}, \boldsymbol{C}_{k_j})\}_{j=1}^{J})$

feedback information encoded with $\frac{N_{\mathrm{L}}}{N_{\mathrm{E}}} \log_2 C$ bits back to the BS. After passing $\boldsymbol{f}_{k_j}$ through the decoder at the BS side, the acquired statistical information about each MT is used in the following way to design precoders.

We utilize the sample generation capabilities of VQ-VAEs, treating the MT channels as random variables, and apply the SWMMSE algorithm from [106] for precoder design, similar to the GMM, see Section 3.2.2.2. In each iteration step, the SWMMSE requires samples that follow the channel distribution of each MT. Therefore, based on each MT's feedback information, we propose generating samples using the VQ-VAE as

$$\tilde{\boldsymbol{h}}_j \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_{k_j}, \boldsymbol{C}_{k_j}), \tag{3.29}$$

resembling samples from the channel distribution of MT j, and feed them to the SWMMSE to design the precoders. A summary is provided in Algorithm 5.

We utilize a similar encoder-decoder architecture as described in [69], adapting both the input and latent dimensions accordingly. Consequently, the computational complexity of the VQ-VAE inference is $\mathcal{O}(LN^2)$, where L represents the number of layers.

3.3 A Versatile Limited Feedback Precoder Design Scheme

3.3.1 Discussion on the Versatility

As discussed earlier, after fitting the GMM at the BS, codebooks to support the point-to-point transmission mode can be constructed, see Section 3.1.2. Then, the GMM of the channels, cf. (2.10), is offloaded to every MT within the coverage area of the BS. In Section 3.1.3, we discussed how model-based insights can help to effectively reduce the offloading overhead. While the BS could offload many GMMs with different sizes to allow for different numbers of feedback bits, the corresponding signaling overhead might be unaffordable in practice. To address this, Section 3.1.4 explores different mixture reduction techniques, enabling adaptable feedback bit lengths using pruning or merging methods. Due to the simplicity of the pruning strategy, it can be straightforwardly implemented at the MTs. The merging strategy can be either conducted at the MTs or, to reduce their computational burden, the BS could offload additional lists with the respective merging updates.

In the online phase, the BS regularly sends pilots to the MTs. Depending on the SNR and the pilots, the GMM of the observations, see (3.5), can be straightforwardly constructed from the offloaded GMM of the channels. Using the GMM of the observations, the received observations $\boldsymbol{y}_j$ at each MT are processed to a feedback index $k_j^\star$ through MAP estimation using the responsibilities via (3.19). Given the feedback information of each MT, the BS determines whether to operate in point-to-point or multi-user mode. In the point-to-point mode (denoted as P2P in Figure 3.3), the BS simply selects the codebook entry $\boldsymbol{Q}_{k_j^\star}$, cf. (3.10), associated with MT j that should be served for data transmission, and no further processing is required. In the multi-user mode (denoted as MU in Figure 3.3), the BS can represent each MT's channel using the subspace information associated with the respective $\bar{\boldsymbol{X}}_{k_j^\star}^{\mathrm{H}}$ extracted from the high SNR codebook, see (3.17). The BS then jointly designs the precoders for multi-user transmission using either non-iterative methods (RBD or RCI) or iterative approaches (WMMSE), with the choice impacting the required processing time. Alternatively, the BS can utilize the generative capabilities of the GMM and design precoders using the SWMMSE through sampling, as outlined in (3.22). For massive MIMO systems, the combined GNN- and GMM-based precoder design scheme proposed in Section 3.2.3 offers another option to reduce the computational time required for precoder design. The proposed versatile feedback scheme is illustrated in Figure 3.3. Green nodes represent steps executed at the BS, while blue nodes denote processing performed at the MTs.

This flexibility is not provided by the discussed baseline approaches since the

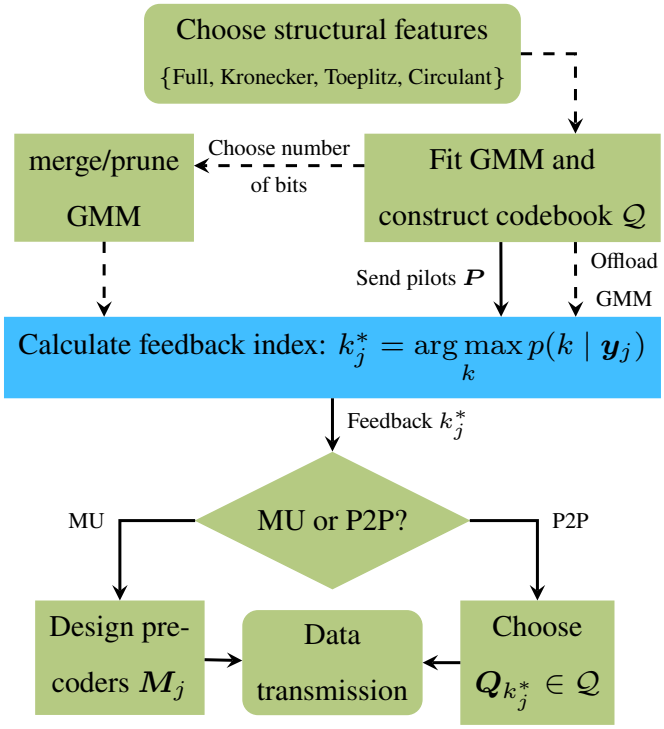


Figure 3.3: Flowchart of the versatile feedback scheme. Green (blue) colored nodes are processed at the BS (MTs), and solid (dashed) arrows indicate online (offline) processing. Figure adapted from [57] (©2023 IEEE).

feedback for the point-to-point mode associated with the current SNR is determined by selecting an element via (3.4) out of a codebook constructed for that SNR, see Section 3.1.1. In contrast, for the multi-user case, the codebook entry selection relies on a different codebook, such as the random quantization codebook, as outlined in (3.16), or the directional codebook, as described in (3.18).

3.3.2 Complexity Analysis

The online computational complexity of the proposed GMM-based precoder design scheme can be divided into two parts:

Inference of the Feedback Information: The calculation of responsibilities in (3.6) involves evaluating Gaussian densities. Since the GMM parameters remain unchanged for observations at fixed SNR levels, the inverse and the determinant of the densities can be pre-computed. Therefore, the online evaluation of the responsibilities in (3.6)

is dominated by matrix-vector multiplications, with a per-component complexity of $\mathcal{O}(N_{\mathrm{rx}}^2 n_{\mathrm{p}}^2)$. Correspondingly, determining the feedback using the GMM via (3.7) has a total complexity of $\mathcal{O}(K N_{\mathrm{rx}}^2 n_{\mathrm{p}}^2)$. One significant advantage is that the complexity does not scale with the number of transmit antennas N_{tx}. This is particularly beneficial in large-scale MIMO systems. Moreover, the proposed method allows for parallelization with respect to the number of components K, i.e., all of the K responsibilities can be evaluated in parallel. The proposed scheme avoids explicit channel estimation, unlike conventional methods where the complexity depends on both channel estimation and the selection of the best codebook entry using (3.4). Among the conventional approaches, the GMM estimator from (3.30) combined with the selection method from (3.4) performed best in simulations, see Section 3.4. Evaluating the GMM estimator from (3.30) involves a complexity of $\mathcal{O}(K N_{\mathrm{rx}}^2 n_{\mathrm{p}}^2 + K N_{\mathrm{rx}}^2 N_{\mathrm{tx}} n_{\mathrm{p}})$ which accounts for the responsibilities and the linear minimum mean square error (LMMSE) filters in (3.31) [33]. Notably, the responsibilities for feedback index selection using (3.7) are also used for the GMM estimator. However, the additional evaluation of LMMSE filters increases complexity, making the proposed method inherently less complex for the MT. Additionally, with conventional methods, the estimated channel must also be further processed to a feedback index via (3.4), which has a complexity of $\mathcal{O}(K N_{\mathrm{tx}} N_{\mathrm{rx}}^2 + K N_{\mathrm{rx}}^3)$ when using the QR decomposition [108, Section 5.2]. This analysis similarly holds for the multi-user mode.

Precoder Design: The computational complexity of the online precoder design depends on the transmission mode, i.e., single- or multi-user mode. In the single-user mode, the complexity is $\mathcal{O}(K)$, as it only involves traversing the pre-computed set of transmit covariance matrices $\mathcal{Q}$, see Section 3.1.2. In the multi-user mode, the computational complexity depends on the selected precoder design algorithm. Depending on latency requirements, the BS can either employ non-iterative algorithms (RBD or RCI), iterative approaches (WMMSE, SWMMSE), or, in massive MIMO systems, the parallelizable combined GNN- and GMM-based precoder design scheme from Section 3.2.3. The GNN module exhibits a computational complexity of $\mathcal{O}(\sum_{\ell=1}^{L} M_\ell M_{\ell-1})$ [39].

3.4 Simulation Results

Conventionally, the DL channel is estimated firstly at each MT, and subsequently, the best-fitting codebook entry is determined. Thus, in Section 3.4.1, the baseline channel estimators considered throughout the simulations are briefly discussed. In Section 3.4.2 we analyze point-to-point MIMO systems, while in Section 3.4.3 we

evaluate MU-MIMO systems. Finally, simulations for massive MIMO systems, i.e., systems with multiple single-antenna MTs, are presented in Section 3.4.4.

3.4.1 Baseline Channel Estimators

Since channel estimation is performed independently at each MT, we present the estimators from a MT's perspective, omitting the index j for brevity.

As a baseline, we adopt the recently proposed GMM-based channel estimator from [33], described in (3.30). This estimator is shown to asymptotically converge to the MSE-optimal conditional mean estimator (CME) as the number of GMM components K increases to infinity, provided that $\boldsymbol{A} = (\boldsymbol{P} \otimes \mathbf{I}_{N_{\mathrm{rx}}})$ is invertible, see (2.2), which requires $n_{\mathrm{p}} = N_{\mathrm{tx}}$. Even if $\boldsymbol{A}$ is not invertible ($n_{\mathrm{p}} < N_{\mathrm{tx}}$) and for a moderate K, the GMM-based channel estimator is highly effective as demonstrated in [33]. This estimator uses the same GMM utilized in the proposed GMM-based feedback scheme, see Section 3.1.2. Specifically, the MT uses the offloaded GMM to estimate the channel by computing:

$$\hat{\boldsymbol{h}}_{\mathrm{GMM}}(\boldsymbol{y}) = \sum_{k=1}^{K} p(k \mid \boldsymbol{y}) \hat{\boldsymbol{h}}_{\mathrm{LMMSE},k}(\boldsymbol{y}), \tag{3.30}$$

where $p(k \mid \boldsymbol{y})$ are the responsibilities from (3.6), and

$$\hat{\boldsymbol{h}}_{\mathrm{LMMSE},k}(\boldsymbol{y}) = \boldsymbol{C}_k \boldsymbol{A}^{\mathrm{H}} (\boldsymbol{A}\boldsymbol{C}_k\boldsymbol{A}^{\mathrm{H}} + \boldsymbol{\Sigma})^{-1} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}_k) + \boldsymbol{\mu}_k. \tag{3.31}$$

Thus, $\hat{\boldsymbol{h}}_{\mathrm{GMM}}$ is a convex combination of per-component LMMSE estimators.

Another baseline is the LMMSE estimator, which utilizes the sample covariance matrix $\boldsymbol{C}_s = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{h}^{(m)}\boldsymbol{h}^{(m),\mathrm{H}}$, calculated from the same training samples used for fitting the GMM (cf. [18, 33]):

$$\hat{\boldsymbol{h}}_{\mathrm{LMMSE}} = \boldsymbol{C}_s \boldsymbol{A}^{\mathrm{H}} (\boldsymbol{A}\boldsymbol{C}_s\boldsymbol{A}^{\mathrm{H}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{y}. \tag{3.32}$$

Lastly, we consider a compressive sensing approach for channel estimation using the orthogonal matching pursuit (OMP) [109] algorithm. This method finds a sparse vector $\boldsymbol{s}$, and constructs the estimated channel as

$$\hat{\boldsymbol{h}}_{\mathrm{OMP}} = \boldsymbol{O}\boldsymbol{s}, \tag{3.33}$$

where $\boldsymbol{O} = \boldsymbol{O}_{\mathrm{rx}} \otimes (\boldsymbol{O}_{\mathrm{tx,h}} \otimes \boldsymbol{O}_{\mathrm{tx,v}})$ is a dictionary formed from oversampled DFT matrices $\boldsymbol{O}_{\mathrm{rx}}$, $\boldsymbol{O}_{\mathrm{tx,h}}$, and $\boldsymbol{O}_{\mathrm{tx,v}}$ (cf., e.g., [110]), assuming a URA at the BS and a ULA at the MT. Since the algorithm's performance is sensitive to the sparsity order, which is unknown, we use a genie-aided approach as a performance bound, i.e., the true channel (perfect CSI) is used to determine the optimal sparsity order, as in [80].

3.4.2 Point-to-Point MIMO Systems

In Sections 3.4.2 and 3.4.3, we utilize channel data generated with QuaDRiGa, as detailed in Section 2.2.2. The BS equipped with a URA has in total $N_{\mathrm{tx}} = N_{\mathrm{tx,h}} N_{\mathrm{tx,v}}$ antenna elements, with $N_{\mathrm{tx,h}}$ horizontal and $N_{\mathrm{tx,v}}$ vertical elements. At the MT, we have a ULA with N_{rx} elements. We consider B feedback bits and thus $K = 2^B$. We generate datasets with $30 \cdot 10^3$ channels for both the UL and DL domain of the scenario described in Section 2.2.2: $\mathcal{H}^{\mathrm{UL}}$ and $\mathcal{H}^{\mathrm{DL}}$. The data samples are normalized such that $\mathrm{E}[\|\boldsymbol{h}\|^2] = N = N_{\mathrm{tx}} N_{\mathrm{rx}}$ holds for the vectorized channels. We further set $\rho = 1$, which allows us to define the SNR as $\frac{1}{\sigma_n^2}$. We split the two sets $\mathcal{H}^{\mathrm{UL}}$ and $\mathcal{H}^{\mathrm{DL}}$ into a training set with $M = 20 \cdot 10^3$ samples, and the remaining samples constitute an evaluation set, denoted by $\mathcal{H}^{\mathrm{UL}}_{\mathrm{train}}$, $\mathcal{H}^{\mathrm{UL}}_{\mathrm{eval}}$, $\mathcal{H}^{\mathrm{DL}}_{\mathrm{train}}$, and $\mathcal{H}^{\mathrm{DL}}_{\mathrm{eval}}$. The following transmit strategies are always evaluated on $\mathcal{H}^{\mathrm{DL}}_{\mathrm{eval}}$, i.e., in the DL domain. When we fit the GMM based on $\mathcal{H}^{\mathrm{UL}}_{\mathrm{train}}$, we transpose all elements of the set to emulate the DL domain, cf. [10, 16, 18, 20]. We fit GMMs with Kronecker structure if not mentioned otherwise, see Section 3.1.3.

In the single-user case, we depict the normalized spectral efficiency (NSE) as performance measure. The spectral efficiency achieved with a given transmit covariance matrix is normalized by the spectral efficiency achieved with the optimal transmit covariance matrix, which is given by decomposing the channel into N_{rx} parallel streams and employing water-filling [83]. The empirical complementary cumulative distribution function (cCDF) $P(\mathrm{NSE} > s)$ of the normalized spectral efficiency corresponds to the empirical probability that NSE exceeds a specific value s.

We consider the following baseline transmit strategies: The curves labeled "uni pow" represent uniform power allocation where the transmit covariance matrix is given by $\boldsymbol{Q} = \frac{\rho}{N_{\mathrm{tx}}} \mathbf{I}$. In this case, no CSI knowledge or codebook is used. Moreover, "uni pow eigsp" corresponds to the transmit strategy where a transmit covariance matrix is calculated by allocating equal power on the eigenvectors of the channel. That is, the channel is decomposed into N_{rx} parallel streams and $\frac{\rho}{N_{\mathrm{rx}}}$ power is allocated to each stream. Note that this approach is infeasible because the BS would require full knowledge of the DL channel (or its eigenvectors).

In the following, the simulation parameters are $N_{\mathrm{tx}} = 32$ ($N_{\mathrm{tx,h}} = 8$, $N_{\mathrm{tx,v}} = 4$), $N_{\mathrm{rx}} = 16$ and $B = 6$ bits, thus $K = 64$ ($K_{\mathrm{tx}} = 16$, $K_{\mathrm{rx}} = 4$). In Figure 3.4(a), we set $\mathrm{SNR} = 0\,\mathrm{dB}$. The conventional codebook construction approach, see Section 3.1.1, is denoted by "Lloyd UL/DL," depending on whether $\mathcal{H}^{\mathrm{UL}}_{\mathrm{train}}$ or $\mathcal{H}^{\mathrm{DL}}_{\mathrm{train}}$ is used as training data to construct the codebooks, respectively. With these approaches, the codebook is known to the BS and the MT and, additionally, perfect CSI is assumed at the MT. Each

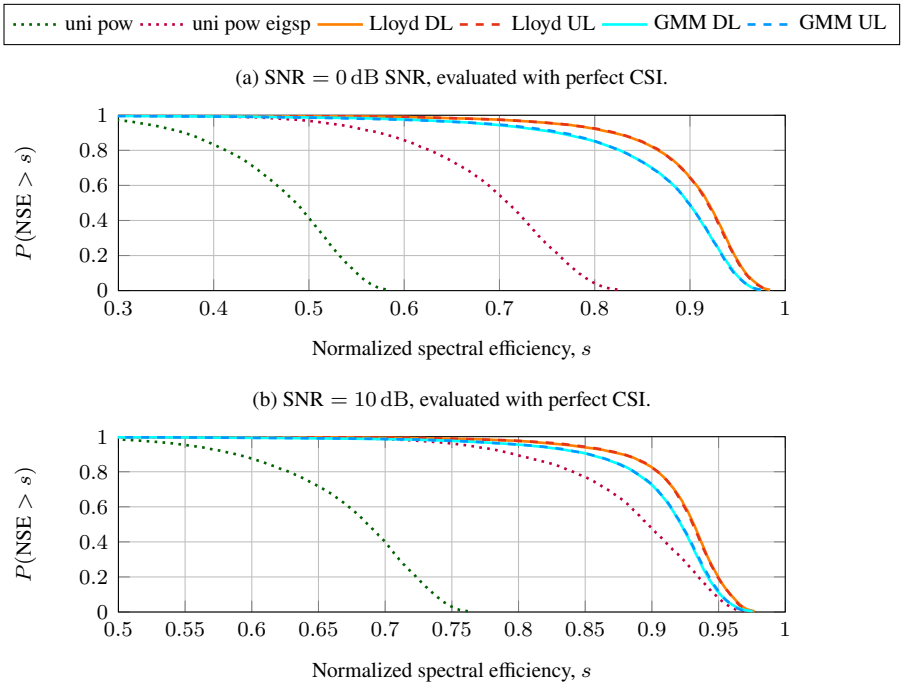


Figure 3.4: Empirical cCDFs of the normalized (by the optimal transmit strategy) spectral efficiencies achieved with different codebooks and transmit strategies evaluated with perfect CSI, for a system with $N_{\mathrm{tx}} = 32$, $N_{\mathrm{rx}} = 16$, and $B = 6$ bits. Figure adapted from [55] (©2022 IEEE) and [56].

MT then selects the best possible codebook entry by evaluating (3.4). We can observe that, using DL or UL training data, results in approximately the same performance. The proposed codebook construction and encoding scheme is denoted by "GMM UL/DL," where we either use $\mathcal{H}_{\mathrm{train}}^{\mathrm{UL}}$ or $\mathcal{H}_{\mathrm{train}}^{\mathrm{DL}}$ as training data to fit the GMM and to construct the codebook as described in Section 3.1.2. With our proposed approach, the knowledge of the codebook at the MT is not required. After offloading the GMM to the MT and given perfect CSI knowledge, the MT can then determine the feedback index by evaluating (3.8). Again, using DL or UL training data results in approximately the same performance, which is in accordance with the findings from [10, 16, 18, 20]. The proposed GMM approach performs slightly worse than the Lloyd clustering approach, with the perfect CSI assumption. In Figure 3.4(b), we set SNR $= 10\,\mathrm{dB}$, and observe similar results.

However, assuming perfect CSI at the MT is not feasible. In the following, we consider imperfect CSI, i.e., systems with reduced pilot overhead ($n_{\mathrm{p}} \leq N_{\mathrm{tx}}$). In the

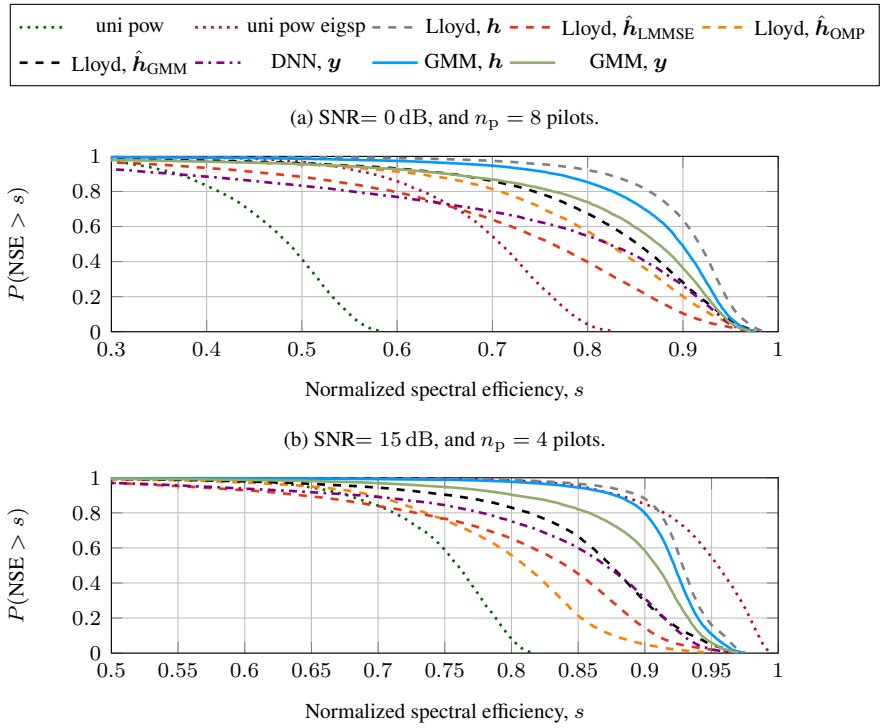


Figure 3.5: Empirical cCDFs of the normalized (by the optimal transmit strategy) spectral efficiencies achieved with different codebooks and transmit strategies evaluated for a system with $N_{\text{tx}} = 32$, $N_{\text{rx}} = 16$, different SNRs, different number of pilots n_{p}, and $B = 6$ bits. Figure adapted from [55] (©2022 IEEE) and [56].

remainder, we consider UL training data exclusively. Thus, we omit writing "UL" in the legend from now on.

In Figure 3.5(a), the SNR $= 0\,\text{dB}$ and we have $n_{\text{p}} = 8$. We depict results for the conventional Lloyd codebook construction approach, see Section 3.1.1, where we first estimate the channel either via OMP (3.33), the LMMSE approach (3.32), or via the GMM estimator (3.30), and then select a transmit covariance matrix by evaluating (3.4) given the estimated channel: "Lloyd, $\hat{\boldsymbol{h}}_{\text{OMP}}$," "Lloyd, $\hat{\boldsymbol{h}}_{\text{LMMSE}}$," or "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$," respectively. Moreover, we compare to the SNR-independent DNN approach from [10], denoted by "DNN $\boldsymbol{y}$," where a classifier is employed to directly map the observation to a feedback index that specifies an element from the Lloyd codebook. During the training phase, the DNN was provided with input-output pairs for an SNR range of $0\,\text{dB}$ to $25\,\text{dB}$, with $5\,\text{dB}$ steps. We employ random search [111] to determine the hyperparameters of the DNN. The DNN consists of D_{CM}

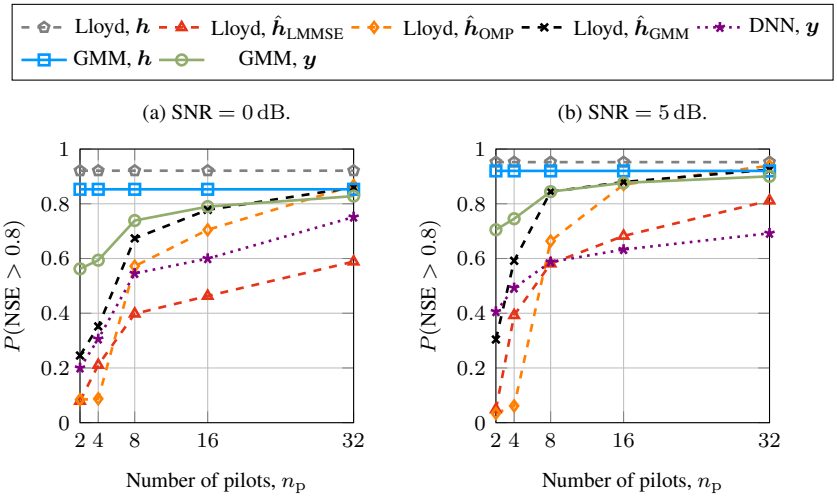


Figure 3.6: The probability that the NSE of a certain transmit strategy exceeds $s = 80\%$ of the optimal transmit strategy's spectral efficiency for a varying number of pilots, for a system with $N_{\text{tx}} = 32$, $N_{\text{rx}} = 16$, and $B = 6$ bits. Figure adapted from [55] (©2022 IEEE) and [56].

convolutional modules, which comprise a convolutional layer, a batch normalization, and an activation function, where D_{CM} is randomly chosen from the range $[2, 5]$. Each of the convolutional layers consists of D_{K} kernels, where D_{K} is randomly chosen within $[16, 64]$. After a subsequent two-dimensional max-pooling, the features are flattened, and a fully connected layer is employed with an output dimension of K. Depending on the randomly drawn parameters, the number of DNN parameters is at least the same as or even higher than the number of GMM parameters, and increases with the number of pilots. Moreover, note that a separate DNN per pilot configuration is needed. The complexity of the DNN approach is $\mathcal{O}(KD_{\text{K}}N_{\text{rx}}n_{\text{p}})$. Further details can be found in [10]. As can be seen, estimating the channel via the GMM estimator gives the best performance when considering the conventional channel estimation-based approaches. The DNN approach, which also does not require any codebook knowledge at the MT, similar to our proposed GMM-based feedback scheme, achieves comparable performance as "Lloyd, $\hat{\boldsymbol{h}}_{\text{OMP}}$" or "Lloyd, $\hat{\boldsymbol{h}}_{\text{LMMSE}}$." In contrast, with the proposed approach, see Section 3.1.2, denoted by "GMM, $\boldsymbol{y}$," where we bypass channel estimation and directly evaluate (3.7) for determining a feedback index, we achieve even better performance as compared to any of the conventional approaches. With the curves "Lloyd, $\boldsymbol{h}$" and "GMM, $\boldsymbol{h}$," we depict the results for the case of

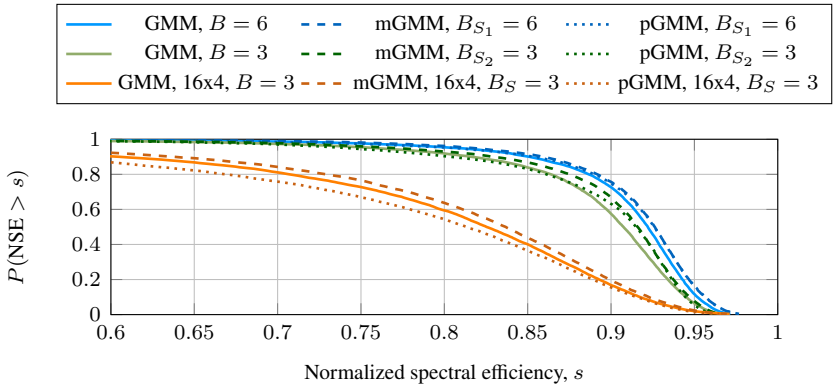


Figure 3.7: Empirical cCDFs of the normalized spectral efficiencies for two setups using either mixture reduction techniques or the direct fitting approach assuming perfect CSI. Setup A: $N_{\text{tx}} = 32$ ($N_{\text{tx,h}} = 8, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 16$, and SNR $= 10\,\text{dB}$. Setup B: $N_{\text{tx}} = 16$ ($N_{\text{tx,h}} = 4, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 4$, and SNR $= 0\,\text{dB}$. Figure adapted from [57] (©2023 IEEE).

assuming perfect CSI knowledge at the MT. Although the Lloyd approach performs well if perfect CSI is available at the MT, the performance suffers significantly from CSI imperfections (due to noise and low pilot overhead). In contrast, the proposed GMM-based feedback scheme is superior in case of imperfect CSI available at the MT and thus is a robust solution for practical system deployments. A similar observation can also be made in Figure 3.5(b), where the SNR $= 15\,\text{dB}$ and we only have $n_{\text{p}} = 4$ pilots.

In Figure 3.6(a), we set SNR $= 0\,\text{dB}$ and in Figure 3.6(b), we have SNR $= 5\,\text{dB}$, where we fix $s = 0.8$, thus, we consider $P(\text{NSE} > 0.8)$ for a varying number of pilots n_{p}. We see that our proposed low-complexity feedback scheme is beneficial in the low number of pilots regime and outperforms the conventional approaches, which either require both channel estimation and the feedback evaluation via (3.4) or the DNN-based approach.

Next, we consider simulations involving mixture reduction techniques, which enable variable feedback bit lengths using the GMM, see Section 3.1.4. In Figure 3.7, we denote by "Setup A" a system with $N_{\text{tx}} = 32$ ($N_{\text{tx,h}} = 8, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 16$, and SNR $= 10\,\text{dB}$, and with "Setup B" a system with $N_{\text{tx}} = 16$ ($N_{\text{tx,h}} = 4, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 4$, and SNR $= 0\,\text{dB}$. We assume perfect CSI, i.e., the feedback information is determined via (3.8). For "Setup A" we fit a Kronecker-structured GMM with $K = 256$ ($B = 8$, $K_{\text{tx}} = 32$, and $K_{\text{rx}} = 8$) components and first reduce to a GMM with $K_{S_1} = 64$ ($B_{S_1} = 6$) components and continue the reduction to another one

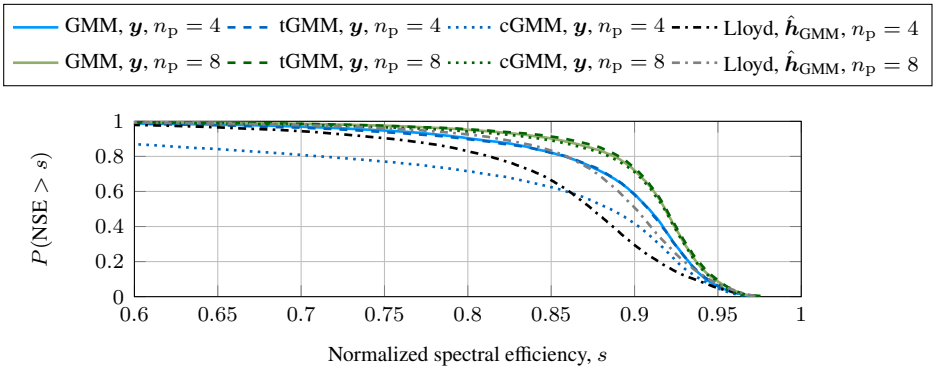
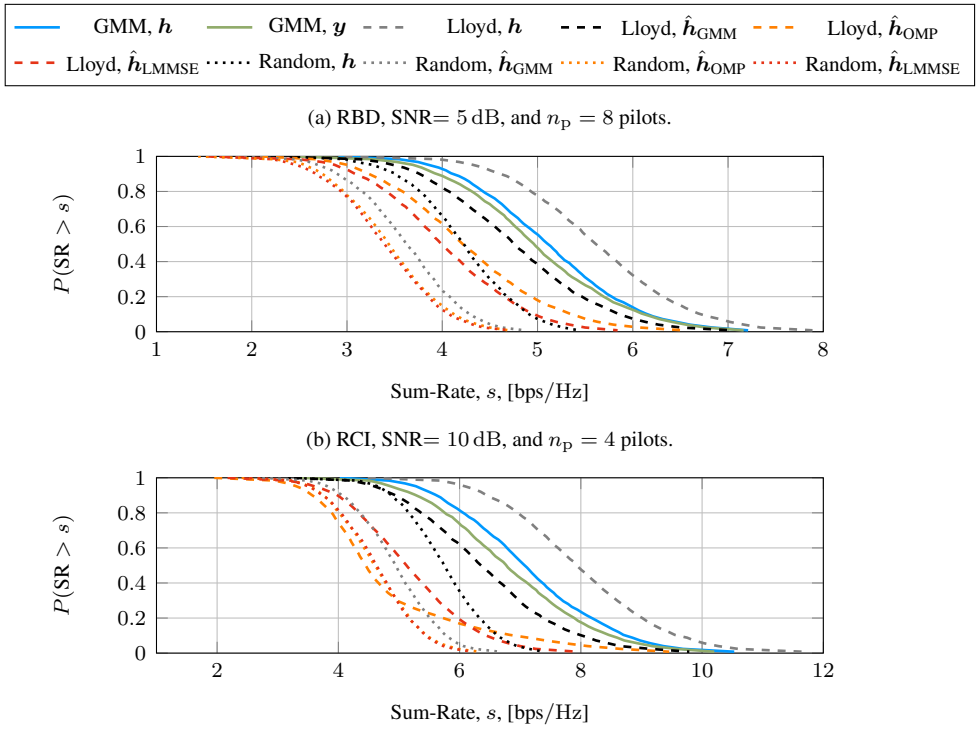


Figure 3.8: Empirical cCDFs of the normalized spectral efficiencies using differently structured GMMs for a system with $N_\mathrm{tx} = 32$, $N_\mathrm{rx} = 16$, SNR $= 15\,\mathrm{dB}$, different number of pilots n_p, and $B = 6$ bits. Figure adapted from [57] (©2023 IEEE).

with $K_{S_2} = 8$ ($B_{S_2} = 3$) components, by either applying the merging strategy using Algorithm 1, denoted as "mGMM," or the pruning strategy using Algorithm 2, denoted by "pGMM." Moreover, with "GMM" we denote the case of directly fitting a GMM with $K = 64$ ($B = 6$, $K_\mathrm{tx} = 16$, and $K_\mathrm{rx} = 4$), or with $K = 8$ ($B = 3$, $K_\mathrm{tx} = 4$, and $K_\mathrm{rx} = 2$) components. We can observe that both mixture reduction approaches not only enable the feedback scheme to support variable bit lengths but, at the same time, provide at least a similar performance as the respective direct fitting approaches. In case of "Setup B," highlighted by "16x4" in the legend of Figure 3.7, we fit a GMM with $K = 256$ components and reduce to one with $K_S = 8$ ($B_S = 3$) components and compare to a GMM directly fitted with $K = 8$ ($K_\mathrm{tx} = 4$, and $K_\mathrm{rx} = 2$) components. In this setup, the merging strategy performs best, whereas the low-complexity pruning strategy yields a slightly worse performance as compared to the direct fitting approach. Altogether, the applied mixture reduction techniques enable variable bit lengths and, in some cases, slightly improve the performance as a consequence of the enhanced clustering ability, cf. [98].

In Figure 3.8, we have $N_\mathrm{tx} = 32$ ($N_\mathrm{tx,h} = 8, N_\mathrm{tx,v} = 4$), $N_\mathrm{rx} = 16$, SNR $=$ $15\,\mathrm{dB}$, and $n_\mathrm{p} \in \{4, 8\}$ pilots and depict differently structured GMMs with $K = 64$ ($K_\mathrm{tx} = 16$, and $K_\mathrm{rx} = 4$) components. With "tGMM" or "cGMM" we denote Kronecker GMMs constructed by (block) Toeplitz or (block) circulant transmit- and receive-side GMMs, respectively. The approach "tGMM," with a reduced number of parameters, achieves the same performance as compared to the Kronecker GMM with full transmit- and receive-side GMMs denoted by "GMM," irrespective of the number of pilots. With $n_\mathrm{p} = 4$, "tGMM" attains a similar performance as the conventional



(a) RBD, SNR= 5 dB, and $n_{\text{p}} = 8$ pilots.

(b) RCI, SNR= 10 dB, and $n_{\text{p}} = 4$ pilots.

Figure 3.9: Empirical cCDFs of the sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) achieved with different feedback approaches and precoding techniques for different SNRs, different number of pilots n_{p}, and $B = 6$ bits. Figure adapted from [56].

Lloyd approach with twice as many pilots, denoted by "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$, $n_{\text{p}} = 8$." This shows the great potential of the enhanced GMM-based feedback scheme in systems with reduced pilot overhead. Interestingly, in the case of the circulant approximation "cGMM," a performance degradation can be observed if $n_{\text{p}} = 4$. It seems that the expressivity of the circulant GMM with a very low number of parameters is too restrictive to be applied for systems with a few pilots only.

3.4.3 Multi-user MIMO Systems

In this section, we evaluate MU-MIMO systems and use the sum-rate as the performance measure, which is given by $\sum_{j=1}^{J} R_j^{\text{inst}}$, cf. (2.6). We depict the results for 2,500 scenarios, where for each scenario, we draw J MTs randomly from our evaluation set $\mathcal{H}_{\text{eval}}^{\text{DL}}$. The empirical cCDF $P(\text{SR} > s)$ of the sum-rate, is used to depict the empirical probability that the sum-rate (SR) exceeds a specific value s.

In the following, with "GMM, $\boldsymbol{h}$" and "GMM, $\boldsymbol{y}$" we denote the cases, where either perfect CSI $\boldsymbol{h}_j$ is assumed or the observations $\boldsymbol{y}_j$ are used at each MT j to determine a feedback index using the GMM feedback encoding approach, cf. (3.19). We omit the index j in the legend for notational convenience. The channel of each MT is then represented by the subspace information extracted from the high-SNR GMM codebook, see Section 3.2.2.1. With "Lloyd, $\boldsymbol{h}$," "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$," "Lloyd, $\hat{\boldsymbol{h}}_{\text{OMP}}$," and "Lloyd, $\hat{\boldsymbol{h}}_{\text{LMMSE}}$," or with "Random, $\boldsymbol{h}$," "Random, $\hat{\boldsymbol{h}}_{\text{GMM}}$," "Random, $\hat{\boldsymbol{h}}_{\text{OMP}}$," and "Random, $\hat{\boldsymbol{h}}_{\text{LMMSE}}$," we denote the cases where either perfect CSI is assumed at each MT, or the channel is firstly estimated at each MT and then the index of the best fitting subspace entry of the high-SNR Lloyd codebook or of the random codebook, is fed back from each MT to the BS, see Sections 3.2.2.1 and 3.2.1.

The above-mentioned approaches are evaluated using either RBD, RCI, or the iterative WMMSE to jointly design the precoders $\boldsymbol{M}_j,\ \forall j \in \mathcal{J}$. In the case of RBD and RCI, the used regularization factor is $\frac{JN_{\text{rx}}\sigma_n^2}{\rho}$, and the precoders are normalized to satisfy the transmit power constraint, cf. [86–88]. In the case of the iterative WMMSE, we use [90, Algorithm 1]. Additionally, with "GMM samples, $\boldsymbol{h}$" and "GMM samples, $\boldsymbol{y}$," we denote the cases where we generate samples that represent each MT's distribution using the most responsible component of the GMM and feed them to the SWMMSE algorithm, see Section 3.2.2.2. In all iterative approaches, we set $I_{\max} = 300$ iterations.

In the following, the simulation parameters are $N_{\text{tx}} = 16$ ($N_{\text{tx,h}} = 4, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 4$ and $B = 6$ bits, thus $K = 64$ ($K_{\text{tx}} = 16$, $K_{\text{rx}} = 4$). Accordingly, we have $J = 4$ MTs. In Figure 3.9(a), the SNR $= 5\,\text{dB}$ and we have $n_{\text{p}} = 8$ pilots. In this case, we use RBD in order to jointly design the precoders. We can observe that the random codebook performs worst. Even with perfect CSI assumed at the MTs, the random codebook approach cannot compete with the environment-aware approaches. The Lloyd directional codebook approach yields the best performance, if the GMM-based channel estimator from (3.30) is used prior to codebook entry selection. Similar to the point-to-point MIMO case, we can observe that the chosen channel estimator significantly impacts the performance. That is, using the LMMSE estimator from (3.32) or the genie OMP from (3.33) yield worse results as compared to the GMM-based channel estimator. Furthermore, our proposed GMM-based feedback approach "GMM, $\boldsymbol{y}$" even outperforms the best conventional approach "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$." A similar trend is observed in Figure 3.9(b), where the SNR is increased to $10\,\text{dB}$, the number of pilots is reduced to $n_{\text{p}} = 4$, and RCI is employed for precoder design.

In the following two figures, Figure 3.10 and Figure 3.11, we restrict our analysis to RBD as the precoder design algorithm for sake of brevity. The purpose of the next

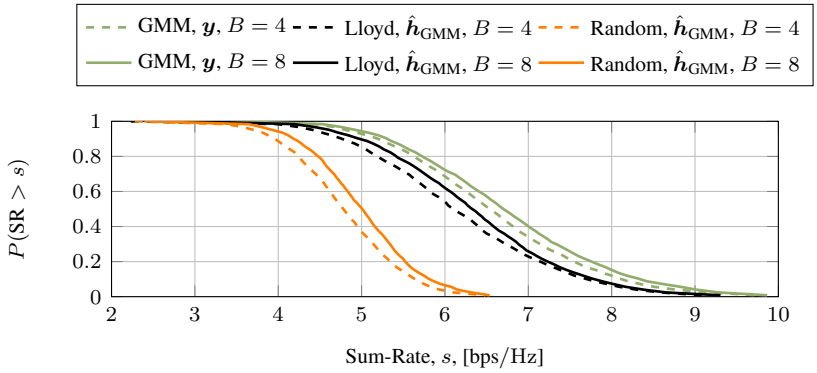


Figure 3.10: Empirical cCDFs of the sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) achieved with different feedback approaches when RBD is employed for a system with an SNR $= 10\,\text{dB}$, $n_{\text{p}} = 4$ pilots, and different bits B. Figure adapted from [56].

two figures is to quantify the performance gains obtained with our proposed approach from different perspectives. In particular, in Figure 3.10 we have SNR $= 10\,\text{dB}$ and $n_{\text{p}} = 4$, and we consider systems with $B = 4$ bits, thus, $K = 16$ ($K_{\text{tx}} = 8$, $K_{\text{rx}} = 2$), or $B = 8$ bits, thus $K = 256$ ($K_{\text{tx}} = 32$, $K_{\text{rx}} = 8$) and compare the performance of our proposed GMM-based feedback approach "GMM, $\boldsymbol{y}, B \in \{4, 8\}$" to the best performing conventional Lloyd directional codebook "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}, B \in \{4, 8\}$" and random codebook "Random, $\hat{\boldsymbol{h}}_{\text{GMM}}, B \in \{4, 8\}$" approaches, which use the GMM-based channel estimator in the channel estimation phase. We can observe that our proposed feedback approach is superior to the conventional methods. In particular, with only $B = 4$ bits "GMM, $\boldsymbol{y}, B = 4$" even outperforms the conventional approach "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}, B = 8$" with twice as many bits.

In Figure 3.11, we consider a system with more transmit antennas and, accordingly, more MTs. The simulation parameters are $N_{\text{tx}} = 64$ ($N_{\text{tx,h}} = 8, N_{\text{tx,v}} = 8$), $N_{\text{rx}} = 4$, and $B = 6$ bits, thus, $K = 64$ ($K_{\text{tx}} = 16$, $K_{\text{rx}} = 4$), and the SNR $= 5\,\text{dB}$. Accordingly, we have $J = 16$ MTs. This time, we depict the performances for a varying number of pilots n_{p}. For a fixed number of pilots n_{p}, our proposed approach "GMM, $\boldsymbol{y}, n_{\text{p}} \in \{2, 6, 12\}$" outperforms the conventional approaches "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}, n_{\text{p}} \in \{2, 6, 12\}$" and "Lloyd, $\hat{\boldsymbol{h}}_{\text{OMP}}, n_{\text{p}} = 12$." Moreover, with only $n_{\text{p}} = 2$ pilots, "GMM, $\boldsymbol{y}, n_{\text{p}} = 2$," almost achieves the same performance as the conventional approaches, which require $n_{\text{p}} = 6$ pilots in the case of "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}, n_{\text{p}} = 6$," or even $n_{\text{p}} = 12$ pilots in the case of "Lloyd, $\hat{\boldsymbol{h}}_{\text{OMP}}, n_{\text{p}} = 12$" (i.e., the MTs are unaware of the GMM and use the OMP channel estimator). When random

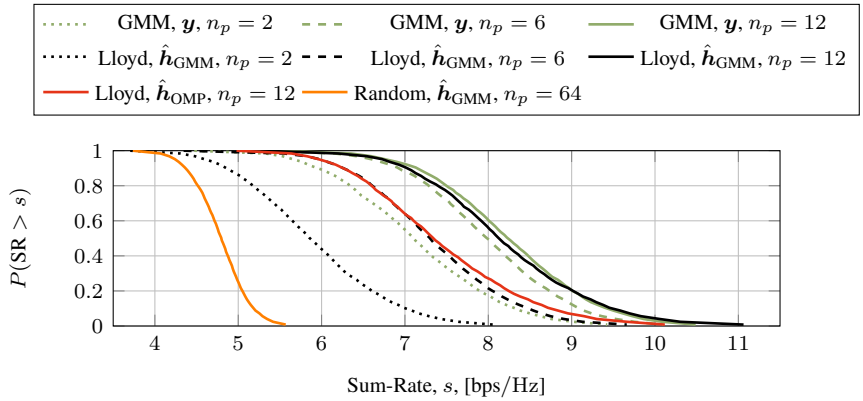


Figure 3.11: Empirical cCDFs of the sum-rate ($N_{\text{tx}} = 64$, $N_{\text{rx}} = 4$, and $J = 16$ MTs) achieved with different feedback approaches when RBD is employed for systems with an SNR $= 5\,\text{dB}$, different number of pilots n_{p}, and $B = 6$ bits. Figure adapted from [56].

codebooks are used, the performance remains poor even with a large pilot overhead, i.e. $n_{\text{p}} = 64$ pilots ("Random, $\hat{\boldsymbol{h}}_{\text{GMM}}, n_{\text{p}} = 64$"). Thus, with our proposed approach, systems with lower pilot overhead can be deployed, which would inherently increase the system throughput. Additionally, with fewer pilots, the complexity of determining the feedback index at the MTs with the proposed GMM-based approach decreases, see Section 3.3.2.

So far, we have only considered non-iterative precoding algorithms, i.e., RBD and RCI. In the remainder, we will focus our analysis on the iterative WMMSE and the SWMMSE precoding techniques. Due to the exclusive usage of channel directional information, i.e., no channel magnitude information is fed back to the BS, as in [7], see Section 3.2.1, in case of random codebooks, or the Lloyd directional codebook from Section 3.2.2.1, or the GMM directional codebook from Section 3.2.2.1, changing the number of streams d impacts the overall sum-rate which can be obtained using the iterative WMMSE. In fact, we observed that depending on the SNR, the number of pilots n_{p}, the chosen channel estimator (for the conventional approaches), and the selected codebook, the performance can be improved by varying $d \in \{1, 2, \cdots, N_{\text{rx}}\}$, and then setting d to the value which gives the best average performance. In a practical deployment, the parameter d can be pre-adjusted in the offline phase at the BS by emulating a DL system (for example using the set $\mathcal{H}^{\text{UL}}_{\text{eval}}$). However, the situation differs for the SWMMSE. In this case, we observed that setting $d = N_{\text{rx}}$ consistently yields the best performance. Intuitively, due to the sampling process involved in the design

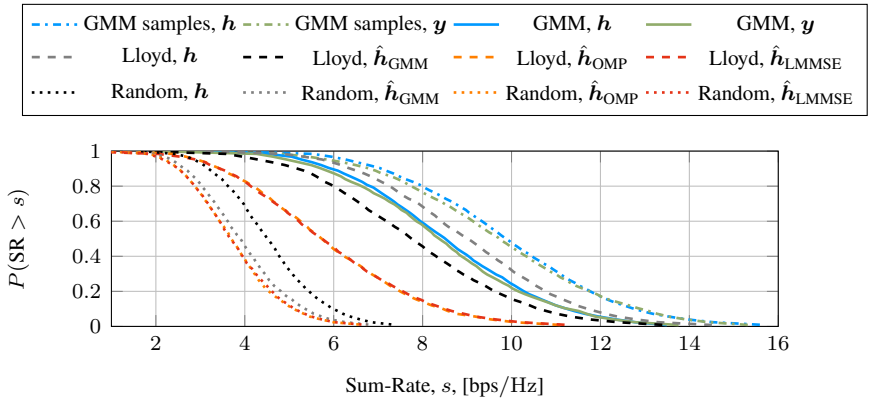


Figure 3.12: Empirical cCDFs of the sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) achieved with different feedback approaches when the iterative WMMSE or the SWMMSE are employed, for a system with an SNR $= 5\,\text{dB}$, $n_{\text{p}} = 8$ pilots, and $B = 6$ bits. Figure adapted from [56].

procedure of the SWMMSE, (average) channel magnitude information is provided to and exploited by the SWMMSE algorithm to appropriately adjust stream powers.

In the remainder of this section, the simulation parameters are $N_{\text{tx}} = 16$ ($N_{\text{tx,h}} = 4, N_{\text{tx,v}} = 4$), $N_{\text{rx}} = 4$, yielding $J = 4$ MTs, and $B = 6$ bits, thus, $K = 64$ ($K_{\text{tx}} = 16$, $K_{\text{rx}} = 4$). In Figure 3.12, the SNR $= 5\,\text{dB}$ and we have $n_{\text{p}} = 8$ pilots. This is the same simulation setting as in Figure 3.9(a), where RBD was used in order to design the precoders. By comparing Figure 3.12 and Figure 3.9(a), we can conclude that by using the iterative precoding techniques, the performances of the conventional and the proposed feedback approaches are improved tremendously. We can observe, that also in the case of iterative precoding techniques, the random codebook approach performs worst. Also in this case, the Lloyd directional codebook approach yields the best performance, if the GMM-based channel estimator from (3.30) is used prior to codebook entry selection whereas using the LMMSE estimator from (3.32) or the genie OMP from (3.33) deteriorates the performance. Furthermore, our proposed low-complexity GMM-based feedback approach "GMM, $\boldsymbol{y}$" again outperforms the best conventional approach "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$." With our generative modeling-based approach from Section 3.2.2.2, i.e., the SWMMSE with samples generated by the GMM, denoted by "GMM samples, $\boldsymbol{h}$" or "GMM samples, $\boldsymbol{y}$," we even outperform the performance bound of the Lloyd directional codebook approach "Lloyd, $\boldsymbol{h}$" (which uses the iterative WMMSE) with perfect CSI assumed at the MTs. This shows the great potential of the generative modeling ability of the GMM.

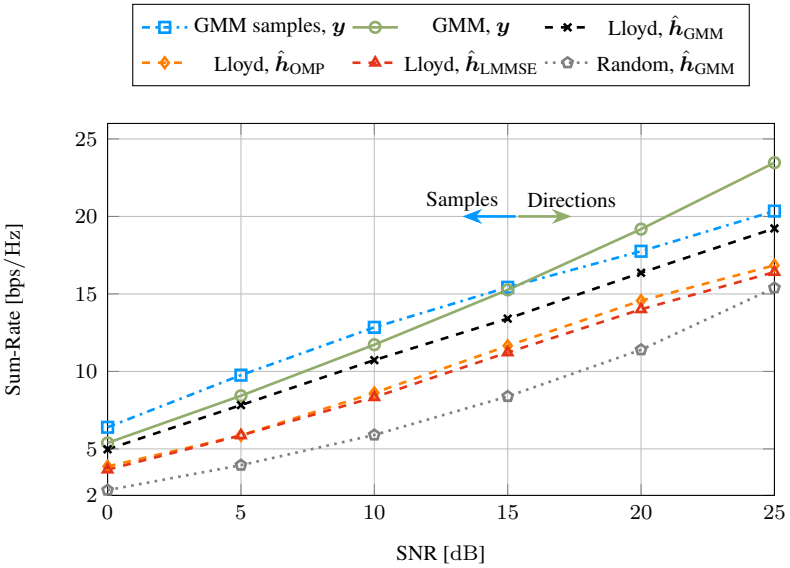


Figure 3.13: The average sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) over the SNR achieved with different feedback approaches when the iterative WMMSE or the SWMMSE are employed, for $n_{\text{p}} = 8$ pilots, and $B = 6$ bits. Figure adapted from [56].

In Figure 3.13, we also consider a setting with $n_{\text{p}} = 8$ pilots but vary the SNR. We depict the sum-rate averaged over all scenarios. We can see that our proposed GMM-based feedback approach, with either exploiting directional information "GMM, $\boldsymbol{y}$" or the generative modeling-based approach "GMM samples, $\boldsymbol{y}$" outperform the conventional approaches. We can observe, that in this case, up to an SNR of approximately $15\,\text{dB}$, the generative modeling-based method performs better, and for larger SNR values, the directional approach is superior. This is illustrated by the blue and green arrows in Figure 3.13. Thus, the results suggest that jointly designing the precoders by solving the ergodic sum-rate maximization problem from (3.21) by utilizing the SWMMSE, is beneficial for low to medium SNR values, and the directional approach, which exploits the iterative WMMSE, is superior for larger SNR values.

This observation is also supported by the results in Figure 3.14, where we depict the performances of our proposed approaches "GMM, $\boldsymbol{y}$" and "GMM samples, $\boldsymbol{y}$" and compare them to the best performing conventional method "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$," and the random codebook-based approach which uses the OMP estimator "Random, $\hat{\boldsymbol{h}}_{\text{OMP}}$" (i.e., no environment awareness) for a varying SNR and $n_{\text{p}} \in \{2, 8, 16\}$. There, dotted curves represent $n_{\text{p}} = 2$, dashed curves $n_{\text{p}} = 8$, and solid curves $n_{\text{p}} = 16$ pilots. The

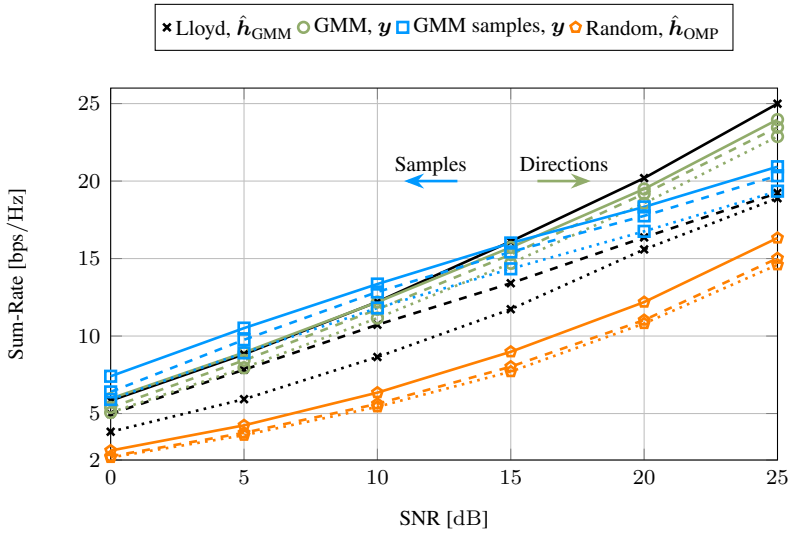


Figure 3.14: The average sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) over the SNR achieved with different feedback approaches using the iterative WMMSE or the SWMMSE, for different number of pilots n_{p} (dotted: $n_{\text{p}} = 2$, dashed: $n_{\text{p}} = 8$, solid: $n_{\text{p}} = 16$), and $B = 6$ bits. Figure adapted from [56].

approaches with no environment awareness, i.e., "Random, $\hat{\boldsymbol{h}}_{\text{OMP}}$" perform worst. Both of our proposed approaches, i.e., "GMM, $\boldsymbol{y}$" and "GMM samples, $\boldsymbol{y}$," with only $n_{\text{p}} = 2$ pilots, outperform the conventional "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$" approach with four times more, i.e., $n_{\text{p}} = 8$, pilots. When the number of pilots is equal to the number of transmit antennas (large pilot overhead), i.e., $n_{\text{p}} = N_{\text{tx}} = 16$, our proposed approaches which solely require the GMM at the MTs, at least can compete with the conventional "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$" approach, which requires both the GMM and the Lloyd directional codebook at each MT. For SNRs up to about $15\,\text{dB}$ our proposed generative modeling-based approach even outperforms the conventional method based on the Lloyd directional codebook.

In Figure 3.15, we depict how the sum-rate evolves over the number of iterations in the case of the iterative WMMSE (solid curves), or over the number of drawn samples (per MT) in the case of the SWMMSE (dashed curves) for a setting with an SNR $= 5\,\text{dB}$ and $n_{\text{p}} = 8$ pilots. In comparison, we depict the performance of the case, where we applied RBD (dotted curves) to jointly design the precoders. Since RBD is a non-iterative approach, the respective curves are constant over the iterations. We can observe that in the case of random codebooks "Random, $\hat{\boldsymbol{h}}_{\text{GMM}}$" the performance

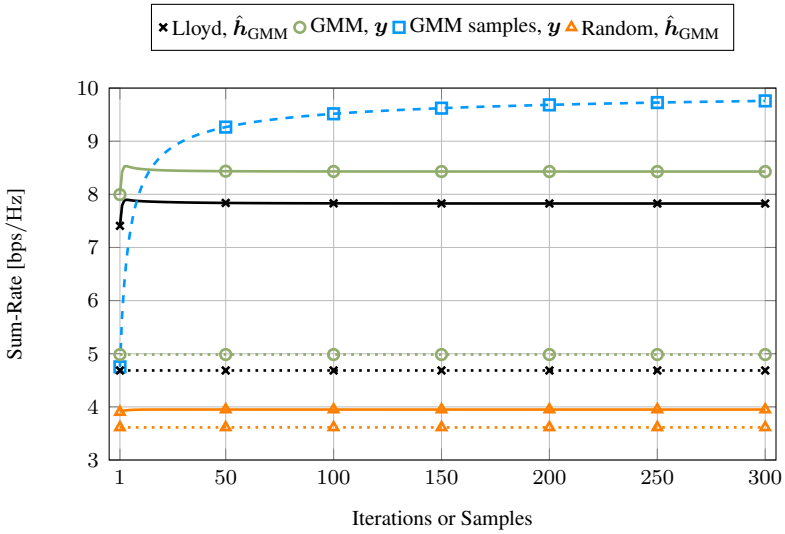


Figure 3.15: The average sum-rate ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, and $J = 4$ MTs) over the number of iterations/samples achieved with different feedback approaches and precoding techniques (dotted: RBD (remains constant, since non-iterative), dashed: SWMMSE, solid: WMMSE) for a system with an SNR $= 5\,\text{dB}$, $n_{\text{p}} = 8$ pilots, and $B = 6$ bits. Figure adapted from [56].

gains achieved by using the iterative WMMSE, compared to using RBD, are relatively small. In contrast, using the "Lloyd, $\hat{\boldsymbol{h}}_{\text{GMM}}$" or "GMM, $\boldsymbol{y}$" directional codebook approaches, we obtain huge performance gains when we use the iterative precoding techniques. In these cases, a small number of iterations is already enough to reach the maximum performance. Interestingly, a small overshoot can be observed. This artifact is possibly due to the fact that the iterative WMMSE is designed for perfect CSI, but here we are restricted to using quantized directional information due to the limited feedback. In contrast, when we use the generative modeling-based approach "GMM samples, $\boldsymbol{y}$" we can observe that the performance steadily improves over the number of drawn samples. We observed this behavior consistently for different SNR values and numbers of pilots.

In the remainder of this section, we combine both, the Toeplitz approximation and the mixture reduction techniques, in a setup with $N_{\text{tx}} = 16$ ($N_{\text{tx,h}} = 4, N_{\text{tx,v}} = 4$) and $N_{\text{rx}} = 4$, and $J = 4$ MTs. In particular, we fit a Kronecker GMM constructed by (block) Toeplitz transmit- and receive-side GMMs, with $K = 256$ components and reduce to a GMM with $K_S = 64$ ($B_S = 6$) components denoted by "mtGMM"

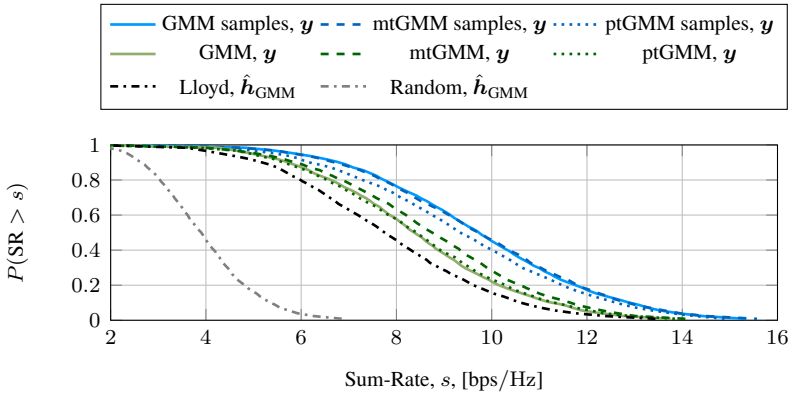


Figure 3.16: Empirical cCDFs of the sum-rate using either mixture reduction techniques ($B = 8$ to $B_S = 6$) combined with Toeplitz structured GMMs or the direct fitting approach ($B = 8$) when the iterative WMMSE or the SWMMSE are utilized, for a system with $N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, $J = 4$ MTs, SNR $= 5\,\text{dB}$, and $n_{\text{p}} = 8$ pilots. Figure adapted from [57] (©2023 IEEE).

(merged) and "ptGMM" (pruned), and compare to a GMM directly fitted with 64 ($K_{\text{tx}} = 16$, and $K_{\text{rx}} = 4$) components with full transmit- and receive-side GMMs, denoted by "GMM." In Figure 3.16, with SNR $= 5\,\text{dB}$ and $n_{\text{p}} = 8$ pilots, we can see that the pruning and merging strategy in combination with the Toeplitz structured covariances in the case of the subspace approach, i.e., "ptGMM, $\boldsymbol{y}$" and "mtGMM, $\boldsymbol{y}$," perform similarly or even better than the "kGMM, $\boldsymbol{y}$" and outperform the conventional methods "{Lloyd, Random}, $\hat{\boldsymbol{h}}_{\text{GMM}}$." With the generative modeling approach, the merging method "mtGMM samples, $\boldsymbol{y}$" performs equally well as "GMM samples, $\boldsymbol{y}$," whereas the pruning method "ptGMM samples, $\boldsymbol{y}$" performs slightly worse.

A similar behaviour is present in Figure 3.17, where we consider the same setting with $n_{\text{p}} = 8$ pilots, but vary the SNR and depict the sum-rate averaged over all scenarios. As also observed in Figure 3.13, adopting the generative modeling approach for joint precoder design is advantageous in scenarios characterized by low to moderate SNR levels, and for larger SNR values the directional approach is beneficial. This property can be similarly observed if the merging or pruning procedures are applied, and is illustrated by the arrows in Figure 3.17. Altogether, the enhanced GMM-based feedback scheme exhibits superior performance compared to the baselines in a multi-user system with reduced pilot overhead.

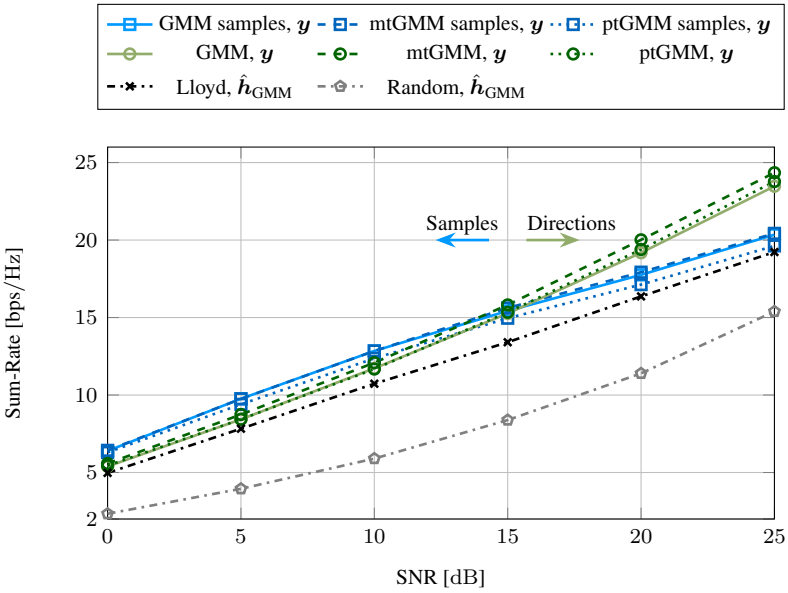


Figure 3.17: The average sum-rate over the SNR using either mixture reduction techniques ($B = 8$ to $B_S = 6$) combined with Toeplitz structured GMMs or the direct fitting approach ($B = 6$) when the iterative WMMSE or the SWMMSE are employed, for a system with $N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$, $J = 4$ MTs, and $n_{\text{p}} = 8$ pilots. Figure adapted from [57] (©2023 IEEE).

3.4.4 Massive MIMO Systems

Next, we consider massive MIMO systems and evaluate the performance of the GMM-based feedback scheme combined with the SWMMSE for precoder design and compare to the combined GNN- and GMM-based scheme from Section 3.2.3. We consider either the spatial channel model, see Section 2.2.1, or real-world measurement data, see Section 2.2.3. For both data sources, the channels are normalized to satisfy $\mathrm{E}[\|\boldsymbol{h}\|^2] = N$. Additionally, we set $\rho = 1$, enabling the SNR to be defined as $\frac{1}{\sigma_n^2}$.

Due to MTs being equipped with single antennas only, as a conventional baseline solution, we similarly apply the baseline channel estimators from Section 3.4.1 but consider now well-established DFT codebook-based approaches [112–114]. Based on the estimated channel, each MT determines the directional information as (see, e.g., [115]):

$$k_j^\star = \arg\max_k |\boldsymbol{d}_k^{\mathrm{H}} \hat{\boldsymbol{h}}_j|, \tag{3.34}$$

where $\boldsymbol{d}_k \in \mathcal{D}$, with the codebook $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{K_{\text{Dir}}}\}$ of cardinality $|\mathcal{D}| = K = 2^B$. Depending on whether a ULA or URA is deployed at the BS, a DFT codebook [112]

or a 2D-DFT codebook [113] is used, respectively. Given the feedback from each MT, the BS represents the channel of each MT as [115],

$$\hat{\boldsymbol{h}}_j^{(q)} = \boldsymbol{d}_{k_j^\star}, \tag{3.35}$$

and uses the iterative WMMSE [90, Algorithm 1] to jointly design the precoders.

In the case of the spatial channel model, see Section 2.2.1, we simulate a BS with $N = N_{\text{tx}} = 64$ antenna elements, assuming a single propagation cluster composed of multiple sub-paths per MT. To train the GNN using perfect statistical knowledge, we generate the set $\mathcal{D}_{\text{genie}}$ with $D = 2{,}400$ scenarios and $J = 16$ MTs. For training the GMM, we create a set $\mathcal{H}$ with $M = 10^5$ samples. As described in Section 3.2.3.2, for training the GNN with GMM-based feedback, we construct the set $\mathcal{D}$ with $D = 2{,}400$ scenarios and $J = 16$ MTs. Note that the GNNs are solely trained for this configuration ($J = 16$ MTs) and effectively generalize to configurations with fewer or more served MTs. The GNN consists of five hidden layers with dimension $M_\ell = 128$ for $\ell = 1, \ldots, L-1$, and we set $\alpha = 0.1/N$, $\beta = 0.1$, and use the ReLU activation function, similar to [39]. We train the GNN for 500 epochs, where the batch size is 100, and use the Adam optimizer with 0.001 as the initial learning rate. For model selection, a validation set with $D_{\text{val}} = 300$ scenarios is used to choose the best-performing GNN configuration. To evaluate performance, we use the average sum-rate metric across similarly generated test sets, each with $D_{\text{test}} = 500$ scenarios, for various numbers of MTs J. When using measurement data, see Section 2.2.3, we use the same dataset sizes for training and validating the GNN, for training the GMM, and for testing.

For clarity, we omit the index j in the legend descriptions below and use a slightly different notation than before to facilitate the distinction of different precoder design methods. In the following discussion, "{GNN, SWMMSE}, genie," refers to the cases where perfect statistical knowledge is assumed to design the precoders, see Section 3.2.3.1. The labels "GNN, GMM, $\{\boldsymbol{h}, \boldsymbol{y}\}$," represent the cases where either perfect CSI, see (3.20), is assumed at each MT, or the observation $\boldsymbol{y}_j$ at each MT is used, see (3.19), for determining a feedback index via the GMM-based feedback scheme and subsequently, the precoders are designed using the corresponding GNN. Similarly, "SWMMSE, GMM, $\{\boldsymbol{h}, \boldsymbol{y}\}$," indicates that the SWMMSE is used instead of the GNN to design the precoders based on the GMM feedback, cf. Section 3.2.2.2. The baselines "IWMMSE, DFT, $\{\hat{\boldsymbol{h}}_{\text{GMM}}, \hat{\boldsymbol{h}}_{\text{LS}}\}$," refer to the cases where the channel is estimated at each MT; afterward, the feedback information is determined using the DFT codebook. We either use the LS estimator $\hat{\boldsymbol{h}}_{\text{LS}}$, or the GMM-based estimator $\hat{\boldsymbol{h}}_{\text{GMM}}$, see (3.30). For both the SWMMSE and the iterative WMMSE, the maximum

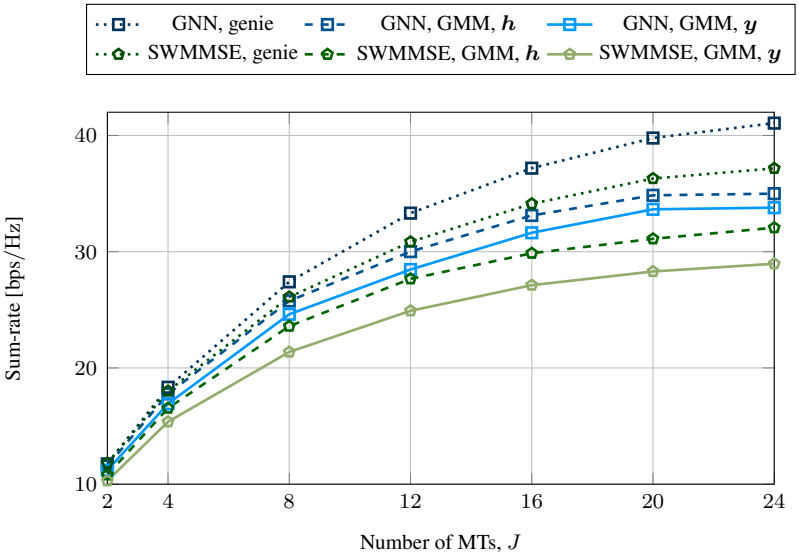


Figure 3.18: The average sum-rate over the number of MTs J for a system with $B = 6$ feedback bits, $n_{\mathrm{p}} = 16$ pilots, and SNR $= 10\,\mathrm{dB}$. Figure adapted from [58].

number of iterations is set to $I_{\max} = 300$, unless otherwise specified.

In Figure 3.18, Figure 3.19(a), and Figure 3.19(b), we use the spatial channel model from Section 2.2.1. In Figure 3.18, we fix $B = 6$ bits, $n_{\mathrm{p}} = 16$ pilots, and SNR $= 10\,\mathrm{dB}$ and vary the number J of served MTs. We can observe that the proposed GNN-based precoder design method using perfect statistical knowledge "GNN, genie" outperforms the baseline "SWMMSE, genie" for all considered numbers J of served MTs. The method "GNN, GMM, $\boldsymbol{h}$" shows performance degradation due to quantized feedback information associated with GMM components, a trend also seen in the "SWMMSE, GMM, $\boldsymbol{h}$" approach. For the respective counterparts solely utilizing the observations for feedback inference, the proposed "GNN, GMM, $\boldsymbol{y}$," significantly outperforms "SWMMSE, GMM, $\boldsymbol{y}$." While the SWMMSE relies on imperfect statistical input from the GMM feedback for optimization, the GNN benefits from access to CSI during the offline training. This allows us to accurately evaluate the loss function of the GNN and perform updates, a capability that is not given by the SWMMSE. Consequently, the GNN holds a significant advantage over the SWMMSE, despite both using the same input during the online phase.

In Figure 3.19(a), we set $B = 6$ bits, $J = 16$ MTs, and $n_{\mathrm{p}} = 8$ pilots and depict the sum-rate over the SNR. While at low SNR values "GNN, genie" and "SWMMSE, genie" perform equally well, with an increasing SNR, the proposed "GNN, genie"

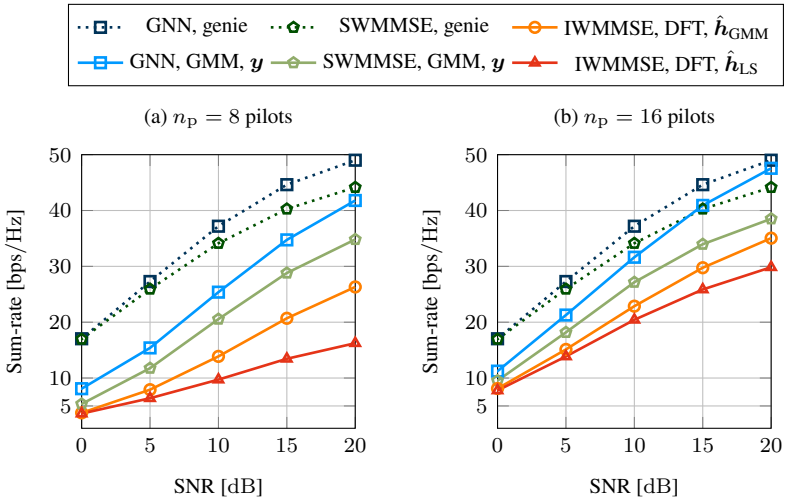


Figure 3.19: The average sum-rate over the SNR for a system with $B = 6$ feedback bits, $J = 16$ MTs, and (a) $n_{\text{p}} = 8$ pilots, or (b) $n_{\text{p}} = 16$ pilots. Figure adapted from [58].

method is superior. Moreover, the proposed "GNN, GMM, $\boldsymbol{y}$" approach outperforms the approach "SWMMSE, GMM, $\boldsymbol{y}$" based on the same GMM feedback and all other baselines "IWMMSE, DFT, $\{\hat{\boldsymbol{h}}_{\text{GMM}}, \hat{\boldsymbol{h}}_{\text{LS}}\}$." While the GMM estimator improves the DFT-based feedback scheme over the LS estimator, the performance is still significantly worse than the "GNN, GMM, $\boldsymbol{y}$" approach. Remarkably, in Figure 3.19(b) where we have the same setup but with $n_{\text{p}} = 16$, the proposed "GNN, GMM, $\boldsymbol{y}$" approach even outperforms the "SWMMSE, genie" in the high SNR regime, due to more accurate GMM feedback with the additional pilots, underlining the effectiveness of the proposed approach in low pilot overhead systems.

In Figure 3.20, we use the measurement data, see Section 2.2.3, to simulate a system with $B = 6$ bits, $J = 8$ MTs, and $n_{\text{p}} = 8$ pilots and depict the sum-rate over the SNR. In this case as well, we observe that the proposed "GNN, GMM, $\boldsymbol{y}$" approach outperforms all other approaches. Since we do not have access to perfect statistics in the case of measured channels, the respective curves are missing. To further emphasize the advantage of the parallelizable GNN, where precoders are determined in a single forward pass, we also evaluate the iterative SWMMSE with reduced maximum iteration counts of 20 and 100 (in addition to $I_{\max} = 300$), denoted as "SWMMSE($\{20, 100, 300\}$), GMM, $\boldsymbol{y}$" for faster precoder calculation. We observe that with fewer iterations, the performance of the SWMMSE baseline degrades.

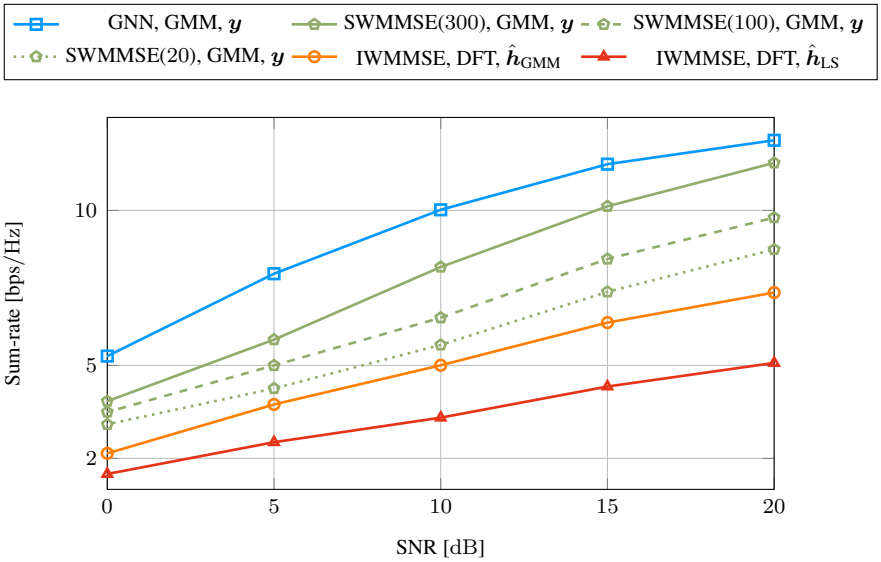


Figure 3.20: The average sum-rate over the SNR for a system with $B = 6$ feedback bits, $J = 8$ MTs, and $n_{\text{p}} = 8$ pilots. Figure adapted from [58].

More simulations involving the GMM-based feedback scheme in combination with the SWMMSE and the subspace counterpart of the method from Section 3.2.2.1 for single-antenna MTs are available in [59].

In the remainder of this section, we use the measurement data, see Section 2.2.3, and evaluate the VQ-VAE-based feedback scheme from Section 3.2.4 in massive MIMO systems. The VQ-VAE allows for higher feedback bit budgets compared to the GMM.

We consider the following baselines using state-of-the-art feedback schemes involving autoencoders (AEs) or VQ-VAEs, where a reconstruction $\bar{\boldsymbol{h}}_j$ of the channel of MT j is obtained at the decoder output, cf. [3,41]. To facilitate a fair comparison, we utilize the same VQ-VAE architecture, encoder input pre-processing, and an embedding $\mathcal{E}$ of the same size. With these baselines, only a mean vector, denoted by $\bar{\boldsymbol{h}}_j$, is inferred. Accordingly, the overall loss function which aims to reconstruct the instantaneous channel is the loss from (2.19) with the MSE as reconstruction loss, see (2.21). In the case of the regular AE, the embedding is omitted, and the overall loss degenerates to the MSE reconstruction loss. Also, in these cases, multi-user operation for any desired number of MTs J is supported, and we train for the same SNR range as in Section 3.2.4 to facilitate operation for different SNR levels. Since the information about each MT j is a reconstruction $\bar{\boldsymbol{h}}_j$, we apply the iterative WMMSE algorithm for

precoder design. Specifically, we use [90, Algorithm 1].

Due to the higher feedback bit budget, the DFT codebook-based baseline additionally contains the channel magnitude information m_j, which is assumed to be the norm of the estimated channel $\hat{\boldsymbol{h}}_j$ and is encoded using B_{Mag} bits [115]. Thus, the overall feedback information is encoded using $B = B_{\text{Dir}} + B_{\text{Mag}}$ bits, where the B_{Dir} bits quantize the directional information using the DFT-based codebook, cf. (3.34). In this case, based on the feedback from each MT, the BS represents the channel of each MT as [115],

$$\hat{\boldsymbol{h}}_j^{(q)} = m_j \boldsymbol{c}_{k_j^\star}, \tag{3.36}$$

and then utilizes the iterative WMMSE [90, Algorithm 1] to jointly design the precoders.

We use $420{,}000$ data samples from the measurement campaign detailed in Section 2.2.3. Specifically, $M = 400{,}000$ samples are used for training the data-based methods, $10{,}000$ for validation, and $10{,}000$ for evaluation. We use a batch size of 256 and the ReLU activation function. The learning rate is randomly drawn in the interval $[3 \cdot 10^{-4}, 9 \cdot 10^{-4}]$ [111]. The models are trained over 500 epochs and the Adam optimizer is used. More details regarding the architecture are available in [69]. As before, the sum-rate averaged over 500 multi-user scenarios is employed as the performance metric. For the SWMMSE as well as the iterative WMMSE, we again set the maximum number of iterations to $I_{\max} = 300$.

In the following discussion, the proposed VQ-VAE-based feedback scheme from Section 3.2.4, which utilizes statistical information for precoder design, is denoted by "VQ-VAE-S." The baseline VQ-VAE feedback scheme, which uses the MSE for instantaneous reconstructions, is denoted by "VQ-VAE-I," and the regular AE with no quantization in the latent space by "AE."

In Figure 3.21, we compare embeddings $\mathcal{E}$ composed of vectors ($N_{\text{E}} = 2$ or $N_{\text{E}} = 4$) with scalar embeddings ($N_{\text{E}} = 1$) for both "VQ-VAE-S" and "VQ-VAE-I," using a fixed feedback bit budget of $B = 40$ bits. The proposed "VQ-VAE-S" consistently outperforms the baseline "VQ-VAE-I" in all cases. For both schemes, using either $N_{\text{E}} = 2$ ($C = 1024$) or $N_{\text{E}} = 1$ ($C = 32$) with a latent dimension of $N_{\text{L}} = 8$ shows minor performance differences. While at high SNRs employing $N_{\text{E}} = 4$ ($C = 1024$) with a larger latent space dimension $N_{\text{L}} = 16$ shows slight gains, this comes at the cost of increased complexity of the forward pass. Thus, scalar embeddings ($N_{\text{E}} = 1$) are adopted in the remainder.

In the subsequent simulations, we fix the feedback budget to $B = 40$ bits, if not mentioned otherwise. This is motivated by a typical compression ratio for the regular AE with $N_{\text{L}} = 8$, cf. [3], where each latent entry is described by 32 bits

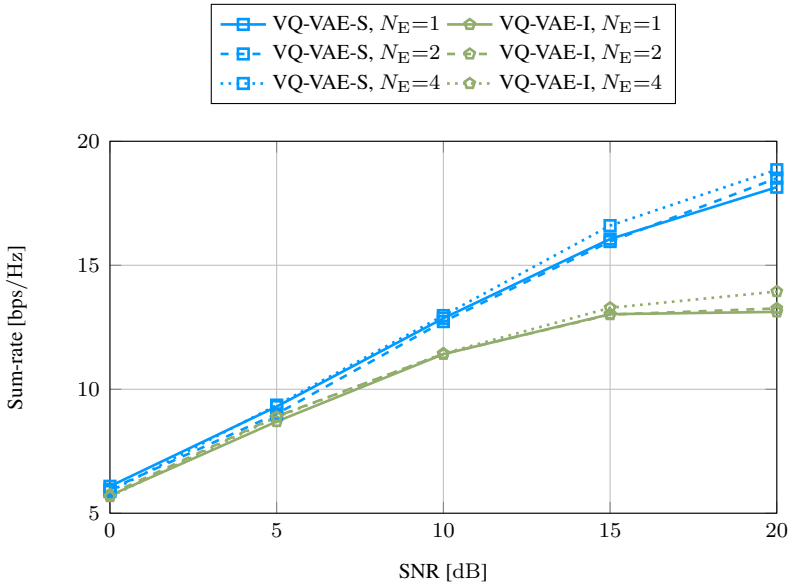


Figure 3.21: The average sum-rate over the SNR for a system with $B = 40$ feedback bits, $J = 8$ MTs, and $n_\mathrm{p} = 8$ pilots for different embedding vector dimensions N_E. Figure adapted from [60].

(single precision), yielding $8 \cdot 32 = 256$ bits in total ("AE, B=256"). We observed, by introducing the embedding with as few as $C = 32$ (i.e., 5 bits per entry), the same performance was achieved with the "VQ-VAE-I," see Figure 3.22. For the VQ-VAE based methods "VQ-VAE-S" and "VQ-VAE-I," we thus fix $N_\mathrm{L} = 8$ and set $C = 32$ ($N_\mathrm{L} \log_2 C = 8 \cdot 5 = 40$ bits). For the DFT codebook-based scheme we allocate $B_\mathrm{Dir} = 8$ bits for the directional information and $B_\mathrm{Mag} = 32$ bits for the magnitude (single precision), ensuring $B_\mathrm{Dir} + B_\mathrm{Mag} = 40$ bits. While different bit allocations for B_Dir and B_Mag are possible to achieve a total of $B = 40$ bits, the chosen allocation maintains reasonable complexity for determining channel directional information. Additionally, the "VQ-VAE-I" method inherently aims for a more efficient bit allocation for improved instantaneous reconstruction compared to the conventional DFT codebook-based scheme.

In Figure 3.22, we have $J = 8$ MTs and $n_\mathrm{p} = 8$ pilots and depict the sum-rate over the SNR. The proposed scheme "VQ-VAE-S" performs best, followed by the "VQ-VAE-I." The conventional approaches "DFT, $\{\hat{\boldsymbol{h}}_\mathrm{GMM}, \hat{\boldsymbol{h}}_\mathrm{LS}\}$" perform poorly, especially at high SNR levels. Despite improvements with the GMM estimator over the LS estimator, the performance remains poor due to significant channel estimation errors. Our proposed scheme with even a reduced feedback bit overhead "VQ-VAE-S,

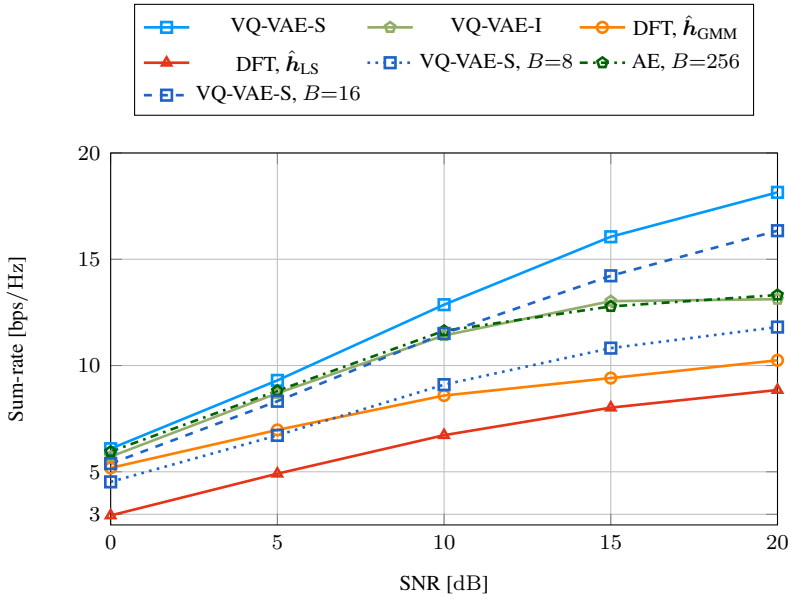


Figure 3.22: The average sum-rate over the SNR for a system with $B = 40$ feedback bits, $J = 8$ MTs, and $n_{\text{p}} = 8$ pilots. Figure adapted from [60].

$B = 16$" ($N_{\text{L}} = 8$, $C = 4$) performs comparable to the "VQ-VAE-I" approach with $B = 40$ bits in the low SNR regime, and is superior for high SNR levels. The proposed scheme with only 8 feedback bits denoted by "VQ-VAE-S, $B = 8$" ($N_{\text{L}} = 8$, $C = 2$) significantly outperforms "DFT, $\hat{\boldsymbol{h}}_{\text{LS}}$" and performs comparable to "DFT, $\hat{\boldsymbol{h}}_{\text{GMM}}$" in the low SNR regime, and is superior for high SNR levels. Accordingly, the performance gains from the proposed VQ-VAE-based precoder design scheme allow for deploying systems with reduced feedback overhead while maintaining performance.

In Figure 3.23, we examine the impact of the number of pilots n_{p} on the performance with $J = 8$ MTs for a fixed SNR of $15\,\text{dB}$. The proposed scheme "VQ-VAE-S" performs best for all considered numbers of pilots n_{p}. With only $n_{\text{p}} = 6$ pilots, the "VQ-VAE-S" approach performs almost as well as the "VQ-VAE-I" baseline with $n_{\text{p}} = 16$ pilots. Additionally, with $n_{\text{p}} = 4$ pilots, the "VQ-VAE-S" approach performs almost as well as the "DFT, $\hat{\boldsymbol{h}}_{\text{GMM}}$" approach and outperforms "DFT, $\hat{\boldsymbol{h}}_{\text{LS}}$" each with $n_{\text{p}} = 16$ pilots. This analysis reveals that systems with lower pilot overhead can be deployed without sacrificing performance compared to the baseline schemes.

In Figure 3.24, we set $n_{\text{p}} = 8$ pilots, and SNR $= 15\,\text{dB}$ and vary the number J of served MTs. It can be seen that the VQ-VAE-based feedback scheme "VQ-VAE-S" is superior as compared to the baselines for all numbers of MTs except for a setup with only

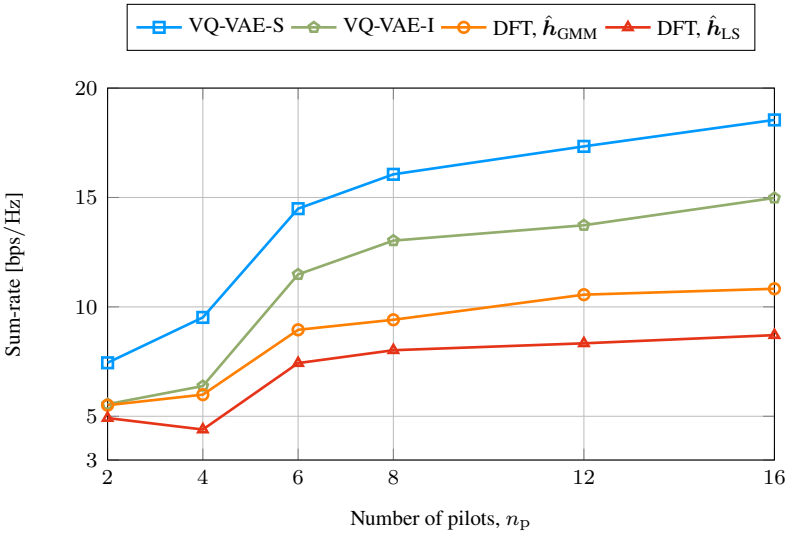


Figure 3.23: The average sum-rate over the number of pilots n_{p} for a system with $J = 8$ MTs, $B = 40$ feedback bits, and SNR $= 15\,\text{dB}$. Figure adapted from [60].

two MTs. Remarkably, the sum-rate of the "VQ-VAE-S" approach steadily increases with an increasing number of MTs. The remaining approaches exhibit saturation or even slightly degrade with an increasing number of MTs. This can be reasoned by the fact that with an increasing number of MTs, resolving the interference present in the scenario becomes more difficult, particularly when restricted to a directional codebook of finite size. In contrast, with the proposed approach "VQ-VAE-S," due to the involved sampling procedure (see Algorithm 5), a different representative interference scenario is provided to and exploited by the SWMMSE to design the precoders. For a system with only two MTs, the interference is less severe, and thus, the "VQ-VAE-I" aiming for an instantaneous reconstruction is beneficial.

Lastly, in Figure 3.25, we set $J = 8$ MTs, $n_{\text{p}} = 8$ pilots, and SNR $= 15\,\text{dB}$ and investigate the impact of the number B of feedback bits on the system performance for the two SNR levels $10\,\text{dB}$ and $20\,\text{dB}$. For $B \leq 40$, we adopt $N_{\text{L}} = 8$ and correspondingly $C \in \{2, 4, 8, 32\}$ (i.e., $\{1, 2, 3, 5\}$ bits per entry). For $B \geq 64$, we adopt $N_{\text{L}} = 32$ and $C \in \{4, 16\}$ (i.e., $\{2, 4\}$ bits per entry). The proposed scheme "VQ-VAE-S" outperforms the baseline "VQ-VAE-I" for all feedback bit configurations by a large margin, especially for SNR $= 20\,\text{dB}$. Once again, the results indicate that the proposed scheme "VQ-VAE-S" can achieve superior performance than the baseline approach "VQ-VAE-I" with fewer bits, reducing both feedback overhead and

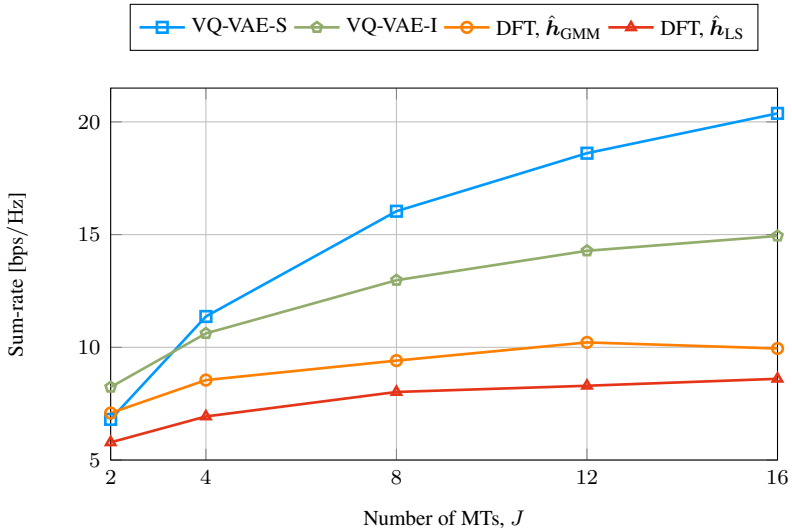


Figure 3.24: The average sum-rate over the number of MTs J for a system with $B = 40$ feedback bits, $n_{\text{p}} = 8$ pilots, and SNR $= 15\,\text{dB}$. Figure adapted from [60].

processing requirements at the MTs. In Figure 3.25, we also depict the performance of the GMM-based feedback scheme, denoted by "GMM samples" with full covariance matrices or "tGMM samples" with block Toeplitz structure. While the GMM-based feedback scheme performs well in systems with a few feedback bits, it suffers from overfitting issues for $B > 9$ bits in the considered setup. Remarkably, both GMM variants with a few bits only, achieves a similar performance or even outperforms the "VQ-VAE-I" baseline with significantly more feedback bits. Although the "VQ-VAE-S" performs slightly worse than the GMM-based approaches at $B = 8$ bits, it demonstrates performance improvements for systems with $B \geq 16$ feedback bits, underlining the effectiveness of the proposed "VQ-VAE-S" scheme.

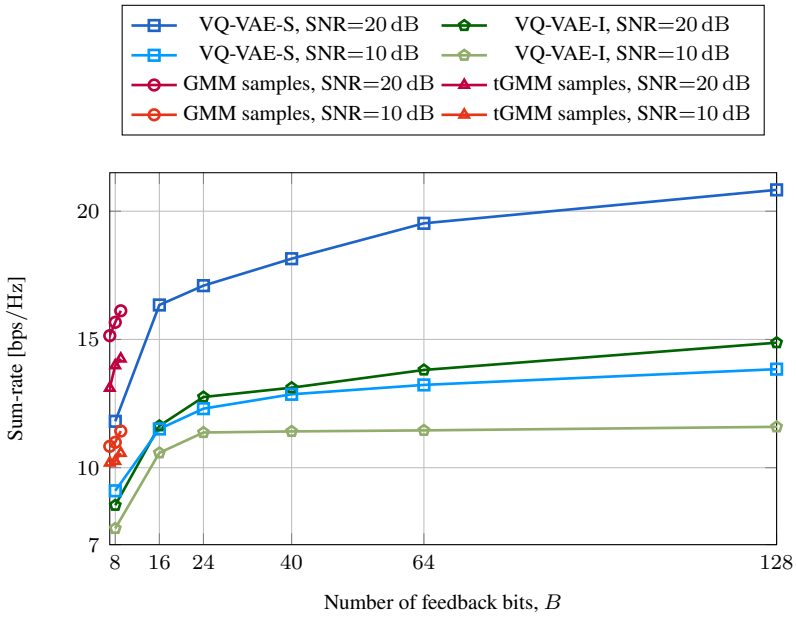


Figure 3.25: The average sum-rate over the number of feedback bits B, for a system with $J = 8$ MTs, $n_{\mathrm{p}} = 8$ pilots, and SNR $\in \{10\,\mathrm{dB}, 20\,\mathrm{dB}\}$. Figure adapted from [60].

Generative Modeling-based Feedback Schemes for Pilot Design

4

This chapter focuses on the design of pilot matrices for single- and multi-user block fading systems to enhance the DL channel estimation performance. Similarly to Chapter 3, the BS relies on feedback information provided by the MTs, subject to a budget of B bits to design the pilot matrices.

The main contributions discussed in this chapter are organized as follows:

- First, we review conventional methods that require perfect statistical knowledge about the channel of the MT for designing pilot matrices in point-to-point MIMO systems. Inspired by the conventional methods, we propose leveraging GMMs for feedback inference and for pilot matrix design in point-to-point MIMO systems. Parts of Sections 4.1.1 and 4.1.2 and the accompanying simulations have been published in [61] (©2024 IEEE) and [62].

- Next, we extend the sum CMI pilot optimization framework that uses perfect statistical knowledge originally designed for multi-user systems with single antenna MTs to MU-MIMO systems. We further propose a low-complexity alternative that optimizes a lower bound on the sum CMI in MU-MIMO systems, serving as a proxy objective. We briefly discuss how a state-of-the-art scheme achieves common pilot knowledge at the BS and across all MTs, using a feedforward link. We propose to leverage quantized information associated with the component selection of the GMM to establish common pilot knowledge. Parts of Sections 4.2.1, 4.2.2 and 4.2.3 and the accompanying simulations have been published in [62].

- The versatility of the proposed GMM-based pilot design scheme is discussed, showcasing its adaptability to single- or multi-user scenarios and its flexibility in adapting to any SNR level and pilot configuration without retraining. Additionally, a detailed complexity analysis is provided. Parts of Section 4.3 have been published in [61] (©2024 IEEE) and [62].

- We conclude this chapter with simulations showcasing the superior performance of the proposed methods for pilot matrix design.

4.1 Single-user Systems

4.1.1 Pilot Optimization with Perfect Statistical Knowledge

In the case of a point-to-point MIMO system, we drop the index j in (2.2) for notational convenience, yielding

$$\boldsymbol{y}_t = (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})\boldsymbol{h}_t + \boldsymbol{n}_t. \tag{4.1}$$

Given genie knowledge of $\boldsymbol{\delta}$, see Section 2.2.1, which fully characterizes the channel statistically, i.e., $\boldsymbol{h}_t|\boldsymbol{\delta} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\delta}})$ for $t = 0, \ldots, T$, since the covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}}$ is assumed to remain constant over $T+1$ blocks, the observation $\boldsymbol{y}_t$ is jointly Gaussian with the channel $\boldsymbol{h}_t$ for all blocks $t = 0, \ldots, T$. Thus, we can compute a genie LMMSE channel estimate via [80]

$$\hat{\boldsymbol{h}}_{t,\mathrm{gLMMSE}} = \mathrm{E}[\boldsymbol{h}_t \mid \boldsymbol{y}_t, \boldsymbol{\delta}] \tag{4.2}$$

$$= \boldsymbol{C}_{\boldsymbol{\delta}}(\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}}((\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})\boldsymbol{C}_{\boldsymbol{\delta}}(\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} + \boldsymbol{\Sigma})^{-1}\boldsymbol{y}_t. \tag{4.3}$$

The goal of pilot optimization is to design the pilot matrix $\boldsymbol{P}_t$ such that the MSE between $\hat{\boldsymbol{h}}_{t,\mathrm{gLMMSE}}$ and the actual channel $\boldsymbol{h}_t$ is minimized [44, 47, 48]:

$$\boldsymbol{P}_t^{\star} = \underset{\boldsymbol{P}_t}{\arg\min}\ \mathrm{E}[\|\hat{\boldsymbol{h}}_{t,\mathrm{gLMMSE}} - \boldsymbol{h}_t\|^2], \tag{4.4}$$

where the pilot matrix $\boldsymbol{P}_t$ typically satisfies either a total power constraint as in [47,48], or an equal power per pilot vector constraint as in [44]. In this work, we consider the latter constraint. Given genie knowledge of $\boldsymbol{\delta}$, i.e., $\boldsymbol{h}_t|\boldsymbol{\delta} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\delta}})$ for $t = 0, \ldots, T$, the optimal pilot matrix $\boldsymbol{P}_t^{\star}$ for every block is the same, i.e., $\boldsymbol{P}_t^{\star} = \boldsymbol{P}_{\text{genie}}$ for $t = 0, \ldots, T$. In particular, $\boldsymbol{P}_{\text{genie}}$ is a sub-unitary matrix [44]

$$\boldsymbol{P}_{\text{genie}} = \sqrt{\rho}\boldsymbol{U}_{\boldsymbol{\delta}}^{\mathrm{H}}[:n_{\mathrm{p}}, :], \tag{4.5}$$

which is composed of the n_{p} dominant eigenvectors of the transmit-side covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}}^{\mathrm{tx}} = \boldsymbol{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\boldsymbol{U}_{\boldsymbol{\delta}}^{\mathrm{H}}$ corresponding to the n_{p} largest eigenvalues, where ρ denotes the transmit power per pilot vector.

Note that power loading across pilot vectors generally performs better but requires additional processing. Moreover, with a sub-unitary pilot design, our proposed scheme discussed in Section 4.1.2 yields a codebook consisting of pilot matrices that do not depend on the SNR, resolving the burden of saving SNR-dependent pilot codebooks.

4.1.2 GMM-based Pilot Design and Downlink Channel Estimation

Throughout this chapter, we enforce the means of the GMM-components to zero, i.e., $\boldsymbol{\mu}_k = \mathbf{0}$ for all $k \in \{1, \ldots, K\}$, with $K = 2^B$, see [76, 78], and impose a Kronecker factorization on the covariance matrices of the GMM, i.e., $\boldsymbol{C}_k = \boldsymbol{C}_k^{\mathrm{tx}} \otimes \boldsymbol{C}_k^{\mathrm{rx}}$. Beyond the regularization benefits and model transfer overhead reduction, see Section 3.1.3, imposing the Kronecker constraint ensures having access to a transmit-side covariance during pilot design in the online phase, see (4.11). Note that the parameters of the GMM, i.e., $\{\pi_k, \boldsymbol{C}_k\}_{k=1}^K$, remain constant across all blocks. Once a GMM is fitted to the training dataset $\mathcal{H}$, see (2.7), the joint Gaussian nature of each GMM component, see (2.10), combined with the AWGN, see (2.2), enables straightforward computation of the GMM of the observations based on the GMM from (2.10) as

$$p_{\mathrm{GMM}}(\boldsymbol{y}_t) = \sum\nolimits_{k=1}^{K} \pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_t; \mathbf{0}, (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}}) \boldsymbol{C}_k (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} + \boldsymbol{\Sigma}). \tag{4.6}$$

Consequently, the MT can compute the responsibilities based on the observations $\boldsymbol{y}_t$ as

$$p(k \mid \boldsymbol{y}_t) = \frac{\pi_k \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_t; \mathbf{0}, (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}}) \boldsymbol{C}_k (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} + \boldsymbol{\Sigma})}{\sum_{i=1}^{K} \pi_i \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_t; \mathbf{0}, (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}}) \boldsymbol{C}_i (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} + \boldsymbol{\Sigma})}, \tag{4.7}$$

and the feedback information $k_t^\star$ is then determined through MAP estimation, cf. (3.19), as

$$k_t^\star = \arg\max_k \; p(k \mid \boldsymbol{y}_t), \tag{4.8}$$

where the index of the component with the highest responsibility for the observation $\boldsymbol{y}_t$ serves as the corresponding feedback information and is thus encoded by B bits. Hence, the feedback information is simply the index of the GMM component that best explains the underlying channel $\boldsymbol{h}_t$ for a given observation $\boldsymbol{y}_t$. Subsequently, the responsibilities are utilized to obtain a channel estimate via the GMM by calculating a convex combination of per-component LMMSE estimates, as discussed in [33, 56]. In particular, the MT estimates the channel by computing, cf. (3.30),

$$\hat{\boldsymbol{h}}_{t,\mathrm{GMM}}(\boldsymbol{y}_t) = \sum\nolimits_{k=1}^{K} p(k \mid \boldsymbol{y}_t) \hat{\boldsymbol{h}}_{t,\mathrm{LMMSE},k}(\boldsymbol{y}_t), \tag{4.9}$$

using the responsibilities $p(k \mid \boldsymbol{y}_t)$ from (4.7) and

$$\begin{aligned} &\hat{\boldsymbol{h}}_{t,\mathrm{LMMSE},k}(\boldsymbol{y}_t) \\ &= \boldsymbol{C}_k (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} ((\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}}) \boldsymbol{C}_k (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{y}_t. \end{aligned} \tag{4.10}$$

Note that parallelization concerning the number of components K is possible during the feedback inference and the application of the LMMSE filters. Although the GMM

estimator is a natural choice due to the usage of the responsibilities, which are anyway computed to determine the feedback information, in principle, any other channel estimator can be used.

Consider the eigenvalue decomposition of each of the GMM's transmit-side covariances, i.e., $\boldsymbol{C}_k^{\mathrm{tx}} = \boldsymbol{U}_k \boldsymbol{\Lambda}_k \boldsymbol{U}_k^{\mathrm{H}}$. For $t > 0$, given the feedback information $k_{t-1}^{\star}$ of the MT from the preceding block $t-1$, i.e., the most recent feedback information, we propose employing the pilot matrix $\boldsymbol{P}_t$ at the BS for the following block t as, cf. (4.5),

$$\boldsymbol{P}_t = \sqrt{\rho} \boldsymbol{U}_{k_{t-1}^{\star}}^{\mathrm{H}}[:n_{\mathrm{p}}, :], \tag{4.11}$$

i.e., the n_{p} dominant eigenvectors of the $k_{t-1}^{\star}$-th transmit-side covariance matrix $\boldsymbol{C}_{k_{t-1}^{\star}}^{\mathrm{tx}}$ are selected as the pilot matrix. Since the GMM-covariances remain fixed, we can store a set of pilot matrices

$$\mathcal{P} = \{\sqrt{\rho} \boldsymbol{U}_k^{\mathrm{H}}[:n_{\mathrm{p}}, :]\}_{k=1}^{K}, \tag{4.12}$$

and the online pilot design utilizing the proposed GMM-based scheme simplifies to a simple selection task based on the feedback information $k_{t-1}^{\star}$ from the previous block. For the initial block $t = 0$, we employ a DFT-based pilot matrix.

4.2 Multi-user Systems

4.2.1 Pilot Optimization with Perfect Statistical Knowledge

Two well-established possibilities for designing pilot sequences for multi-user setups with single-antenna MTs were introduced in [49, 50]. The method proposed in [49] utilizes the close relationship between the mutual information and the MSE [52] to design pilot sequences that maximize the sum CMI expression subject to a transmit power budget. Accordingly, the sum mutual information between each MT's channel $\boldsymbol{h}_{j,t}$ and the observations $\boldsymbol{y}_{j,t}$ conditioned on the transmitted pilot matrix $\boldsymbol{P}_t$ is considered. We refer to [49] for systems where the MTs are equipped with a single antenna. In the following, we extend the framework from [49] to systems where the MTs are equipped with multiple antennas, i.e., to MU-MIMO systems.

The sum CMI maximization problem for MU-MIMO systems for a given number of pilots n_{p} and transmit power budget ρ is expressed as

$$\begin{aligned} \max_{\boldsymbol{P}_t} \; & \sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}}) \boldsymbol{C}_{\boldsymbol{\delta}_j} (\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}}\right) \\ & \text{s.t.} \quad \mathrm{tr}\left(\boldsymbol{P}_t \boldsymbol{P}_t^{\mathrm{H}}\right) \leq \rho n_{\mathrm{p}}. \end{aligned} \tag{4.13}$$

For given $\{\boldsymbol{\delta}_j\}_{j=1}^J$, the optimal pilot matrix $\boldsymbol{P}_t^\star$ for every block is the same, i.e., $\boldsymbol{P}_t^\star = \boldsymbol{P}_{\text{genie}}$ for all $t = 0, \ldots, T$, and we establish the following theorem.

Theorem 1. *The pilot matrix $\boldsymbol{P}_t^\star$ that maximizes the sum CMI in a MU-MIMO system provided in* (4.13) *satisfies the condition*

$$\sum_{j=1}^{J}\sum_{i=1}^{I} \tfrac{1}{\sigma_n^2}\boldsymbol{V}_j\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}}\boldsymbol{V}_j^{\mathrm{H}}\boldsymbol{P}_t^\star\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \operatorname{tr}\left(\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}\boldsymbol{T}_j\right) = \lambda\boldsymbol{P}_t^\star, \tag{4.14}$$

with the SVDs for all $j \in \mathcal{J}$,

$$\boldsymbol{P}_t^\star\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}\boldsymbol{P}_t^{\star,\mathrm{H}} = \boldsymbol{V}_j\boldsymbol{S}_j\boldsymbol{V}_j^{\mathrm{H}}, \tag{4.15}$$

$$\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}} = \boldsymbol{W}_j\boldsymbol{T}_j\boldsymbol{W}_j^{\mathrm{H}}. \tag{4.16}$$

Further, let $\sum_{i=1}^{I}\alpha_{j,i}\boldsymbol{\gamma}_{j,i}\boldsymbol{\beta}_{j,i}^{\mathrm{T}}$ with $I = \min(n_{\mathrm{p}}, N_{\mathrm{rx}})$, be the SVD of the matrix $\operatorname{unvec}_{N_{\mathrm{rx}},n_{\mathrm{p}}}\left(\operatorname{diag}\left((\mathbf{I} + \frac{1}{\sigma_n^2}\boldsymbol{S}_j \otimes \boldsymbol{T}_j)^{-1}\right)\right)$. Then, the diagonal matrices $\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \in \mathbb{R}^{n_{\mathrm{p}}\times n_{\mathrm{p}}}$ and $\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \in \mathbb{R}^{N_{\mathrm{rx}}\times N_{\mathrm{rx}}}$ for all $j \in \mathcal{J}$, in (4.14)*, are defined as*

$$\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} = \alpha_{j,i}\operatorname{diag}(\boldsymbol{\beta}_{j,i}), \tag{4.17}$$

$$\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} = \operatorname{diag}(\boldsymbol{\gamma}_{j,i}), \tag{4.18}$$

and $\lambda \geq 0$ is a constant chosen to satisfy the power constraint.

Proof. See Appendix A.1. □

In the following corollary, we consider the special case with $n_p \leq N_{\mathrm{rx}}$, which allows for a simplification.

Corollary 1.1. *The pilot matrix $\boldsymbol{P}_t^\star$ that maximizes the sum CMI in a MU-MIMO system provided in* (4.13) *with $n_{\mathrm{p}} \leq N_{\mathrm{rx}}$ satisfies the condition*

$$\sum_{j=1}^{J}\sum_{i=1}^{n_{\mathrm{p}}} \tfrac{1}{\sigma_n^2}\boldsymbol{v}_{j,i}\boldsymbol{v}_{j,i}^{\mathrm{H}}\boldsymbol{P}_t^\star\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \operatorname{tr}\left(\boldsymbol{D}_{j,\bar{i}}\boldsymbol{T}_j\right) = \lambda\boldsymbol{P}_t^\star, \tag{4.19}$$

utilizing the SVDs for all $j \in \mathcal{J}$ from (4.15) *and* (4.16)*, where $\boldsymbol{v}_{j,i}$ denotes the i-th column of $\boldsymbol{V}_j$, and*

$$\begin{aligned}&(\mathbf{I} + \tfrac{1}{\sigma_n^2}\boldsymbol{S}_j \otimes \boldsymbol{T}_j)^{-1}\\ &= \operatorname{diag}\left([d_{j,1}, \ldots, d_{j,n_{\mathrm{p}}N_{\mathrm{rx}}}]^{\mathrm{T}}\right) = \sum_{i=1}^{n_{\mathrm{p}}}\operatorname{diag}(\mathbf{e}_i)\otimes\boldsymbol{D}_{j,\bar{i}},\end{aligned} \tag{4.20}$$

with the n_{p}-dimensional vectors $\mathbf{e}_i$, and $\boldsymbol{D}_{j,\bar{i}} = \operatorname{diag}([d_{j,(i-1)N_{\mathrm{rx}}+1}, \ldots, d_{j,iN_{\mathrm{rx}}}]^{\mathrm{T}})$, cf. (A.7)*, where $\lambda \geq 0$ is a constant chosen to satisfy the power constraint.*

Algorithm 6 Iterative Pilot Matrix Design for MU-MIMO Systems. Adapted from [62, Algortihm 1].

1: For $\ell = 0$, initialize pilot matrix $\boldsymbol{P}_t^{(0)}$, and set maximum number of iterations $L_{\max}$. Compute SVDs for all $j \in \mathcal{J}$, $\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}} = \boldsymbol{W}_j \boldsymbol{T}_j \boldsymbol{W}_j^{\mathrm{H}}$. Set ε and set $\ell = 1$.
2: **repeat**
3: Compute SVDs for all $j \in \mathcal{J}$,
$\boldsymbol{P}_t^{(\ell-1)} \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \boldsymbol{P}_t^{(\ell-1),\mathrm{H}} = \boldsymbol{V}_j^{(\ell-1)} \boldsymbol{S}_j^{(\ell-1)} \boldsymbol{V}_j^{(\ell-1),\mathrm{H}}$
4: Compute real-valued SVDs for all $j \in \mathcal{J}$, $I = \min(n_{\mathrm{p}}, N_{\mathrm{rx}})$,
$\sum_{i=1}^{I} \alpha_{j,i}^{(\ell-1)} \boldsymbol{\gamma}_{j,i}^{(\ell-1)} \boldsymbol{\beta}_{j,i}^{(\ell-1),\mathrm{T}}$
$= \mathrm{unvec}_{N_{\mathrm{rx}},n_{\mathrm{p}}} \left(\mathrm{diag} \left((\mathbf{I} + \frac{1}{\sigma_n^2} \boldsymbol{S}_j^{(\ell-1)} \otimes \boldsymbol{T}_j)^{-1} \right) \right)$
5: Set for all j and i,
$\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}}^{(\ell-1)} = \alpha_{j,i}^{(\ell-1)} \mathrm{diag} \left(\boldsymbol{\beta}_{j,i}^{(\ell-1)} \right)$
$\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}^{(\ell-1)} = \mathrm{diag} \left(\boldsymbol{\gamma}_{j,i}^{(\ell-1)} \right)$
6: Compute intermediate pilot matrix $\tilde{\boldsymbol{P}}_t^{(\ell)}$
$= \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\mathrm{tr}(\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}^{(\ell-1)} \boldsymbol{T}_j)}{\sigma_n^2} \boldsymbol{V}_j^{(\ell-1)} \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}}^{(\ell-1)} \boldsymbol{V}_j^{(\ell-1),\mathrm{H}} \boldsymbol{P}_t^{(\ell-1)} \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}$
7: Apply normalization to fulfill power constraint,
$\boldsymbol{P}_t^{(\ell)} = \sqrt{\frac{\rho n_{\mathrm{p}}}{\mathrm{tr}\left(\tilde{\boldsymbol{P}}_t^{(\ell)} \tilde{\boldsymbol{P}}_t^{(\ell),\mathrm{H}} \right)}} \tilde{\boldsymbol{P}}_t^{(\ell)}$
8: $\ell \leftarrow \ell + 1$
9: **until** Spectral norm $\| \boldsymbol{P}_t^{(\ell)} - \boldsymbol{P}_t^{(\ell-1)} \|_2 < \varepsilon$ or $\ell \geq L_{\max}$.

Note that, in (4.20), we utilized that any diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n_{\mathrm{p}} N_{\mathrm{rx}} \times n_{\mathrm{p}} N_{\mathrm{rx}}}$ with $n_p \leq N_{\mathrm{rx}}$ can be decomposed as $\boldsymbol{D} = \mathrm{diag}\left([d_1, \ldots, d_{n_{\mathrm{p}} N_{\mathrm{rx}}}]^{\mathrm{T}} \right) = \sum_{i=1}^{n_{\mathrm{p}}} \mathrm{diag}(\mathbf{e}_i) \otimes \boldsymbol{D}_{\bar{i}}$, with the n_{p}-dimensional vectors $\mathbf{e}_i$, and the $(N_{\mathrm{rx}} \times N_{\mathrm{rx}})$-dimensional diagonal matrices $\boldsymbol{D}_{\bar{i}} = \mathrm{diag}([d_{(i-1)N_{\mathrm{rx}}+1}, \ldots, d_{iN_{\mathrm{rx}}}]^{\mathrm{T}})$.

In general, the optimization problem in (4.13) is non-convex (see the discussion at the end of this section). Thus, similarly to the algorithm for MTs with single antennas [49], we outline an iterative algorithm that designs the pilot matrix based on the first-order Karush-Kuhn-Tucker (KKT) condition for MU-MIMO systems, cf. (4.14), in Algorithm 6. Throughout this work, we set $\varepsilon = 10^{-3}$. As initialization, we apply either a random pilot matrix or a random selection of columns of an (oversampled) DFT matrix for $\boldsymbol{P}_t^{(0)}$. We discuss the effect of the initialization as well as the specific selection of the maximum number of iterations $L_{\max}$ on the algorithm's performance in Section 4.4.3.

To reduce the computational overhead of optimizing (4.13), we establish a lower

bound on the sum CMI expression serving as a proxy objective. We also present conditions under which this lower bound exhibits the largest gap to the sum CMI and when it is tight. Accordingly, one can maximize this lower bound instead of the sum CMI.

Theorem 2. *The sum CMI expression of a MU-MIMO system in* (4.13) *can be lower bounded by*

$$\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}\boldsymbol{P}_t^{\mathrm{H}})\right). \tag{4.21}$$

Proof. See Appendix A.2. □

Proposition 2.1. *The lower bound in* (4.21) *is tight in a MU-MIMO system as formulated in* (4.13) *for the same but arbitrary* $\boldsymbol{P}_t$ *if and only if for all* $j \in \mathcal{J}$

$$\operatorname{rk}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}) = 1. \tag{4.22}$$

Proof. See Appendix A.3. □

Proposition 2.2. *The lower bound expression from* (4.21) *attains the largest gap to the sum CMI in a MU-MIMO system provided in* (4.13) *for the same but arbitrary* $\boldsymbol{P}_t$*, with fixed traces* $\tau_j = \operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})$*, if and only if for all* $j \in \mathcal{J}$

$$\tfrac{1}{\tau_j}\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}} = \tfrac{1}{N_{\mathrm{rx}}}\mathbf{I}. \tag{4.23}$$

Proof. See Appendix A.4. □

Corollary 2.1. *In conjunction with [49, Th. 1], the optimal pilot matrix* $\boldsymbol{P}_t^{\star}$*, which maximizes the lower bound expression in* (4.21)*, satisfies the condition*

$$\sum_{j=1}^{J} \tfrac{1}{\sigma_n^2}\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\boldsymbol{P}_t^{\star}\operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}\boldsymbol{P}_t^{\star,\mathrm{H}}\right)^{-1}\boldsymbol{P}_t^{\star}\operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} = \lambda\boldsymbol{P}_t^{\star}. \tag{4.24}$$

Due to Proposition 2.1, there are cases where the lower bound provided in Theorem 2 matches the MU-MISO sum CMI expression (with additional scaling by $\operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})$ for all $j \in \mathcal{J}$) in [49], and consequently, [49, Remark 1] can be utilized to show that the MU-MIMO sum CMI maximization problem in (4.13) is not a convex problem.

The main advantage of maximizing the lower bound is the reduced computational complexity, see Section 4.3.2, compared to solving the original problem in (4.13), since as a consequence of Theorem 2 and Corollary 2.1, the MU-MISO optimizer

from [49] can be used for pilot design. Specifically, the steps outlined in [49, Eq. (18)] and [49, Eq. (19)] are similarly applied successively and repeatedly, with the additional scaling of the transmit-side covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}}$ by $\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}})$ for all $j \in \mathcal{J}$. The same stopping criterion as described in step nine of Algorithm 6 is used.

4.2.2 Conventional Multi-user Pilot Matrix Signaling

In general, each of the MTs is unaware of the statistics of the other MTs. Thus, after designing the pilot matrix at the BS following either the iterative Algorithm 6 outlined in Section 4.2.1 utilizing Theorem 1, or using the optimizer from [49] due to Theorem 2, the pilot matrix needs to be transferred from the BS to the MTs. Transferring the complete pilot matrix to the MTs might result in a significant signaling overhead that contradicts the initial goal of training overhead reduction [50, 116]. The state-of-the-art solution, introduced in [50], employs a quantization codebook to avoid transferring the complete pilot matrix. A codebook $\mathcal{C}_{\text{RMP}}$, with $|\mathcal{C}_{\text{RMP}}| = 2^{B_{\text{RMP}}}$ codewords, shared between the BS and all MTs is used to quantize the n_{p} pilot vectors of the pilot matrix. Subsequently, the selected codeword indices are broadcasted to all MTs using a limited feedforward link. The method proposed in [50] avoids brute force search of the codewords and instead relies on the so-called range matching pursuit (RMP) algorithm [50, Algorithm 2]. Thereby, the span of the resulting pilot matrix, which can be encoded by a total of $n_{\mathrm{p}} B_{\text{RMP}}$ bits, provides the best approximation of the span of the originally designed pilot matrix. As suggested in [50, 116], the N_{tx}-dimensional columns of an oversampled DFT matrix can be selected as the elements of the codebook $\mathcal{C}_{\text{RMP}}$.

4.2.3 GMM-based Multi-user Pilot Matrix Design and Signaling

In the multi-user case, we can exploit the same GMM as in the single-user case and, thus, require no additional training or adjustment. We propose that MT j determines its feedback information via a MAP estimation, similarly to the single-user case, as

$$k_{j,t}^{\star} = \arg\max_{k} \; p(k \mid \boldsymbol{y}_{j,t}). \tag{4.25}$$

For $t > 0$, given the feedback information $\{k_{j,t-1}^{\star}\}_{j=1}^{J}$ of all MTs from the preceding block $t-1$, we propose designing the pilot matrix $\boldsymbol{P}_t$ at the BS for the following block t by utilizing the respective GMM-component's covariance information, cf. Section 4.1.2,

$$\{\boldsymbol{C}_{k_{j,t-1}^{\star}} = \boldsymbol{C}_{k_{j,t-1}^{\star}}^{\mathrm{tx}} \otimes \boldsymbol{C}_{k_{j,t-1}^{\star}}^{\mathrm{rx}}\}_{j=1}^{J} \tag{4.26}$$

associated with the most recent feedback indices $\{k^{\star}_{j,t-1}\}_{j=1}^{J}$. At the BS, the covariance information obtained with the help of the GMM can then be leveraged to design the pilot matrix $\boldsymbol{P}_t$ by either applying the iterative algorithm (Algorithm 6) outlined in Section 4.2.1 following Theorem 1, or using the iterative algorithm from [49] following Corollary 2.1. For the initial block $t = 0$, we employ a random selection of columns of an (oversampled) DFT matrix as the pilot matrix. The MTs are, in general, unaware of the channel statistics of the other MTs, or in this particular case, the covariance information associated with the GMM components. In contrast to the conventional feedforward signaling, see Section 4.2.2, we propose to broadcast the feedback information $\{k^{\star}_{j,t-1}\}_{j=1}^{J}$ of all MTs using a feedforward link to all MTs, i.e., the BS first collects the feedback indices of all MTs and subsequently feedforwards them to all MTs. Accordingly, the feedforward information is encoded by JB bits. This approach requires each MT to design the pilot matrix locally to obtain the common pilot matrix knowledge required for the next fading block. Thereby, the computational effort at the MTs can be kept low by either utilizing the lower bound optimization approach or restricting the maximum number of iterations $L_{\max}$ of the iterative algorithm outlined in Section 4.2.1. Furthermore, the initialization of the pilot matrix of the iterative algorithm needs to be based on the same initial pilot matrix. We use the oversampled DFT matrix, which is used in the zeroth block, or a random pilot matrix as the initial pilot matrix. The initialization, the selection of $L_{\max}$, and the choice of the optimization approach allow for a complexity-to-performance trade-off, and are discussed in Sections 4.3.2 and 4.4.3.

4.3 A Versatile Limited Feedback Pilot Design Scheme

4.3.1 Discussion on the Versatility

After fitting the GMM at the BS offline, the model from (2.10) is offloaded, i.e., the model parameters are transferred to every MT within the coverage area of the BS. In the online phase, at the MTs, the GMM of the observations, see (4.6), can be constructed depending on the pilots and SNR level and does not require retraining. By utilizing the GMM of the observations, each MT can infer its feedback information $k^{\star}_{j,t}$ by evaluating (4.25), which is subsequently fed back to the BS. In particular, the BS can then decide to either serve one MT in the point-to-point mode or all MTs simultaneously in the multi-user mode. Thus, the GMM-based pilot design scheme facilitates a handover- and scheduling-friendly scheme for pilot design in FDD systems. In the single-user mode (denoted as P2P in Figure 4.1), the BS simply selects the pilot matrix associated with the MT's feedback information of the preceding block $k^{\star}_{j,t-1}$

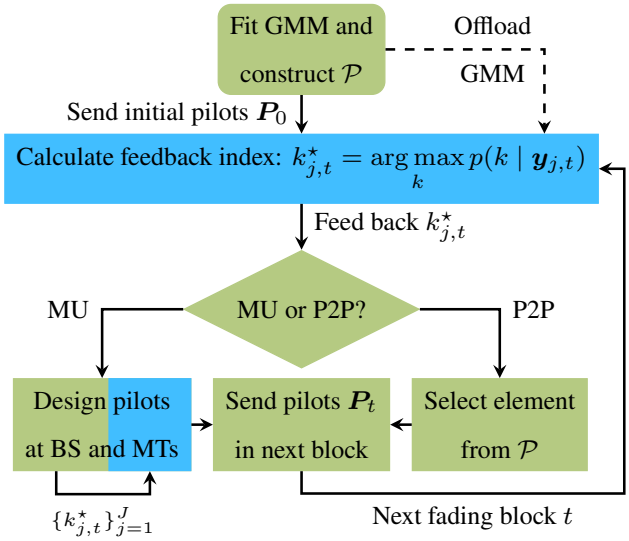


Figure 4.1: Flowchart of the versatile pilot matrix design scheme. Green (blue) colored nodes are processed at the BS (MTs), and solid (dashed) arrows indicate online (offline) processing. Figure adapted from [62].

from the pilot codebook $\mathcal{P}$, see Section 4.1.2 for details, and no further processing is required. In the multi-user mode (denoted as MU in Figure 4.1), a feedforward link is required to broadcast the feedback information $\{k^\star_{j,t-1}\}_{j=1}^J$ from the BS to all MTs. Then, the pilot matrix $\boldsymbol{P}_t$ for the subsequent fading block can be obtained at the BS and MTs utilizing either the iterative algorithm (Algorithm 6) outlined in Section 4.2.1 following Theorem 1, or using the iterative algorithm from [49] following Corollary 2.1. The flexibility regarding the transmission mode, together with the possibility to serve any desired number of MTs J, in combination with the straightforward adaptation to any number of pilots n_p and SNR level, highlights the versatility of the proposed GMM-based pilot design scheme. The proposed versatile pilot design scheme is summarized as a flowchart in Figure 4.1, where green (blue) colored nodes represent processing steps that are performed at the BS (MTs).

4.3.2 Complexity Analysis

The online computational complexity of the proposed GMM-based scheme can be divided into three parts:

Inference of the Feedback Information: The inference of the feedback information using (4.25) in the online phase at each MT has a complexity of $\mathcal{O}(KN_\mathrm{rx}^2 n_\mathrm{p}^2)$, where parallelization concerning the number of components K is possible, cf. Section 3.3.2

for details.

Pilot Matrix Design: The computational complexity of the online pilot matrix design depends on the transmission mode, i.e., single- or multi-user mode. In the single-user mode, the computational complexity is $\mathcal{O}(K)$, as it only involves traversing the pre-computed set of pilot matrices $\mathcal{P}$. Thus, in the online phase, our scheme avoids computing an eigenvalue decomposition, which is required for solving the optimization problem from (4.4). The computational complexity in the multi-user mode depends on the applied optimization algorithm. In the case of the iterative algorithm outlined in Section 4.2.1 following Theorem 1, the computational complexity per iteration is $\mathcal{O}(J(n_{\mathrm{p}}N_{\mathrm{tx}}^2 + n_{\mathrm{p}}^2 N_{\mathrm{tx}} \min(n_{\mathrm{p}}, N_{\mathrm{rx}})))$. Alternatively, using the iterative algorithm from [49] following Corollary 2.1 exhibits a complexity of $\mathcal{O}(Jn_{\mathrm{p}}N_{\mathrm{tx}}^2)$ per iteration.

Channel Estimation: Processing the responsibilities to an estimated channel with the GMM via (4.9) exhibits a computational complexity of $\mathcal{O}(KN_{\mathrm{rx}}^2 N_{\mathrm{tx}} n_{\mathrm{p}})$, since also the LMMSE filters for a given SNR level can be pre-computed. Similar to the feedback inference, parallelization concerning the number of components K is possible when applying the LMMSE filters. As discussed before, in general, any other channel estimator can be used.

4.4 Simulation Results

4.4.1 Channel Estimation and Pilot Design Baselines

Since channel estimation is performed independently at each MT, we discuss the estimators from a single MT perspective and omit the index j for simplicity. In addition to the genie-aided LMMSE approach in (4.2), where we assume perfect knowledge of $\boldsymbol{\delta}$ at the BS to design the optimal pilots and at the MT to apply the genie-aided LMMSE estimator, we consider the following channel estimators and pilot matrices.

Firstly, we consider the LMMSE estimator $\hat{\boldsymbol{h}}_{\mathrm{LMMSE}}$, see Section 3.4.1. The main purpose of this baseline is to demonstrate that a Gaussian approximation cannot effectively characterize the overall PDF of the communications environment.

Secondly, we consider a compressive sensing estimation method $\hat{\boldsymbol{h}}_{\mathrm{OMP}}$ employing the OMP algorithm, see Section 3.4.1.

Additionally, we compare to an end-to-end DNN approach for DL channel estimation with a jointly learned pilot matrix $\boldsymbol{P}_{\mathrm{DNN}}$, similarly to [53, 54]. To determine the hyperparameters of the DNN, we utilize random search [111], with the MSE serving as the loss function. The DNN architecture comprises D_{CM} convolutional modules, each consisting of a convolutional layer, batch normalization, and an activation function, where D_{CM} is randomly selected within $[3, 9]$. Each convolutional layer

contains D_{K} kernels, where D_{K} is randomly selected within $[32, 64]$. The activation function in each convolutional module is the same and is randomly selected from {ReLu, sigmoid, PReLU, Leaky ReLU, tanh, swish}. Following a subsequent two-dimensional max-pooling, the features are flattened, and a fully connected layer is employed with an output dimension of $2N_{\mathrm{tx}}N_{\mathrm{rx}}$ (concatenated real and imaginary parts of the estimated channel). We train a distinct DNN for each pilot configuration and SNR level, running 50 random searches per pilot configuration and SNR level and selecting the best-performing DNN for each setup.

Lastly, as further baseline pilot matrices, we utilize a DFT sub-matrix $\boldsymbol{P}_{\mathrm{DFT}}$ as the pilot matrix, see, e.g., [81], and alternatively, we consider random pilot matrices denoted by $\boldsymbol{P}_{\mathrm{RND}}$, see, e.g., [48].

4.4.2 Point-to-Point MIMO Systems

Throughout the simulations, we consider the spatial channel model assuming a single propagation cluster composed of multiple sub-paths per MT, see Section 2.2.1, and use a set $\mathcal{H}$, see (2.7), with $M = 10^5$ samples for fitting the GMM and all other data-aided baselines. We use a different data set of $M_{\mathrm{eval}} = 10^4$ channel samples per block t for evaluation purposes, where we set $T = 10$. The data samples are normalized to satisfy $\mathrm{E}[\|\boldsymbol{h}\|^2] = N = N_{\mathrm{tx}}N_{\mathrm{rx}}$. Additionally, we fix $\rho = 1$, enabling the definition of the SNR as $\frac{1}{\sigma_n^2}$. We employ the normalized MSE (NMSE) as the performance measure. Specifically, for every block t, we compute a corresponding channel estimate $\hat{\boldsymbol{h}}^{(m_{\mathrm{eval}})}$ for each test channel in the set $\{\boldsymbol{h}^{(m_{\mathrm{eval}})}\}_{m_{\mathrm{eval}}=1}^{M_{\mathrm{eval}}}$, and calculate $\mathrm{NMSE} = \frac{1}{NM_{\mathrm{eval}}} \sum_{m_{\mathrm{eval}}=1}^{M_{\mathrm{eval}}} \|\boldsymbol{h}^{(m_{\mathrm{eval}})} - \hat{\boldsymbol{h}}^{(m_{\mathrm{eval}})}\|^2$. If not mentioned otherwise, we consider the block with index $t = 5$ in the subsequent simulations.

In Figure 4.2, we simulate a system with $N_{\mathrm{tx}} = 16$ BS antennas, $N_{\mathrm{rx}} = 4$ MT antennas, and $n_{\mathrm{p}} = 4$ pilots. Since we have a MIMO setup, we fit a Kronecker-structured GMM with a total of $K = 2^7 = 128$ components ($B = 7$ feedback bits), where $K_{\mathrm{tx}} = 32$ and $K_{\mathrm{rx}} = 4$. The proposed GMM-based pilot design scheme, see Section 4.1.2, denoted by "GMM, $\boldsymbol{P}_{\mathrm{GMM}}$" outperforms all baselines "{GMM, OMP, LMMSE}, {$\boldsymbol{P}_{\mathrm{DFT}}$, $\boldsymbol{P}_{\mathrm{RND}}$}" by a large margin, where the GMM estimator, the OMP-based estimator, or the LMMSE estimator, are used in combination with either DFT-based pilot matrices or random pilot matrices. The proposed scheme also outperforms the DNN based approach denoted by "DNN, $\boldsymbol{P}_{\mathrm{DNN}}$," which jointly learns the estimator and a global pilot matrix for the whole scenario, i.e., it cannot provide an MT-adaptive pilot matrix. This highlights the advantage of the proposed model-based technique over the end-to-end learning approach, which is even trained

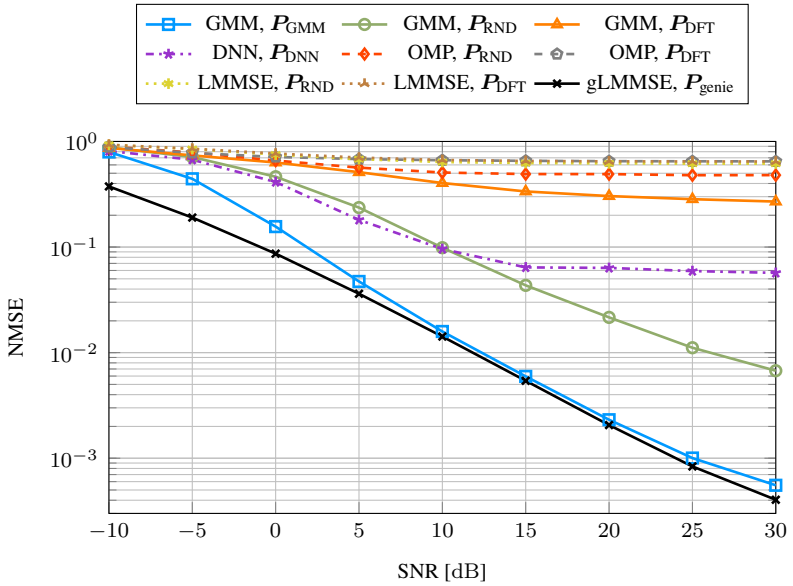


Figure 4.2: The NMSE over the SNR for a MIMO system ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$) with $B = 7$ feedback bits and $n_{\text{p}} = 4$ pilots. Figure adapted from [61] (©2024 IEEE) and [62].

separately for each SNR level and pilot configuration. Furthermore, we can observe that the GMM-based pilot design scheme performs only slightly worse than the genie-aided baseline with perfect statistical information at the BS and MT, see (4.2) and (4.5), denoted by "gLMMSE, $\boldsymbol{P}_{\text{genie}}$." We can observe a larger gap in the low SNR regime, where the feedback information obtained through the responsibilities of a given observation, see (4.8), is less accurate due to high noise.

In Figure 4.3, we analyze the performance of the proposed GMM-based pilot design scheme over the block index t for the same setup as before at three different SNR levels, i.e., SNR $\in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. As discussed in Section 4.1.2, at $t = 0$, we utilize DFT-based pilots. At $t = 1$, we can already see a significant gain in performance of the proposed GMM-based pilot design scheme due to the feedback of the index. The results further reveal that with an increasing SNR, fewer blocks are required to achieve a performance close to the genie-aided baseline "gLMMSE, $\boldsymbol{P}_{\text{genie}}$" which requires perfect statistical knowledge at the BS and the MT.

In the remainder of this section, we consider a MISO system with $N_{\text{tx}} = 64$ BS antennas and $N_{\text{rx}} = 1$ antenna at the MT. In Figure 4.4, we utilize a GMM with $K = 2^6 = 64$ components ($B = 6$ feedback bits) and consider setups with

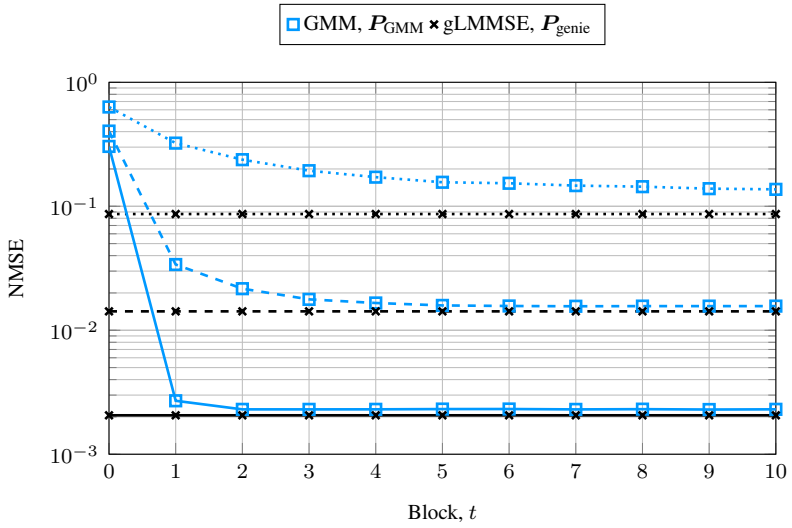


Figure 4.3: The NMSE over the block index t for a MIMO system ($N_{\text{tx}} = 16$, $N_{\text{rx}} = 4$) with $B = 7$ feedback bits, $n_{\text{p}} = 4$ pilots, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [61] (©2024 IEEE) and [62].

$n_{\text{p}} \in \{16, 32, 48\}$ pilots. In this case, the proposed scheme "GMM, $\boldsymbol{P}_{\text{GMM}}$, $n_{\text{p}} = 16$," performs only slightly worse than the genie-aided approach "gLMMSE, $\boldsymbol{P}_{\text{genie}}$." Moreover, the GMM-based pilot design scheme outperforms the baselines "GMM, $\{\boldsymbol{P}_{\text{DFT}}, \boldsymbol{P}_{\text{RND}}\}$, $n_{\text{p}} \in \{16, 32, 48\}$." In particular, the proposed scheme with only $n_{\text{p}} = 16$ pilots outperforms random pilot matrices with twice as many pilots ($n_{\text{p}} = 32$), or in case of DFT-based pilots even three times as many pilots ($n_{\text{p}} = 48$).

In Figure 4.5, we examine the same GMM with $K = 2^6 = 64$ components ($B = 6$ feedback bits) and evaluate setups with varying numbers of pilots n_{p} and consider three different SNR levels, i.e., SNR $\in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. The proposed scheme "GMM, $\boldsymbol{P}_{\text{GMM}}$" and the genie-aided approach "gLMMSE, $\boldsymbol{P}_{\text{genie}}$," similarly benefit from more pilots while a saturation can be observed for both schemes due to finite SNR levels. As the SNR increases, the gap between the proposed scheme and the genie-aided approach shrinks because feedback information derived from processing the responsibilities of a given observation via (4.8) becomes more accurate with low noise.

Lastly, in Figure 4.6, we analyze the effect of varying the number of GMM-components K and accordingly feedback bits B, on the scheme's performance, where we set $n_{\text{p}} = 16$, and consider three different SNR levels, i.e., SNR $\in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. We can observe that the estimation error decreases with

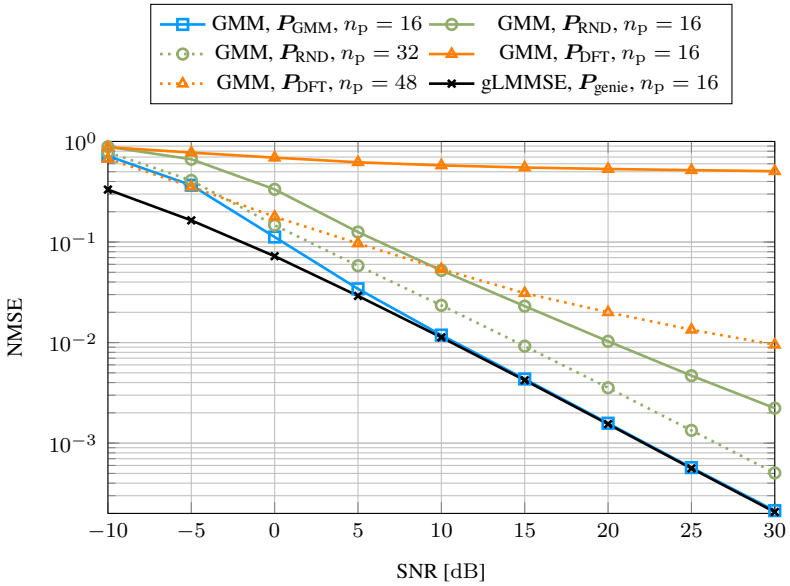


Figure 4.4: The NMSE over the SNR for a MISO system ($N_{\text{tx}} = 64$, $N_{\text{rx}} = 1$) with $B = 6$ feedback bits and $n_{\text{p}} \in \{16, 32, 48\}$ pilots. Figure adapted from [61] (©2024 IEEE) and [62].

an increasing number of components K. Moreover, as the SNR increases, the gap to the genie-aided approach "gLMMSE, $\boldsymbol{P}_{\text{genie}}$" shrinks. Above $K = 32$ components, a saturation can be observed. These results suggest that varying the number of GMM components K, allows for a performance-to-complexity trade-off without sacrificing too much in performance for $K \geq 16$ components.

4.4.3 Multi-user MIMO Systems

In the following multi-user simulations, the performance is measured by the NMSE averaged over the number of MTs J and 500 scenarios, with J MTs randomly drawn from our evaluation set for each scenario, cf. [50]. If not mentioned otherwise, we consider the block with index $t = 5$ in the simulations, do not specify a maximum number of iterations $L_{\max}$ for the pilot optimization algorithm (Algorithm 6) from Section 4.2.1, and initialize with a random pilot matrix.

In Figure 4.7, we consider a MU-MIMO system with $J = 4$ MTs, $N_{\text{tx}} = 32$ BS antennas, $N_{\text{rx}} = 4$ MT antennas, and $n_{\text{p}} = 8$ pilots. Due to the MIMO setup, we fit a Kronecker-structured GMM with a total of $K = 2^7 = 128$ components ($B = 7$ feedback bits per MT), where $K_{\text{tx}} = 32$ and $K_{\text{rx}} = 4$. The proposed GMM-based

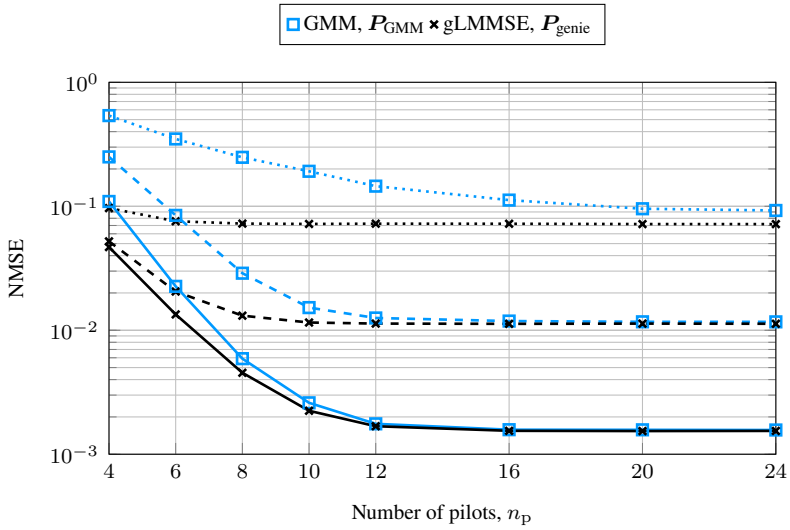


Figure 4.5: The NMSE over the number of pilots n_{p} for a MISO system ($N_{\text{tx}} = 64$, $N_{\text{rx}} = 1$) with $B = 6$ feedback bits, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [62].

pilot design approach "GMM, $\boldsymbol{P}_{\text{GMM}}$," see Section 4.2.3, outperforms the baselines "GMM, $\{\boldsymbol{P}_{\text{DFT}}, \boldsymbol{P}_{\text{RND}}\}$" by a large margin, where the GMM estimator is used in combination with either random pilot matrices or a random selection of columns of the two times oversampled DFT matrix as the pilot matrix. The baselines "gLMMSE, $\{\boldsymbol{P}_{\text{DFT}}, \boldsymbol{P}_{\text{RND}}\}$" assume that each MT has perfect statistical knowledge of its channel to apply the genie-aided LMMSE approach, see (4.2), for channel estimation in combination with random or DFT-based pilots as described above. The GMM-based scheme outperforms the baseline "gLMMSE, $\boldsymbol{P}_{\text{DFT}}$" over the whole SNR range, and the baseline "gLMMSE, $\boldsymbol{P}_{\text{RND}}$" for SNR values larger than $0\,\text{dB}$. In addition to the assumption of perfect statistical knowledge at each MT about its channel, the baseline "gLMMSE, $\boldsymbol{P}_{\text{genie}}$" further assumes perfect statistical knowledge of the channels of all MTs at the BS to design the pilot matrix using Algorithm 6 from Section 4.2.1. The GMM-based scheme performs only slightly worse than this genie-aided baseline. As in the single-user case, a larger gap is present in the low SNR regime, where the feedback information of each MT obtained through the responsibilities, see (4.25), is less accurate due to high noise.

In Figure 4.8, we analyze the performance of the GMM-based pilot design scheme over the block index t for the same MU-MIMO setup as before at three different SNR levels, i.e., $\text{SNR} \in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. As discussed in Section 4.2.3, at $t = 0$, we

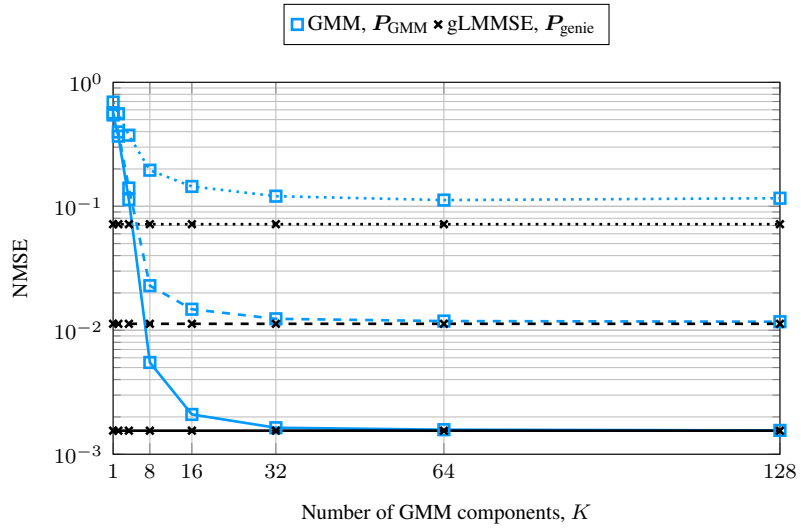


Figure 4.6: The NMSE over the number of components $K = 2^B$ for a MISO system ($N_{\text{tx}} = 64$, $N_{\text{rx}} = 1$) with $n_{\text{p}} = 16$ pilots, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [61] (©2024 IEEE) and [62].

utilize a random selection of columns of the two times oversampled DFT matrix as the pilot matrix. Similarly to the single-user case, see Figure 4.3, after only one block, we can see a significant gain in performance of the proposed GMM-based pilot design scheme due to the feedback of the index of each MT. Additionally, with an increasing SNR, fewer blocks are required to achieve a performance close to the genie-aided baseline "gLMMSE, $\boldsymbol{P}_{\text{genie}}$," which requires perfect statistical knowledge at the BS and the MTs.

In Figure 4.9, for the same MU-MIMO system with $n_{\text{p}} = 16$ pilots, the effect of using the iterative algorithm from [49] following Corollary 2.1 due to the established lower bound in Theorem 2 instead of the iterative approach (Algorithm 6) outlined in Section 4.2.1, subject to a maximum number of iterations $L_{\max}$, is analyzed. Moreover, in addition to the initialization of the iterative algorithm with a random pilot matrix, we consider a random selection of columns of the two times oversampled DFT matrix as an alternative initialization of the pilot matrix. Accordingly, the curves labeled "{GMM, gLMMSE}, $\boldsymbol{P}_{\text{GMM}}$, {LB-RND, LB-DFT}" utilize the lower bound maximization, with either random or DFT-based initialization using the GMM-based feedback scheme, or perfect statistical knowledge at the BS and the MTs, respectively. With "gLMMSE, $\boldsymbol{P}_{\text{genie}}$," we denote the baseline where the iterative algorithm from Section 4.2.1 is used in combination with random initialization and without specifying

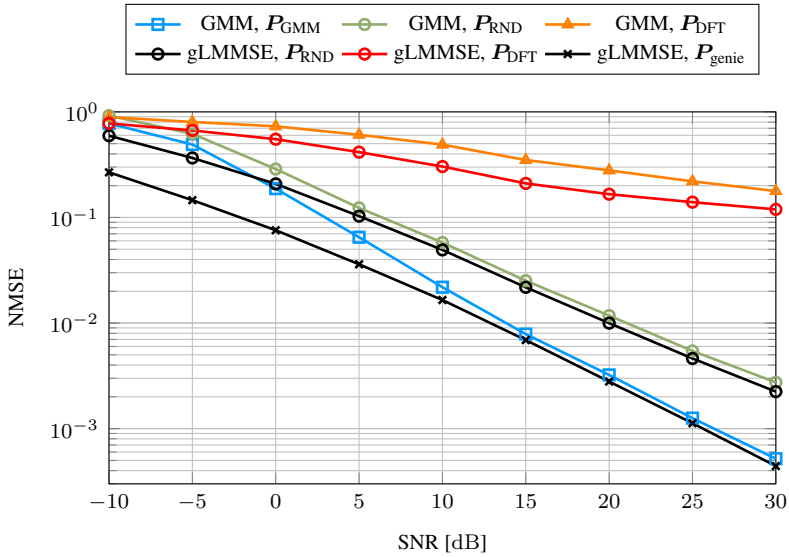


Figure 4.7: The NMSE over the SNR for a MU-MIMO system ($N_{\text{tx}} = 32$, $N_{\text{rx}} = 4$) with $J = 4$ MTs, $B = 7$ feedback bits and $n_{\text{p}} = 8$ pilots. Figure adapted from [62].

the maximum number of iterations $L_{\max}$. In Figure 4.9, we can see that for low values for $L_{\max}$, a random initialization of the iterative lower-bound maximization is advantageous as compared to a DFT-based initialization, irrespective of whether we use the GMM-based scheme or require perfect statistical knowledge. For a high SNR value of $20\,\text{dB}$ (solid), the DFT-based initialization requires larger values of $L_{\max}$ to achieve a similar performance as random initialization. However, for the GMM-based scheme, where an information feedforward from the BS to the MTs is required, DFT-based initialization is advantageous as it eliminates the need for seed exchange and a common random number generator between the BS and MTs, enhancing practicality. We can see further that for an SNR of $0\,\text{dB}$ (dotted), a gap is present between the GMM-based scheme compared to the case that requires perfect statistical knowledge, since the feedback information of each MT obtained through the responsibilities given an observation, see (4.25), is less accurate due to high noise. We have observed this behavior for different numbers of pilots n_{p}. Overall, this analysis suggests that with the lower bound maximization using the iterative algorithm from [49] following Corollary 2.1, together with a relatively low number of maximum iterations $L_{\max}$, a pilot matrix design of relatively low complexity is possible without sacrificing too much performance. In the next simulations, we focus our analysis on DFT-based initialization due to superior practicality compared to random initialization.

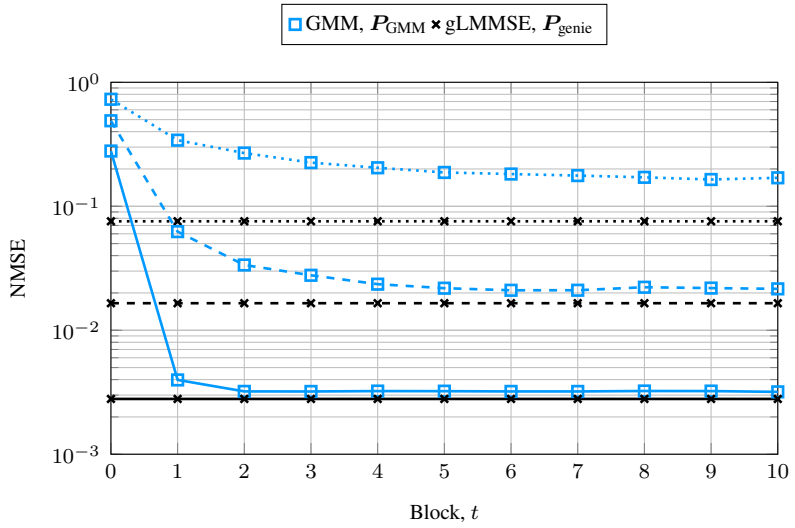


Figure 4.8: The NMSE over the block index t for a MU-MIMO system ($N_{\text{tx}} = 32$, $N_{\text{rx}} = 4$) with $J = 4$ MTs, $B = 7$ feedback bits, $n_{\text{p}} = 8$ pilots, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [62].

In Figure 4.10(a) and Figure 4.10(b), we again consider the MU-MIMO setup from before, with $n_{\text{p}} = 8$ pilots or $n_{\text{p}} = 16$ pilots, respectively. The proposed GMM-based pilot design approaches labeled "GMM, $\boldsymbol{P}_{\text{GMM}}$, $L_{\max} \in \{1, 5, 10, 50\}$" utilize the lower-bound maximization by using the iterative algorithm from [49] following Corollary 2.1 together with DFT-based initialization with different maximum numbers of iterations $L_{\max}$. We can observe that even one iteration ($L_{\max} = 1$) is enough to outperform the baselines "{GMM, gLMMSE}, $\boldsymbol{P}_{\text{DFT}}$," by a large margin. In Figure 4.10(a), for SNR values larger than approximately $5\,\text{dB}$ and even $L_{\max} = 5$, the proposed scheme outperforms "gLMMSE, $\boldsymbol{P}_{\text{genie}}$, RMP," i.e., the state-of-the-art pilot matrix signaling scheme from Section 4.2.2 with $B_{\text{RMP}} = 7$ bits; there, each MT has perfect statistical knowledge of its channel to apply the genie-aided LMMSE approach from (4.2) for channel estimation, and perfect statistical knowledge of the channels of all MTs at the BS is assumed to design the pilot matrix using the iterative algorithm from Section 4.2.1 prior to quantization with the RMP algorithm. The error floor with the RMP-based scheme was already similarly observed and discussed in [50]. In Figure 4.10(b), with $n_{\text{p}} = 16$ pilots, the error floor of the RMP-based scheme is present for SNR values larger than $25\,\text{dB}$. The proposed GMM-based scheme, which does not require any *a priori* statistics, performs very well also in this

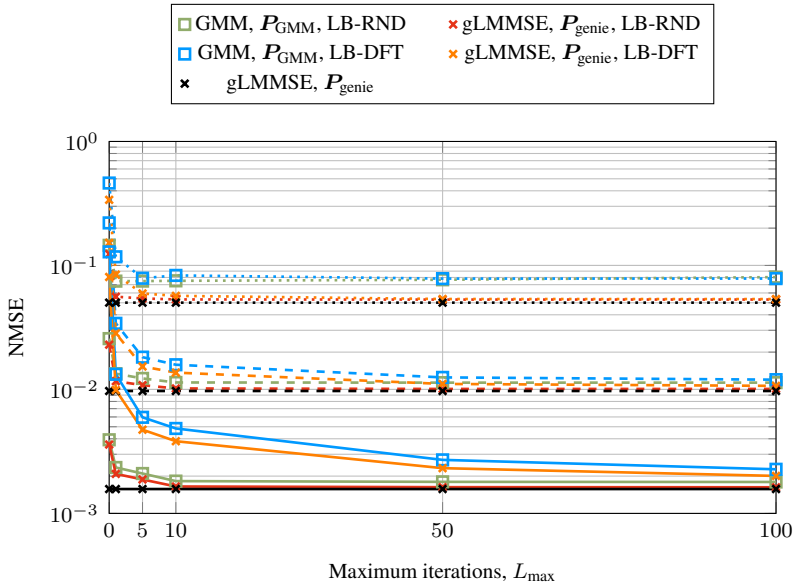


Figure 4.9: The NMSE over the number of maximum iterations $L_{\max}$ for a MU-MIMO system ($N_{\text{tx}} = 32$, $N_{\text{rx}} = 4$) with $J = 4$ MTs, $B = 7$ feedback bits, and $n_{\text{p}} = 16$ pilots for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [62].

setup. Note that in these particular setups, the RMP-based scheme requires in total $n_{\text{p}}B_{\text{RMP}}$ feedforward bits, yielding 56 bits for the setup in Figure 4.10(a), and 112 bits for the setup in Figure 4.10(b), whereas the number of feedforward bits in the GMM-based scheme does not depend on the number of pilots n_{p}, and only requires a total of $JB = 28$ bits.

In Figure 4.11, we illustrate the impact of the number of pilots n_{p} on the performance in the MU-MIMO setup from before for three different SNR levels, i.e., SNR $\in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. The proposed GMM-based pilot design approach labeled "GMM, $\boldsymbol{P}_{\text{GMM}}$" utilizes the lower-bound maximization by using the iterative algorithm from [49] following Corollary 2.1 together with DFT-based initialization with the maximum numbers of iterations $L_{\max} = 50$. The proposed scheme significantly outperforms the baseline "gLMMSE, $\boldsymbol{P}_{\text{genie}}$, RMP," i.e., the state-of-the-art pilot matrix signaling scheme from Section 4.2.2 with $B_{\text{RMP}} = 7$ bits with genie knowledge for the SNR levels $10\,\text{dB}$ and $20\,\text{dB}$, particularly in configurations with low numbers of pilots n_{p}. The advantage of the proposed method lies in determining the pilot matrix based on quantized information associated with GMM components obtained

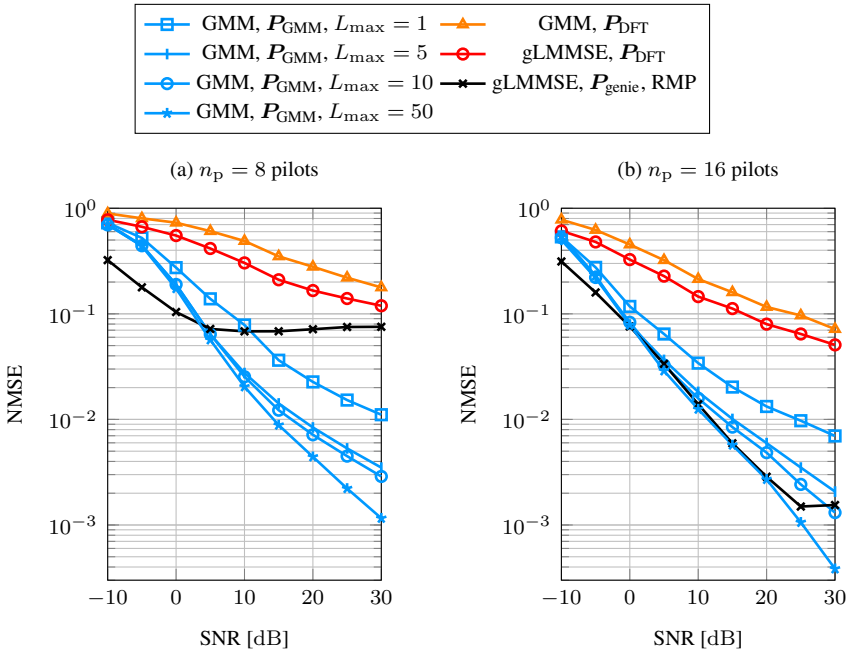


Figure 4.10: The NMSE over the SNR for a MU-MIMO system ($N_{\text{tx}} = 32$, $N_{\text{rx}} = 4$) with $J = 4$ MTs, $B = 7$ feedback bits, and (a) $n_{\text{p}} = 8$ pilots, or (b) $n_{\text{p}} = 16$ pilots. Figure adapted from [62].

through feedforward processing, compared to the baseline's approach of quantizing pilot matrices with DFT vectors after optimization. The baseline suffers from an error floor, particularly in setups with low pilot numbers, as seen in Figure 4.10(a) and discussed in [50]. At a low SNR of $0\,\text{dB}$ the feedback information inferred by each MT through responsibilities given an observation, see (4.25), is less accurate due to high noise, resulting in slightly worse performance compared to the baseline with genie knowledge.

Lastly, in Figure 4.12, we analyze the effect of varying the number of GMM-components K on the scheme's performance in a MU-MISO system with $J = 8$ single-antenna MTs and $N_{\text{tx}} = 64$ BS antennas. We set $n_{\text{p}} = 16$ pilots and consider three different SNR levels, i.e., SNR $\in \{0\,\text{dB}, 10\,\text{dB}, 20\,\text{dB}\}$. Since we consider single-antenna MTs in this setup, we apply the iterative algorithm from [49], utilize DFT-based initialization, and set the maximum number of iterations $L_{\max} = 50$, with the GMM-based scheme. For the "gLMMSE, $\boldsymbol{P}_{\text{genie}}$" baseline, random initialization is applied, and no maximum number of iterations $L_{\max}$ is specified. Similarly to the

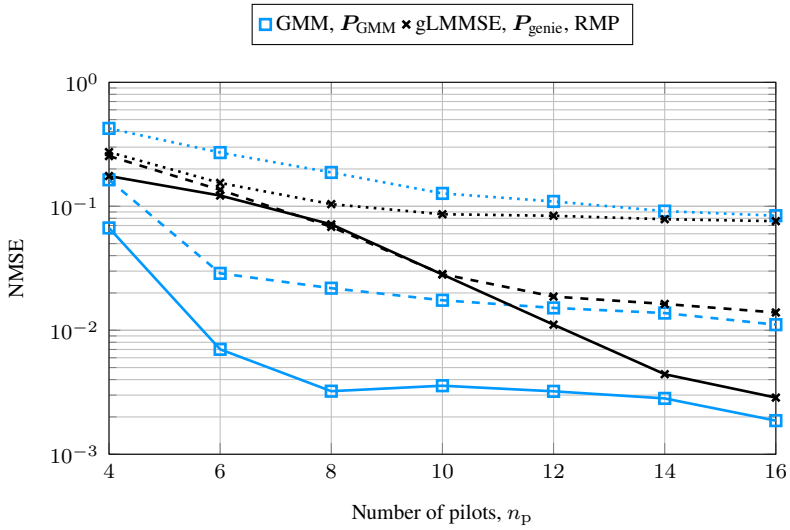


Figure 4.11: The NMSE over the number of pilots n_{p} for a MU-MIMO system ($N_{\text{tx}} = 32$, $N_{\text{rx}} = 4$) with $J = 4$ MTs, $B = 7$ feedback bits, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [62].

single-user case in Figure 4.6, we can observe that the estimation error decreases with an increasing number of components K. With an increasing SNR, the gap to the genie-aided approach "gLMMSE, $\boldsymbol{P}_{\text{genie}}$" decreases. Compared to the single-user case, slightly more GMM-components are needed in a multi-user system. An intuitive explanation is that with the GMM-based scheme, the true but unknown covariance of each MT is approximated, and the approximation error accumulates over multiple MTs. With more components, the approximation error decreases, yielding a better performance. Overall, also in the multi-user case, these results suggest that a performance-to-complexity trade-off can be realized by varying the number of GMM components K.

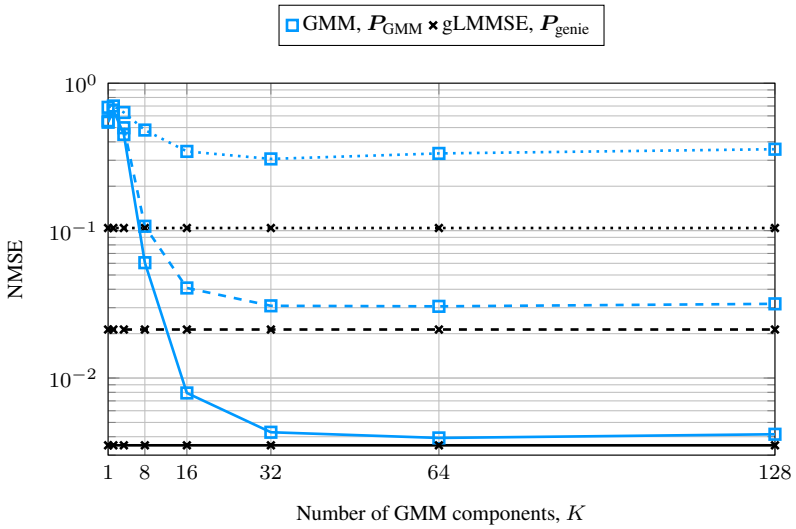


Figure 4.12: The NMSE over the number of components $K = 2^B$ for a MU-MISO system ($N_{\text{tx}} = 64$, $N_{\text{rx}} = 1$) with $J = 8$ MTs, $n_{\text{p}} = 16$ pilots, for different SNR levels (dotted: $0\,\text{dB}$, dashed: $10\,\text{dB}$, solid: $20\,\text{dB}$). Figure adapted from [62].

5 Conclusion and Outlook

In this dissertation, we introduced generative modeling-based feedback schemes for precoder and pilot design in both single- and multi-user MIMO FDD systems.

A novel GMM-based feedback scheme was proposed, which supports codebook construction, feedback encoding, and the generation of channel samples for precoder design. This scheme is characterized by its low computational complexity, enabling parallelization at the MTs, and its adaptability to various SNR levels, numbers of served MTs, pilot configurations, and transmission modes, all without requiring retraining. Building on the GMM-based feedback framework, we proposed methods to adapt a single learned GMM to variable feedback bit budgets by pruning or merging components, thereby enhancing the scheme's versatility. Furthermore, structured GMM variants, such as those employing Kronecker-, Toeplitz-, and circulant-structured covariance matrices, were introduced to reduce memory and model transfer overhead while maintaining high performance. We also developed a GNN-based framework for statistical precoder design that leverages model-based insights to represent statistical channel information compactly. When extended to incorporate GMM-based feedback, this framework demonstrated strong simulation performance, underscoring its potential for efficient feedback and precoding in massive MIMO systems. Additionally, a VQ-VAE-based feedback scheme was proposed as an alternative to the GMM for systems with higher feedback bit budgets, yielding significant gains over conventional approaches. Simulation results confirmed that the proposed schemes enable the deployment of systems with reduced pilot overhead while maintaining performance.

For pilot design, we introduced a GMM-based scheme tailored to single- and multi-user MIMO FDD systems. A key advantage of this approach is its independence from *a priori* knowledge of channel statistics at the BS and the MTs. Instead, it employs a feedback mechanism to establish common knowledge of the pilot matrix in single-user systems. In contrast, in multi-user systems, the pilot design includes an additional feedforward of the GMM component indices. The scheme's versatility allows adaptation to varying SNR levels, pilot configurations, numbers of MTs, and transmission modes without retraining. Simulations confirmed that the performance gains achieved with the proposed scheme allow the deployment of systems with

reduced pilot overhead while maintaining a similar estimation performance.

Based on the findings of this dissertation, the following directions are suggested for future research:

- The GMM-based precoder design scheme assumes perfect CSI for training. Building on the principles proposed in [67], future work could explore training GMMs given noisy data to further improve the scheme's practicality. Additionally, following the approach discussed in [26], learning from noisy data could also be explored for the VQ-VAE-based precoder design scheme.

- The proposed GNN-based framework for precoder design could be extended to systems with multi-antenna MTs by incorporating model-based insights, such as Kronecker factorization, resulting in a scalable and parallelizable solution for MU-MIMO systems.

- The proposed VQ-VAE-based feedback scheme could be integrated into an end-to-end system that jointly learns the pilot matrix and precoder optimization. Similar to the proposed combined GNN- and GMM-based approach, the VQ-VAE could supply feedback information for additional GNN processing, enabling a flexible end-to-end framework that combines generative feedback inference with parallelizable GNN-based precoder design. Furthermore, investigating alternative latent space quantization methods for the VQ-VAE-based scheme, as explored in [42], could facilitate variable feedback bit lengths and reduce computational complexity.

- Future research could extend the GMM-based pilot design scheme to systems with spatio-temporal channel correlation, as studied in [50, 107, 117]. Thereby, the ability of GMMs to exploit temporal channel correlation, validated in [34] for channel prediction, could be further leveraged to improve the pilot design and estimation performance.

- Additionally, the GMM-based pilot design scheme could also be trained using noisy data following [67] to improve its practicality, and incorporating mixture reduction techniques, see Section 3.1.4, could further enhance the scheme's versatility.

- Simulation results revealed that channel estimation benefits from improved GMM-based pilot design across blocks. Investigating the extent to which precoder design via the GMM-based precoder design scheme benefits from the

GMM-based optimized pilots could be pursued. Additionally, the GMM-based pilot design inherently compresses channel information by selecting a specific number of pilots, resulting in lower dimensional observations. Thus, exploring how the enhanced quality of observations impacts precoder design in observation feedback-based approaches, see [2, 4], could also be pursued.

Proofs for Pilot Optimization in MU-MIMO Systems

A

A.1 Proof of Theorem 1

We start by rewriting the sum CMI expression in (4.13) by utilizing for all $j \in \mathcal{J}$, $\boldsymbol{C}_{\boldsymbol{\delta}_j} = \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}$, and the Kronecker product rules, see [118], as

$$\begin{aligned}
&\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})\boldsymbol{C}_{\boldsymbol{\delta}_j}(\boldsymbol{P}_t \otimes \mathbf{I}_{N_{\mathrm{rx}}})^{\mathrm{H}}\right) \\
&= \sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right). \qquad \text{(A.1)}
\end{aligned}$$

The partial derivative of the sum CMI with respect to the entry of the conjugate of the pilot matrix $[\boldsymbol{P}_t^*]_{r,c}$ at row r and column c is then given by

$$\begin{aligned}
&\frac{\partial}{\partial [\boldsymbol{P}_t^*]_{r,c}} \left(\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right) \right) \\
&= \sum_{j=1}^{J} \mathrm{tr}\left(\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right)^{-1} \right. \\
&\qquad \left. \times \frac{\partial}{\partial [\boldsymbol{P}_t^*]_{r,c}} \left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right) \right) \qquad \text{(A.2)} \\
&= \sum_{j=1}^{J} \mathrm{tr}\left(\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right)^{-1} \right. \\
&\qquad \left. \times \left(\tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}^{\mathrm{tx}}_{\boldsymbol{\delta}_j} \mathbf{e}_c \mathbf{e}_r^{\mathrm{T}}) \otimes \boldsymbol{C}^{\mathrm{rx}}_{\boldsymbol{\delta}_j}\right) \right), \qquad \text{(A.3)}
\end{aligned}$$

where (A.2) follows from the linearity of the sum, and using the derivation rule

provided in [119, Eq. (46)]. By utilizing the SVDs

$$\boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \boldsymbol{P}_t^{\mathrm{H}} = \boldsymbol{V}_j \boldsymbol{S}_j \boldsymbol{V}_j^{\mathrm{H}}, \tag{A.4}$$

$$\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}} = \boldsymbol{W}_j \boldsymbol{T}_j \boldsymbol{W}_j^{\mathrm{H}}, \tag{A.5}$$

using $\mathbf{I} = (\boldsymbol{V}_j \otimes \boldsymbol{W}_j)(\boldsymbol{V}_j \otimes \boldsymbol{W}_j)^{\mathrm{H}}$, and applying basic Kronecker product rules, we have

$$\begin{aligned} &\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}\right)^{-1} \\ &= (\boldsymbol{V}_j \otimes \boldsymbol{W}_j)(\mathbf{I} + \tfrac{1}{\sigma_n^2}\boldsymbol{S}_j \otimes \boldsymbol{T}_j)^{-1}(\boldsymbol{V}_j \otimes \boldsymbol{W}_j)^{\mathrm{H}}. \end{aligned} \tag{A.6}$$

Note that $\boldsymbol{S}_j \otimes \boldsymbol{T}_j$ is diagonal. Let $\sum_{i=1}^{I} \alpha_{j,i} \boldsymbol{\gamma}_{j,i} \boldsymbol{\beta}_{j,i}^{\mathrm{T}}$ with $I \triangleq \min(n_{\mathrm{p}}, N_{\mathrm{rx}})$, be the SVD of the matrix $\mathrm{unvec}_{N_{\mathrm{rx}}, n_{\mathrm{p}}}\left(\mathrm{diag}\left((\mathbf{I} + \frac{1}{\sigma_n^2}\boldsymbol{S}_j \otimes \boldsymbol{T}_j)^{-1}\right)\right)$ for all $j \in \mathcal{J}$. Then, it holds for all $j \in \mathcal{J}$,

$$\begin{aligned} &(\mathbf{I} + \tfrac{1}{\sigma_n^2}\boldsymbol{S}_j \otimes \boldsymbol{T}_j)^{-1} \\ &= \sum_{i=1}^{I} \mathrm{diag}(\alpha_{j,i} \boldsymbol{\beta}_{j,i} \otimes \boldsymbol{\gamma}_{j,i}) = \sum_{i=1}^{I} \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \otimes \boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}, \end{aligned} \tag{A.7}$$

which are real-valued Kronecker product SVDs, see [120], with the integer factorization of n_{p} and N_{rx}. Accordingly, the diagonal matrices $\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \in \mathbb{R}^{n_{\mathrm{p}} \times n_{\mathrm{p}}}$ and $\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \in \mathbb{R}^{N_{\mathrm{rx}} \times N_{\mathrm{rx}}}$ for all $j \in \mathcal{J}$, are defined as $\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \triangleq \alpha_{j,i}\,\mathrm{diag}(\boldsymbol{\beta}_{j,i})$ and $\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \triangleq \mathrm{diag}(\boldsymbol{\gamma}_{j,i})$. Note that the specific dimensions (the integer factorization of n_{p} and N_{rx}) are required to proceed with simplifications involving Kronecker products. Incorporating (A.6) and (A.7) into (A.3), we obtain

$$\begin{aligned} \sum_{j=1}^{J} \mathrm{tr}\Bigg(&(\boldsymbol{V}_j \otimes \boldsymbol{W}_j)\left(\sum_{i=1}^{I} \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \otimes \boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}\right)(\boldsymbol{V}_j \otimes \boldsymbol{W}_j)^{\mathrm{H}} \\ &\times \left(\tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \mathbf{e}_c \mathbf{e}_r^{\mathrm{T}}) \otimes \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}\right)\Bigg), \end{aligned} \tag{A.8}$$

which can be further reformulated by exploiting (A.5), basic Kronecker product and

trace rules to

$$\begin{aligned}
&\sum_{j=1}^{J} \operatorname{tr}\left(\left(\sum_{i=1}^{I} \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \otimes \boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}}\right)\right. \\
&\qquad\qquad \left.\times \left(\tfrac{1}{\sigma_n^2}(\boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \mathbf{e}_c \mathbf{e}_r^{\mathrm{T}} \boldsymbol{V}_j) \otimes \boldsymbol{T}_j\right)\right) \\
&= \sum_{j=1}^{J} \operatorname{tr}\left(\sum_{i=1}^{I} \left(\tfrac{1}{\sigma_n^2}(\boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \mathbf{e}_c \mathbf{e}_r^{\mathrm{T}} \boldsymbol{V}_j) \otimes (\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \boldsymbol{T}_j)\right)\right) \qquad \text{(A.9)} \\
&= \sum_{j=1}^{J} \sum_{i=1}^{I} \operatorname{tr}\left(\tfrac{1}{\sigma_n^2} \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \mathbf{e}_c \mathbf{e}_r^{\mathrm{T}} \boldsymbol{V}_j\right) \operatorname{tr}(\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \boldsymbol{T}_j) \qquad \text{(A.10)} \\
&= \sum_{j=1}^{J} \sum_{i=1}^{I} \tfrac{1}{\sigma_n^2} \mathbf{e}_r^{\mathrm{T}} \boldsymbol{V}_j \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \mathbf{e}_c \operatorname{tr}(\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \boldsymbol{T}_j). \qquad \text{(A.11)}
\end{aligned}$$

Accordingly, it holds:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{P}_t^*}\left(\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}(\boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \boldsymbol{P}_t^{\mathrm{H}}) \otimes \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}\right)\right) \\
&= \sum_{j=1}^{J} \sum_{i=1}^{I} \tfrac{1}{\sigma_n^2} \boldsymbol{V}_j \boldsymbol{D}_{\boldsymbol{\beta}_{j,i}} \boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{P}_t \boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{tx}} \operatorname{tr}(\boldsymbol{D}_{\boldsymbol{\gamma}_{j,i}} \boldsymbol{T}_j). \qquad \text{(A.12)}
\end{aligned}$$

It remains to set the derivative of the dual function of the optimization problem (4.13) with respect to $\boldsymbol{P}_t^\star$ to zero to observe the first-order necessary KKT condition provided in (4.14), cf., e.g., [121].

A.2 Proof of Theorem 2

Given the reformulated sum CMI expression from (A.1) and by using (A.4) and (A.5), the sum CMI can be expressed as

$$\begin{aligned}
&\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2} \boldsymbol{S}_j \otimes \boldsymbol{T}_j\right) \\
&= \sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2} \operatorname{tr}(\boldsymbol{T}_j) \boldsymbol{S}_j \otimes \tfrac{1}{\operatorname{tr}(\boldsymbol{T}_j)} \boldsymbol{T}_j\right). \qquad \text{(A.13)}
\end{aligned}$$

By further defining for all $j \in \mathcal{J}$, $\frac{1}{\operatorname{tr}(\boldsymbol{T}_j)}\boldsymbol{T}_j = \sum_{i=1}^{N_{\mathrm{rx}}} t_{j,i} \operatorname{diag}(\mathbf{e}_i)$ with $\sum_{i=1}^{N_{\mathrm{rx}}} t_{j,i} = 1$, we can rewrite the expression provided in (A.13) and lower bound it using the concavity of the log-determinant expression (see [122, Th. 7.6.6]) as

$$\begin{aligned}
&\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j \otimes \left(\sum_{i}^{N_{\mathrm{rx}}} t_{j,i}\operatorname{diag}(\mathbf{e}_i)\right)\right) \\
&= \sum_{j=1}^{J} \log\det \sum_{i}^{N_{\mathrm{rx}}} t_{j,i}\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j \otimes \operatorname{diag}(\mathbf{e}_i)\right) && \text{(A.14)} \\
&\geq \sum_{j=1}^{J}\sum_{i=1}^{N_{\mathrm{rx}}} t_{j,i} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j \otimes \operatorname{diag}(\mathbf{e}_i)\right) && \text{(A.15)} \\
&= \sum_{j=1}^{J}\sum_{i=1}^{N_{\mathrm{rx}}} t_{j,i} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j\right) && \text{(A.16)} \\
&= \sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j\right). && \text{(A.17)}
\end{aligned}$$

Since $\operatorname{tr}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}) = \operatorname{tr}(\boldsymbol{T}_j)$, see (A.5), and by incorporating (A.4) in (4.21), and using $\mathbf{I} = \boldsymbol{V}_j\boldsymbol{V}_j^{\mathrm{H}}$, yields (A.17), finishing the proof.

A.3 Proof of Proposition 2.1

Observe that the inequality in (A.15) becomes an equality if for all $j \in \mathcal{J}$ and any $i', t_{j,i'} = 1$ and $t_{j,i} = 0$ for all $i \neq i'$, i.e., $\operatorname{rk}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}) = 1$ for all $j \in \mathcal{J}$. In any other case, if there is at least one MT j for which $\operatorname{rk}(\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}) > 1$ holds with accordingly $t_{j,i} \in (0,1)$ for all i, and since then

$$\mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j \otimes \operatorname{diag}(\mathbf{e}_i) \neq \mathbf{I} + \tfrac{1}{\sigma_n^2}\operatorname{tr}(\boldsymbol{T}_j)\boldsymbol{S}_j \otimes \operatorname{diag}(\mathbf{e}_{i'}) \tag{A.18}$$

due to the structure of the resulting matrices for $i \neq i'$, the condition for equality between (A.14) and (A.15) according to [122, Th. 7.6.6] is not met (equality in (A.18) is required).

A.4 Proof of Proposition 2.2

The lower bound expression from (4.21) only depends on the trace of each MT's receive-side covariance matrix $\boldsymbol{C}_{\boldsymbol{\delta}_j}^{\mathrm{rx}}$ for all $j \in \mathcal{J}$. We outline next that the sum CMI attains its largest value and, consequently, exhibits the largest gap to the lower bound

for the case of no spatial correlation for all $j \in \mathcal{J}$. Thus, given the reformulated sum CMI from (A.13), we have to show

$$\begin{aligned}&\sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\tau_j \boldsymbol{S}_j \otimes \tfrac{1}{N_{\mathrm{rx}}}\mathbf{I}\right)\\ &\overset{!}{>} \sum_{j=1}^{J} \log\det\left(\mathbf{I} + \tfrac{1}{\sigma_n^2}\tau_j \boldsymbol{S}_j \otimes \tfrac{1}{N_{\mathrm{rx}}}(\mathbf{I} + \mathrm{diag}(\boldsymbol{\kappa}_j))\right),\end{aligned} \tag{A.19}$$

where for all $j \in \mathcal{J}$, the elements of $\boldsymbol{\kappa}_j$ satisfy $\sum_{\ell=1}^{N_{\mathrm{rx}}} \kappa_{j,\ell} = 0$, with $\kappa_{j,\ell} + 1 \geq 0$, and $\boldsymbol{\kappa}_j \neq \mathbf{0}$ (in this way, for all $j \in \mathcal{J}$, the eigenvalues of any valid receive-side covariance matrix, except the weighted identity, are parametrized). Since the involved matrices are diagonal, we have

$$\begin{aligned}&\sum_{j=1}^{J} \log \prod_{i=1}^{n_{\mathrm{p}}} (1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i})^{N_{\mathrm{rx}}}\\ &\overset{!}{>} \sum_{j=1}^{J} \log \prod_{i=1}^{n_{\mathrm{p}}} \prod_{\ell=1}^{N_{\mathrm{rx}}} (1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i}(1 + \kappa_{j,\ell})),\end{aligned} \tag{A.20}$$

which can be further rewritten to

$$\begin{aligned}&\sum_{j=1}^{J} \sum_{i=1}^{n_{\mathrm{p}}} \log(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i})^{N_{\mathrm{rx}}}\\ &\overset{!}{>} \sum_{j=1}^{J} \sum_{i=1}^{n_{\mathrm{p}}} \sum_{\ell=1}^{N_{\mathrm{rx}}} \log(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i}(1 + \kappa_{j,\ell})).\end{aligned} \tag{A.21}$$

For a fixed j and i, we thus have to show

$$N_{\mathrm{rx}} \log(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i}) \overset{!}{>} \sum_{\ell=1}^{N_{\mathrm{rx}}} \log(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i}(1 + \kappa_{j,\ell})), \tag{A.22}$$

which can be reformulated by dividing both sides by N_{rx} as

$$\begin{aligned}&\log\left(\sum_{\ell=1}^{N_{\mathrm{rx}}} \tfrac{1}{N_{\mathrm{rx}}}(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i})\right)\\ &\overset{!}{>} \sum_{\ell=1}^{N_{\mathrm{rx}}} \tfrac{1}{N_{\mathrm{rx}}} \log(1 + \tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}} s_{j,i}(1 + \kappa_{j,\ell})).\end{aligned} \tag{A.23}$$

Observing that the left-hand sides of (A.23) and (A.24) are equal and the following inequality is true due to Jensen's inequality

$$\begin{aligned}&\log\left(\sum_{\ell=1}^{N_{\mathrm{rx}}}\tfrac{1}{N_{\mathrm{rx}}}(1+\tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}}s_{j,i}(1+\kappa_{j,\ell}))\right)\\&>\sum_{\ell=1}^{N_{\mathrm{rx}}}\tfrac{1}{N_{\mathrm{rx}}}\log(1+\tfrac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}}s_{j,i}(1+\kappa_{j,\ell})),\end{aligned}\tag{A.24}$$

and is strict since $\boldsymbol{\kappa}_j \neq \mathbf{0}$, the only condition for which equality can hold is not satisfied, i.e. $1+\frac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}}s_{j,i}(1+\kappa_{j,\ell}) \neq 1+\frac{\tau_j}{\sigma_n^2 N_{\mathrm{rx}}}s_{j,i}(1+\kappa_{j,\ell'})$ for $\ell \neq \ell'$.

Bibliography

[1] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality–What is next? five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3 – 20, 2019, Special Issue on Source Localization in Massive MIMO.

[2] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, 2014.

[3] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018.

[4] D. B. Amor, M. Joham, and W. Utschick, "Asymptotic behavior of zero-forcing precoding based on imperfect channel knowledge for massive MISO FDD systems," in *IEEE Int. Conf. Commun.*, 2023, pp. 1500–1505.

[5] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, 2008.

[6] V. Lau, Y. Liu, and T.-A. Chen, "On the design of MIMO block-fading channels with feedback-link capacity constraint," *IEEE Trans. Commun.*, vol. 52, no. 1, pp. 62–70, 2004.

[7] N. Ravindran and N. Jindal, "Limited feedback-based block diagonalization for the MIMO broadcast channel," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1473–1482, 2008.

[8] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, 2016.

[9] J. Jang, H. Lee, S. Hwang, H. Ren, and I. Lee, "Deep learning-based limited feedback designs for MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 558–561, 2020.

[10] N. Turan, M. Koller, S. Bazzi, W. Xu, and W. Utschick, "Unsupervised learning of adaptive codebooks for deep feedback encoding in FDD systems," in *55th Asilomar Conf. Signals, Syst., Comput.*, 2021, pp. 1464–1469.

[11] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4044–4057, 2021.

[12] Y. Li, K. Li, L. Cheng, Q. Shi, and Z.-Q. Luo, "Digital twin-aided learning to enable robust beamforming: Limited feedback meets deep generative models," in *IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 26–30.

[13] J. Guo, C.-K. Wen, M. Chen, and S. Jin, "Environment knowledge-aided massive MIMO feedback codebook enhancement using artificial intelligence," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4527–4542, 2022.

[14] K. Kong, W.-J. Song, and M. Min, "Knowledge distillation-aided end-to-end learning for linear precoding in multiuser MIMO downlink systems with finite-rate feedback," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11 095–11 100, 2021.

[15] J. Jang, H. Lee, I.-M. Kim, and I. Lee, "Deep learning for multi-user MIMO systems: Joint design of pilot, limited feedback, and precoding," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7279–7293, 2022.

[16] W. Utschick, V. Rizzello, M. Joham, Z. Ma, and L. Piazzi, "Learning the CSI recovery in FDD systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6495–6507, 2022.

[17] V. Rizzello, N. Turan, M. Joham, and W. Utschick, "Two-sample tests for validating the UL-DL conjecture in FDD systems," in *17th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2021, pp. 1–6.

[18] B. Fesl, N. Turan, M. Koller, M. Joham, and W. Utschick, "Centralized learning of the distributed downlink channel estimators in FDD systems using uplink data," in *25th Int. ITG Workshop Smart Antennas (WSA)*, 2021.

[19] M. Baur, V. Rizzello, N. Alvarez Prado, and W. Utschick, "Empirical evaluation of distributional shifts in FDD systems based on ray-tracing," in *WSA & SCC;*

26th Int. ITG Workshop Smart Antennas and 13th Conf. Syst., Commun., Coding, 2023.

[20] N. Turan, M. Koller, V. Rizzello, B. Fesl, S. Bazzi, W. Xu, and W. Utschick, "On distributional invariances between downlink and uplink MIMO channels," in *25th Int. ITG Workshop Smart Antennas (WSA)*, 2021.

[21] 3GPP, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," 3rd Generation Partnership Project (3GPP), Tech. Rep. 38.843 (V18.0.0), Jul. 2024.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[25] E. Balevi, A. Doshi, A. Jalal, A. Dimakis, and J. G. Andrews, "High dimensional channel estimation using deep generative networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 18–30, 2021.

[26] M. Baur, B. Fesl, and W. Utschick, "Leveraging variational autoencoders for parameterized MMSE estimation," *IEEE Trans. Signal Process.*, vol. 72, pp. 3731–3744, 2024.

[27] B. Fesl, M. Baur, F. Strasser, M. Joham, and W. Utschick, "Diffusion-based generative prior for low-complexity MIMO channel estimation," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3493–3497, 2024.

[28] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3133–3143, 2020.

[29] M. Kim, R. Fritschek, and R. F. Schaefer, "Learning end-to-end channel coding with diffusion models," in *WSA & SCC; 26th Int. ITG Workshop Smart Antennas and 13th Conf. Syst., Commun., Coding*, 2023, pp. 1–6.

[30] P. Mukherjee, D. Mishra, and S. De, "Gaussian mixture based context-aware short-term characterization of wireless channels," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 26–40, 2020.

[31] Y. Li, J. Zhang, Z. Ma, and Y. Zhang, "Clustering analysis in the wireless propagation channel with a variational Gaussian mixture model," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 223–232, 2020.

[32] Y. Gu and Y. D. Zhang, "Information-theoretic pilot design for downlink channel estimation in FDD massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2334–2346, 2019.

[33] M. Koller, B. Fesl, N. Turan, and W. Utschick, "An asymptotically MSE-optimal estimator based on Gaussian mixture models," *IEEE Trans. Signal Process.*, vol. 70, pp. 4109–4123, 2022.

[34] N. Turan, B. Böck, K. J. Chan, B. Fesl, F. Burmeister, M. Joham, G. Fettweis, and W. Utschick, "Wireless channel prediction via Gaussian mixture models," in *27th Int. Workshop Smart Antennas (WSA)*, 2024, pp. 1–5.

[35] T. T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan, "Approximation by finite mixtures of continuous density functions that vanish at infinity," *Cogent Math. Statist.*, vol. 7, no. 1, p. 1750861, 2020.

[36] B. Zhao, J. Guo, and C. Yang, "Learning precoding policy: CNN or GNN?" in *IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 1027–1032.

[37] S. Liu, J. Guo, and C. Yang, "Multidimensional graph neural networks for wireless communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3057–3073, 2024.

[38] C. He, Y. Li, Y. Lu, B. Ai, Z. Ding, and D. Niyato, "ICNet: GNN-enabled beamforming for MISO interference channels with statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 73, no. 8, pp. 12 225–12 230, 2024.

[39] V. Rizzello, D. B. Amor, M. Joham, and W. Utschick, "Scalable multi-user precoding and pilot optimization with graph neural networks," in *IEEE Int. Conf. Commun.*, 2024, pp. 2956–2961.

[40] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 6306–6315.

[41] V. Rizzello, M. Nerini, M. Joham, B. Clerckx, and W. Utschick, "User-driven adaptive CSI feedback with ordered vector quantization," *IEEE Wireless Commun. Lett.*, vol. 12, no. 11, pp. 1956–1960, 2023.

[42] J. Shin, Y. Kang, and Y.-S. Jeon, "Vector quantization for deep-learning-based CSI feedback in massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 13, no. 9, pp. 2382–2386, 2024.

[43] J.-H. Lee, J. Lee, and A. F. Molisch, "Generative vs. predictive models in massive MIMO channel prediction," in *58th Asilomar Conf. Signals, Syst., Comput.*, 2024, pp. 1642–1646.

[44] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 5, pp. 802–814, 2014.

[45] J. Kotecha and A. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 546–557, 2004.

[46] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1807–1820, 2010.

[47] J. Pang, J. Li, L. Zhao, and Z. Lu, "Optimal training sequences for MIMO channel estimation with spatial correlation," in *IEEE 66th Veh. Technol. Conf.*, 2007, pp. 651–655.

[48] J. Fang, X. Li, H. Li, and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1935–1947, 2017.

[49] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2882, 2015.

[50] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multiuser massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4732–4744, 2017.

[51] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 6, pp. 1211–1226, 2002.

[52] D. Guo, S. Shamai, and S. Verdu, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[53] X. Ma and Z. Gao, "Data-driven deep learning to design pilot and channel estimator for massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5677–5682, 2020.

[54] M. B. Mashhadi and D. Gündüz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6315–6328, 2021.

[55] N. Turan, M. Koller, B. Fesl, S. Bazzi, W. Xu, and W. Utschick, "GMM-based codebook construction and feedback encoding in FDD systems," in *56th Asilomar Conf. Signals, Syst., Comput.*, 2022, pp. 37–42.

[56] N. Turan, B. Fesl, M. Koller, M. Joham, and W. Utschick, "A versatile low-complexity feedback scheme for FDD systems via generative modeling," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 6251–6265, 2024.

[57] N. Turan, B. Fesl, and W. Utschick, "Enhanced low-complexity FDD system feedback with variable bit lengths via generative modeling," in *57th Asilomar Conf. Signals, Syst., Comput.*, 2023, pp. 363–369.

[58] N. Turan, S. Allaparapu, D. B. Amor, B. Böck, M. Joham, and W. Utschick, "Statistical precoder design in multi-user systems via graph neural networks and generative modeling," *IEEE Wireless Commun. Lett.*, vol. 14, no. 5, pp. 1491–1495, 2025.

[59] N. Turan, B. Fesl, M. Joham, Z. Ma, B. Sheen, W. Xiao, A. C. K. Soong, and W. Utschick, "Limited feedback on measurements: Sharing a codebook or a generative model?" in *IEEE 99th Veh. Technol. Conf. (VTC2024-Spring)*, 2024, pp. 1–7.

[60] N. Turan, M. Baur, J. Li, and W. Utschick, "Feedback design with VQ-VAE for robust precoding in multi-user FDD systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 390–394, 2025.

[61] N. Turan, B. Fesl, B. Böck, M. Joham, and W. Utschick, "Channel-adaptive pilot design for FDD-MIMO systems utilizing Gaussian mixture models," in *19th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2024, pp. 1–6.

[62] N. Turan, B. Böck, B. Fesl, M. Joham, D. Gündüz, and W. Utschick, "A versatile pilot design scheme for FDD systems utilizing Gaussian mixture models," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 4115–4130, 2025.

[63] B. Fesl, N. Turan, M. Koller, and W. Utschick, "A low-complexity MIMO channel estimator with implicit structure of a convolutional neural network," in *IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 11–15.

[64] M. Koller, B. Fesl, N. Turan, and W. Utschick, "An asymptotically optimal approximation of the conditional mean channel estimator based on Gaussian mixture models," in *IEEE Int. Conf. on Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 5268–5272.

[65] N. Turan, B. Fesl, M. Grundei, M. Koller, and W. Utschick, "Evaluation of a Gaussian mixture model-based channel estimator using measurement data," in *Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2022, pp. 1–6.

[66] B. Fesl, M. Joham, S. Hu, M. Koller, N. Turan, and W. Utschick, "Channel estimation based on Gaussian mixture models with structured covariances," in *56th Asilomar Conf. Signals, Syst., Comput.*, 2022, pp. 533–537.

[67] B. Fesl, N. Turan, M. Joham, and W. Utschick, "Learning a Gaussian mixture model from imperfect training data for robust channel estimation," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1066–1070, 2023.

[68] B. Fesl, N. Turan, and W. Utschick, "Low-rank structured MMSE channel estimation with mixtures of factor analyzers," in *57th Asilomar Conf. Signals, Syst., Comput.*, 2023, pp. 375–380.

[69] M. Baur, B. Böck, N. Turan, and W. Utschick, "Variational autoencoder for channel estimation: Real-world measurement insights," in *27th Int. Workshop Smart Antennas*, 2024, pp. 117–122.

[70] M. Baur, N. Turan, B. Fesl, and W. Utschick, "Channel estimation in underdetermined systems utilizing variational autoencoders," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 9031–9035.

[71] B. Fesl, A. Faika, N. Turan, M. Joham, and W. Utschick, "Channel estimation with reduced phase allocations in RIS-aided systems," in *IEEE 24th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2023, pp. 161–165.

[72] F. Weißer, N. Turan, D. Semmler, and W. Utschick, "Data-aided channel estimation utilizing Gaussian mixture models," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 8886–8890.

[73] F. Weißer, D. Semmler, N. Turan, and W. Utschick, "Data-aided MU-MIMO channel estimation utilizing Gaussian mixture models," in *IEEE Int. Conf. Commun.*, 2024, pp. 6684–6689.

[74] F. Weißer, N. Turan, and W. Utschick, "DoA-aided MMSE channel estimation for wireless communication systems," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[75] B. Fesl, M. Koller, N. Turan, and W. Utschick, "Learning a low-complexity channel estimator for one-bit quantization," in *54th Asilomar Conf. Signals, Syst., Comput.*, 2020, pp. 393–397.

[76] B. Fesl, N. Turan, B. Böck, and W. Utschick, "Channel estimation for quantized systems based on conditionally Gaussian latent models," *IEEE Trans. Signal Process.*, vol. 72, pp. 1475–1490, 2024.

[77] N. Turan, M. Koller, and W. Utschick, "One-bit quantized channel prediction with neural networks," in *IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2021, pp. 604–609.

[78] B. Böck, M. Baur, N. Turan, D. Semmler, and W. Utschick, "A statistical characterization of wireless channels conditioned on side information," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3508–3512, 2024.

[79] M. Baur, N. Turan, S. Wallner, and W. Utschick, "Evaluation metrics and methods for generative models in the wireless PHY layer," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 677–689, 2025.

[80] D. Neumann, T. Wiese, and W. Utschick, "Learning the MMSE channel estimator," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2905–2917, 2018.

[81] Y. Tsai, L. Zheng, and X. Wang, "Millimeter-wave beamformed full-dimensional MIMO channel estimation based on atomic norm minimization," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6150–6163, 2018.

[82] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Press, 2005.

[83] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. on Telecommun.*, vol. 10, pp. 585–595, 1999.

[84] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, 2011.

[85] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, 2004.

[86] V. Stankovic and M. Haardt, "Generalized design of multi-user MIMO precoding matrices," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 953–961, 2008.

[87] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[88] H. Lee, K. Lee, B. M. Hochwald, and I. Lee, "Regularized channel inversion for multiple-antenna users in multiuser MIMO downlink," in *IEEE Int. Conf. Commun.*, 2008, pp. 3501–3505.

[89] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, 2008.

[90] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1394–1410, 2021.

[91] 3GPP, "Spatial channel model for multiple input multiple output (MIMO) simulations," 3rd Generation Partnership Project (3GPP), Tech. Rep. 25.996 (V16.0.0), Jul. 2020.

[92] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "Quadriga: A 3-d multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.

[93] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein, "Quadriga: Quasi deterministic radio channel generator, user manual and documentation," Fraunhofer Heinrich Hertz Institute, Tech. Rep., v2.2.0, 2019.

[94] C. Hellings, A. Dehmani, S. Wesemann, M. Koller, and W. Utschick, "Evaluation of neural-network-based channel estimators using measurement data," in *Proc. Int. ITG Workshop on Smart Antennas (WSA)*, 2019, pp. 1–5.

[95] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[96] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[97] R. Hunger, D. A. Schmidt, M. Joham, and W. Utschick, "A general covariance-based optimization framework using orthogonal projections," in *9th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2008, pp. 76–80.

[98] C. Hennig, "Methods for merging Gaussian mixture components," *Adv. Data Anal. Classif.*, vol. 4, p. 3–34, 2010.

[99] A. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 3, pp. 989–999, 2007.

[100] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," in *14th Int. Conf. Inf. Fusion*, 2011.

[101] D. Love and R. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, 2005.

[102] W. Santipach and M. Honig, "Asymptotic capacity of beamforming with limited feedback," in *Int. Symp. Inf. Theory. (ISIT)*, 2004, p. 290.

[103] W. Dai, Y. Liu, and B. Rider, "Quantization bounds on Grassmann manifolds and applications to MIMO communications," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1108–1123, 2008.

[104] F. Mezzadri, "How to generate random matrices from the classical compact groups," *Notices Amer. Math. Soc.*, vol. 54, no. 5, pp. 592 – 604, May 2007.

[105] M. Razaviyayn, M. S. Boroujeni, and Z.-Q. Luo, "A stochastic weighted MMSE approach to sum rate maximization for a MIMO interference channel," in *IEEE 14th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2013, pp. 325–329.

[106] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Math. Program.*, vol. 157, no. 2, p. 515–545, 2016.

[107] S. Noh, M. D. Zoltowski, Y. Sung, and D. J. Love, "Pilot beam pattern design for channel estimation in massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 787–801, 2014.

[108] G. H. Golub and C. F. van Loan, *Matrix Computations*, 2nd ed. Johns Hopkins University Press, 1989.

[109] M. Gharavi-Alkhansari and T. Huang, "A fast orthogonal matching pursuit algorithm," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 3, 1998, pp. 1389–1392.

[110] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 2909–2913.

[111] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, p. 281–305, Feb. 2012.

[112] D. Yang, L.-L. Yang, and L. Hanzo, "DFT-based beamforming weight-vector codebook design for spatially correlated channels in the unitary precoding aided multiuser downlink," in *IEEE Int. Conf. Commun.*, 2010, pp. 1–5.

[113] J. Li, X. Su, J. Zeng, Y. Zhao, S. Yu, L. Xiao, and X. Xu, "Codebook design for uniform rectangular arrays of massive antennas," in *IEEE 77th Veh. Technol. Conf. (VTC Spring)*, 2013, pp. 1–5.

[114] 3GPP, "NR; Physical channels and modulation," 3rd Generation Partnership Project (3GPP), Tech. Spec. 38.211 (V18.0.0), Sep. 2023.

[115] F. Kaltenberger, M. Kountouris, D. Gesbert, and R. Knopp, "On the trade-off between feedback and capacity in measured MU-MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4866–4875, 2009.

[116] S. Bazzi and W. Xu, "On the amount of downlink training in correlated massive MIMO channels," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2286–2299, 2018.

[117] B. Böck, M. Baur, V. Rizzello, and W. Utschick, "Variational inference aided estimation of time varying channels," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[118] K. Schäcke, "On the Kronecker product," Master's thesis, Univ. Waterloo, Waterloo, ON, Canada, 2004.

[119] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Tech. Univ. Denmark, Lyngby, Denmark, 2012.

[120] C. F. V. Loan, "The ubiquitous Kronecker product," *J. Comput. Appl. Math.*, vol. 123, no. 1, pp. 85–100, 2000.

[121] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[122] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.:, Cambridge Univ. Press, 2012.