UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Scaling All-Digital Millimeter-Wave Massive Multiuser MIMO

Permalink

https://escholarship.org/uc/item/5s2879rp

Author

Abdelghany, Mohammed

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Scaling All-Digital Millimeter-Wave Massive Multiuser MIMO

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Mohammed A. Abdelghany

Committee in charge:

Professor Upamanyu Madhow, Chair
Professor Ali M. Niknejad, University of California, Berkeley
Professor Andreas F. Molisch, University of Southern California
Professor Mark J. W. Rodwell
Professor Ramtin Pedarsani

September 2021

The Dissertation of Mohammed A. Abdelghany is approved.

Professor Ali M. Niknejad, University of California, Berkeley

---

Professor Andreas F. Molisch, University of Southern California

---

Professor Mark J. W. Rodwell

Professor Ramtin Pedarsani

Professor Upamanyu Madhow, Committee Chair

July 2021

Scaling All-Digital Millimeter-Wave Massive Multiuser MIMO

Copyright © 2021

by

Mohammed A. Abdelghany

To my loving family, loyal friends, helpful colleagues, and

supportive teachers

Your kindness and generosity have humbled me

Thank You

Acknowledgements

The research work presented in this dissertation would not have been possible without the support of many individuals and entities. Firstly, I would like to express my sincere gratitude to my advisor, Prof. Upamanyu Madhow. Prof. Madhow has spared no effort or expense to remove all obstacles I faced during my Ph.D. journey. I learned so many things from him, academically and on a personal level.

Next, I would like to acknowledge the vital and irreplaceable contribution of our collaborators, Prof. Mark Rodwell, Prof. Antti Tölli, Doctor Maryam Rasekh, and Doctor Ali Farid. Also, I want to thank Doctor Belal Korany, Doctor Ahmed Elshafiy, the WCSL group, Rodwell group, and the whole SRC community for their insightful discussions and help.

I would like to pay all respect and appreciation to the committee in charge, Prof. Ali Niknejad, Prof. Andreas Molisch, Prof. Mark Rodwell, and Prof. Ramtin Pedarsani. They have dedicated valuable time and effort and placed this dissertation under scrutiny to ensure the highest quality of our work.

Many thanks to UCSB faculty members who taught me in class and shared valuable insights and experiences with their students. I especially like to mention Prof. Upamanyu Madhow, Prof. Kenneth Rose, Prof. Ramtin Pedarsani, Prof. João Hespanha, Prof. Shiv Chandrasekaran, and Prof. Jason Marden; I must say it was a privilege and a great pleasure to be once among your students.

I am thankful for the UC system, which has constructed an excellent environment for education and research and is working vigorously to improve the students' quality of life. I want to thank our amazing university staff members for their hard work, dedication, and tenacity, especially the student affairs manager Val De Veyra.

My research was supported in part by the Semiconductor Research Corporation (SRC)

under the JUMP program (2018-JU-2778) and by DARPA (HR0011-18-3-0004). Use was made of the computational facilities administered by the Center for Scientific Computing at the CNSI and MRL (an NSF MRSEC; DMR-1720256) and purchased through NSF CNS-1725797.

Curriculum Vitæ

Mohammed A. Abdelghany

Education

2021	Ph.D. in Electrical and Computer Engineering, University of California, Santa Barbara, USA.
2016	M.Sc. in Electronics and Electrical Communication Engineering, Faculty of Engineering, Cairo University, Egypt.
2012	B.Sc. in Electronics and Electrical Communication Engineering, Faculty of Engineering, Cairo University, Egypt.

Publications

1. **Mohammed A. Abdelghany**, Ali A. Farid, Maryam Eslami Rasekh, Upamanyu Madhow, and Mark J. W. Rodwell. *A design framework for all-digital mmWave massive MIMO with per-antenna nonlinearities.* In IEEE Transactions on Wireless Communications (2021).
2. Ali A. Farid, **Mohammed A. Abdelghany**, Upamanyu Madhow, and Mark J. W. Rodwell. *Dynamic Range Requirements of Digital vs. RF and Tiled Beamforming in mm-Wave Massive MIMO.* In 2021 IEEE Radio and Wireless Symposium (RWS).
3. **Mohammed A. Abdelghany**, Maryam Eslami Rasekh, and Upamanyu Madhow. *Scalable Nonlinear Multiuser Detection for mmWave Massive MIMO.* In 2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).
4. Arda Simsek, Seong-Kyun Kim, **Mohammed A. Abdelghany**, Ahmed S. H. Ahmed, Ali A. Farid, Upamanyu Madhow, and Mark J. W. Rodwell. *A 146.7 GHz Transceiver with 5 GBaud Data Transmission using a Low-Cost Series-Fed Patch Antenna Array Through Wirebonding Integration.* In 2020 IEEE Radio and Wireless Symposium (RWS).
5. **Mohammed A. Abdelghany**, Upamanyu Madhow, and Mark J. W. Rodwell. *An Efficient Digital Backend for Wideband Single-Carrier mmWave Massive MIMO.* In 2019 IEEE Global Communications Conference (GLOBECOM).
6. **Mohammed A. Abdelghany**, Upamanyu Madhow, and Antti Tölli. *Efficient Beamspace Downlink Precoding for mmWave Massive MIMO.* In 2019 53rd Asilomar Conference on Signals, Systems, and Computers.
7. **Mohammed A. Abdelghany**, Upamanyu Madhow, and Antti Tölli. *Beamspace Local LMMSE: An Efficient Digital Backend for mmWave Massive MIMO.* In 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC).

8. Ahmed M. El-Shafiey, **Mohammed A. Abdelghany**, Mohamed E. Farag, Omar A. Nasr, and Hossam A. H. Fahmy. *On Optimization of Mixed-Radix FFT: A Signal Processing Approach*. In 2019 IEEE Wireless Communications and Networking Conference (WCNC).
9. Maryam Eslami Rasekh, **Mohammed A. Abdelghany**, Upamanyu Madhow, and Mark J. W. Rodwell. *Phase Noise Analysis for mmWave Massive MIMO: A Design Framework for Scaling via Tiled Architectures*. In 2019 53rd Annual Conference on Information Sciences and Systems (CISS).
10. **Mohammed A. Abdelghany**, Ali A. Farid, Upamanyu Madhow, and Mark J. W. Rodwell. *Towards All-Digital mmWave Massive MIMO: Designing around Nonlinearities*. In 2018 52nd Asilomar Conference on Signals, Systems, and Computers.
11. Haitham Hassanieh, Omid Abari, Michael Rodriguez, **Mohammed A. Abdelghany**, Dina Katabi, and Piotr Indyk. *Fast Millimeter Wave Beam Alignment*. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication.
12. **Mohammed A. Abdelghany**, Hamed Mohsenian-Rad, and Mahnoosh Alizadeh. *Wholesale Electricity Pricing in the Presence of Geographical Load Balancing*. In 2017 51st Asilomar Conference on Signals, Systems, and Computers.
13. Ezzeldin Hamed, Hariharan Rahul, **Mohammed A. Abdelghany**, and Dina Katabi. *Real-Time Distributed MIMO Systems*. In Proceedings of the 2016 ACM SIGCOMM Conference.
14. **Mohammed A. Abdelghany**, Mohammed Ismail, Mohsen Raafat, Ali A. Farid, Mohammed Raghieb, Nassr Ismail, Sherif Hafez, Ahmed El-Kady, Esmaail El-Sayed, Mohamed Sharaf, Ibrahim Shazly, Wael Abd El-Kawi, Chadi Mohamed, Mohamed Elhidery, Karim Mohammed, and Omar A. Nasr. *A Highly Scalable Vector Oriented ASIP-Based Multi-Standard Digital Receiver*. In 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS).
15. **Mohammed A. Abdelghany**, Ahmed M. El-Shafiey, Mohamed E. Farag, Omar A. Nasr, and Hossam A. H. Fahmy. *Speeding-up Fast Fourier Transform*. In 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS).
16. Ahmed M. El-Shafiey, Mohamed E. Farag, **Mohammed A. Abdelghany**, Omar A. Nasr, and Hossam A. H. Fahmy. *Two-Stage Optimization of CORDIC-Friendly FFT*. In 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS).
17. Ezzeldin Hamed, Hariharan Rahul, **Mohammed A. Abdelghany**, and Dina Katabi. *A Real-Time 802.11 Compatible Distributed MIMO System*. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication.
18. **Mohammed A. Abdelghany**, Omar A. Nasr, and Karim Osama. *A CORDIC-Friendly FFT Architecture*. In 2014 International Wireless Communications and Mobile Computing Conference (IWCMC).

Abstract

Scaling All-Digital Millimeter-Wave Massive Multiuser MIMO

by

Mohammed A. Abdelghany

All-digital architectures enable taking full advantage of the large number of antennas that can be integrated into mmWave transceivers, with fully flexible beamforming that enables the number of simultaneous users K sharing the band to scale with the number of antennas N. The small carrier wavelength in these bands allows realization of antenna arrays with a large number of elements with a relatively small footprint, opening a path to truly massive Multiple Input Multiple Output (MIMO) systems. However, two key bottlenecks to realizing this potential are the cost/power consumption of Radio Frequency (RF) frontends at high carrier frequencies and the high complexity incurred in the digital baseband processing due to the large number of antennas. In this dissertation, we develop approaches for addressing these bottlenecks by adapting signal processing architectures and algorithms to hardware design considerations, while taking advantage of the unique characteristics of the mmWave band. We first develop an analytical model for the impact of nonlinearities such as RF amplifiers and Analog-to-Digital Converters (ADCs) on the performance of a mmWave massive MIMO uplink. We illustrate the utility of this model in providing specific guidelines for hardware design based on desired system-level performance. For example, the framework allows specification of the desired 1 dB compression point for RF amplifiers and the desired number of bits of ADC precision in order for the system outage at a target bit error rate to be below 5%. These hardware design prescriptions depend on coarse system-level parameters such as the number of antennas N, the number of simultaneous users K, and the maximum and minimum link distances to

be supported. An important conclusion from the analytical framework is that hardware specifications can be substantially relaxed by reducing the load factor, defined as the ratio $\beta = K/N$.

We then consider the problem of scaling digital signal processing in this regime. For a relatively small number of antennas and users, the Linear Minimum Mean Squared Error (MMSE) approach is a standard technique for handling multiuser interference at reasonable complexity. However, for fixed load factor β, the complexity of computing the LMMSE detector, as well as the complexity of using it for demodulation, grow polynomially with the number of antennas. We propose complexity reduction techniques that substantially improve scaling by taking advantage of the spatial sparsity of the mmWave channel. Specifically, we use a spatial Discrete Fourier transform (DFT) across antennas to create N discrete beams, transforming from antenna space to *beamspace.* We show that each user's energy is concentrated in a small number of DFT bins in beamspace. Assuming ideal single path channels, we show that each user can be demodulated reliably using a *local* LMMSE detector which employs a beamspace window whose size does not scale with the number of antennas. The local LMMSE detector approaches the performance of standard LMMSE detection at substantially reduced complexity, and these performance-complexity tradeoffs become more favorable at lower system load factor β. For larger load factors, the beamspace window required by the local LMMSE detector increases, but we show that it is possible to scale well in such regimes by adding a layer of nonlinear interference cancellation on top of the local LMMSE receiver.

Next, drawing on the duality between uplink linear multiuser detection and downlink linear precoding, we demonstrate the efficacy of beamspace techniques for linear precoding on the downlink, in order to reduce the interference seen by a given user due to signals directed at other users by the base station.

Finally, we address the problem of simultaneous scaling of bandwidth and number

of antennas. As bandwidth and hence symbol rate increase, the signal from a given user impinging on a large antenna array incurs a multi-symbol delay spread across the array, which smears the spatial frequency for each user across the band. We introduce a novel technique that combines DFTs in the spatial and time domains, together with an interpolation technique that limits the dispersion of spatial frequency across the band. We show that this results in significantly reduced complexity in computing LMMSE weights for uplink multiuser detection.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The potential of massive scale multiuser MIMO for meeting the ever-increasing demand for wireless mobile data is well understood [2, 3]. Massive MIMO becomes particularly attractive as we move up in the frequency spectrum toward millimeter wave (mmWave) and terahertz (THz) frequencies, where bandwidth is plentiful, and the wavelength is small enough to fit a large number of antennas on moderately sized platforms. By utilizing these advantages, mmWave massive MIMO can potentially support tens or hundreds of simultaneous users with per-user data rates of multiple gigabits/second (Gbps).

Two key bottlenecks to realizing this potential are the cost/power consumption of radio frequency (RF) frontends at high carrier frequencies and the high complexity incurred in the digital baseband processing due to the large number of antennas.

- **RF Frontend:** The cost and power consumption of the RF frontends significantly increase in massive MIMO systems due to the increased number of required RF chains. Moreover, the linearity requirement of the RF frontend imposes design challenges in high-frequency RF components. This requirement becomes even more significant for multicarrier systems (such as Orthogonal Frequency Division Multiplexing (OFDM)), since these systems typically suffer from high Peak-to-Average

Power Ratio (PAPR), requiring high dynamic range for the RF frontends. Thus, some researchers have suggested utilizing single-carrier operation for mmWave Massive MIMO systems.

- **Digital Baseband Processing:** Typical receivers for multiuser systems utilize linear interference suppression using Zero Forcing (ZF) or linear Minimum Mean Square Error (LMMSE) criteria, both of which rely on matrix inversions. The complexity of the matrix inversion operations is $O(N^3)$ where N is the number of antennas. This complexity becomes prohibitive for massive MIMO systems, which have huge number of antennas.

In this dissertation, we address the aforementioned bottlenecks as follows. First, we provide an analytical framework to design multiuser massive MIMO receivers in the presence of RF/baseband nonlinearities. More specifically, we provide analytical guidelines for maximum permissible levels of nonlinearities originating from Analog-to-Digital Converters (ADC) and baseband/passband amplifiers, in order to provide the desired system-level performance guarantees (e.g., on outage probabilities).

Second, we focus our attention to reducing the complexity of the digital baseband processing of fully-digital mmWave massive MIMO systems. More specifically, we propose different algorithms that exploit the sparse structure of the mmWave channels in beamspace to reduce the computational complexity of typical linear MMSE receivers, in both uplink and downlink. For even larger number of users, we propose to enhance the receiver's performance by adding a layer of nonlinear interference cancellation on top of the local LMMSE receiver. Furthermore, single-carrier operation, which is envisioned for reducing the cost of RF frontends as described earlier, adds more challenges to the baseband design, since the wideband channel results in multi-symbol delay spread across the receiver array. For such scenarios, we propose novel techniques for limiting the spread

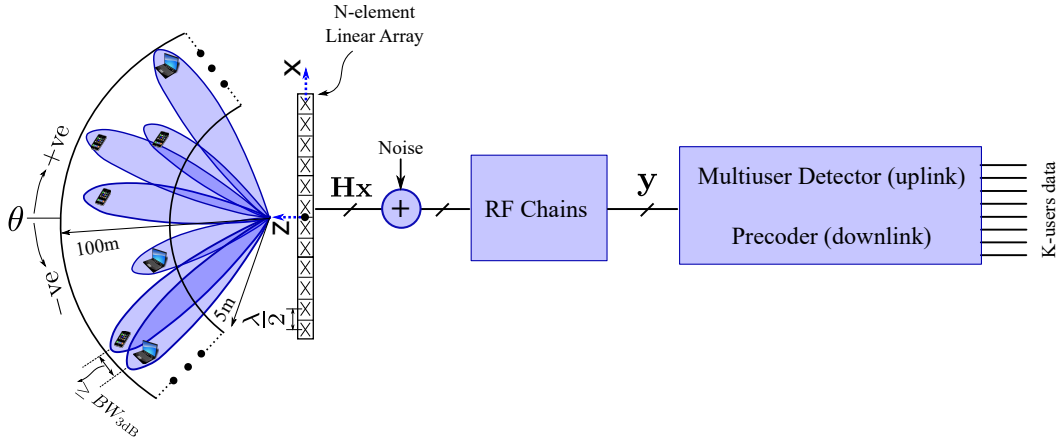


Figure 1.1: The system model considered in this dissertation: The cell size is constrained radially between $R_{\min}$ and $R_{\max}$ and angularly between $-\pi/3 \leq \theta \leq \pi/3$. $\mathrm{BW}_{3\mathrm{dB}}$ stands for the 3 dB beamwidth.

of each user's spatial frequency across the band, which, in turn, will further reduce the complexity of the beamformer weight acquisition.

1.1 Concept System

Fig. 1.1 shows the system model. The base station performs horizontal scanning with a 1D half-wavelength spaced N-element array. Let K denote the number of simultaneous users, and $\beta = \frac{K}{N}$ the *load factor.*

Throughout the dissertation, we assume a line-of-sight (LoS) channel between the base station and each mobile. The direction of arrival (DoA) from the k^{th} mobile is denoted by θ_k, and corresponds to spatial frequency $\Omega_k = 2\pi \frac{d_x}{\lambda} \sin\theta_k$, where λ denotes the carrier wavelength and d_x denotes the inter-element spacing, set to $\frac{\lambda}{2}$ in our numerical results. The $N \times 1$ spatial channel for mobile k is given by

$$\mathbf{h}_k = A_k e^{j\phi_k} \left[1\, e^{j\Omega_k}\, e^{j2\Omega_k}\, \ldots\, e^{j(N-1)\Omega_k}\right]^{\intercal}, \tag{1.1}$$

where ϕ_k is an arbitrary phase shift and $A_k^2 = \left(\frac{\lambda}{4\pi R_k}\right)^2$ depends on the radial location R_k of mobile k, using the Friis formula for path loss [4].

Most of the work reported in this discussion focuses on the uplink from mobiles to the base station, where the main signal processing task is multiuser detection of signals from different mobiles which may interfere with each other. However, analogous concepts also apply to the downlink from base station to mobiles, where the task is to precode the transmitted signal from the base station such that the signal received at a given mobile is minimally corrupted by signals directed at other mobiles.

The cascade of the system's RF nonlinearities is modeled as a complex baseband equivalent nonlinearity $g(\cdot)$. Focusing our discussion on the uplink, the complex baseband received signal vector $\mathbf{z}$ at the base station is therefore given by

$$\mathbf{z} = g(\mathbf{y}) = g\left(\underbrace{\mathbf{H}}_{N\times K}\mathbf{x} + \mathbf{n}\right), \tag{1.2}$$

where $\mathbf{H} = [\mathbf{h}_1\mathbf{h}_2 \ldots \mathbf{h}_K]$ is the channel matrix, $\mathbf{x} = [x_1, ..., x_K]^T$ is the vector of symbols (normalized to unit energy: $\mathbb{E}\left[|x_k|^2\right] = 1$) transmitted by the mobiles, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$ is the thermal AWGN vector, and $g(\cdot)$ is the effective per-antenna nonlinearity in complex baseband.

Note: In scenarios where we analyze the system without considering its RF nonlinearities, we set $g(\mathbf{y}) = \mathbf{y}$, i.e. the complex baseband received signal vector becomes $\mathbf{z} = \mathbf{y} = \underbrace{\mathbf{H}}_{N\times K}\mathbf{x} + \mathbf{n}$.

We consider, as a running example, a 256-element linear array ($N = 256$) at 140 GHz carrier frequency, with a symbol rate of 5 Gbaud with QPSK modulation for each supported user. The load factor β ranges from $\beta = \frac{1}{16}$ ($K = 16$ users) to $\beta = \frac{1}{2}$ ($K = 128$ users). Hence, the aggregate data rate ranges from 160 Gbps to 1.28 Tbps.

1.2 Dissertation Contributions

We now provide a high-level overview of the contributions in this thesis, and how they relate to the state of the art. Detailed discussion of relevant prior work is provided in the individual chapters.

- We develop an analytical model for the impact of RF nonlinearities in mmWave massive MIMO systems, and illustrate its utility in providing hardware design guidelines regarding two key challenges: the low available precision of analog-to-digital conversion at high sampling rates, and nonlinearities in ultra-high speed radio frequency (RF) and baseband circuits. These results have been published in [5, 6]. In comparison with the existing literature, the key conceptual novelty in our work is that we provide an analytical framework for mapping *system-level* performance goals to *hardware design* prescriptions for per-antenna nonlinearities. Thus, while prior work assesses the hardware design tradeoffs in particular scenarios [7, 8], we are able to provide a general framework which provides compact prescriptions that hardware designers can apply to design RF chains jointly with ADCs. Finally, unlike prior work on fading channels, we employ a LoS model which is a more suitable abstraction for mmWave channels [9, 10, 11, 12].

- We propose and investigate a *local* Linear Minimum Mean Square Error (LMMSE) receiver that exploits the sparsity of the mmWave wireless channel. Specifically, in mmWave uplink transmissions, most of a user's transmitted energy reaches the receiver through very few dominant paths. Hence, a spatial Discrete Fourier Transform (DFT) at the receiver array concentrates the energy of each user onto a few DFT bins in *beamspace*, as shown in Figure 1.2. By exploiting this property, one can greatly reduce the complexity of LMMSE detectors by using a small window in beamspace to demodulate each user. Our proposed approach provides 10-fold

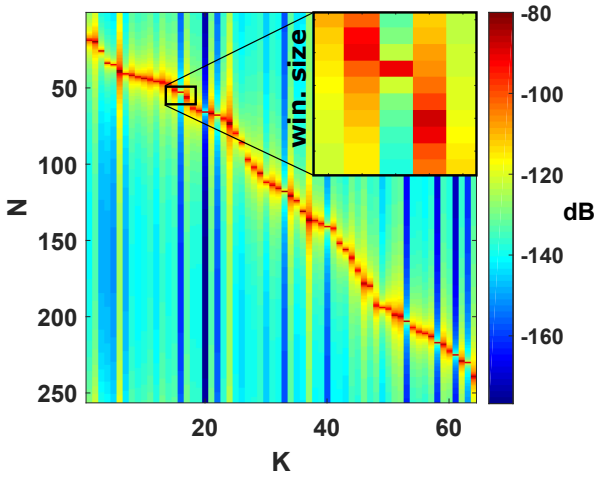


Figure 1.2: Sparsity of the mmWave LoS channel in the beamspace. The vertical axis represents the DFT bin index, while the horizontal axis represents the users indices.

complexity reduction when compared to conventional LMMSE detectors for a system with 256 antennas. This initial exposition of beamspace techniques has been published in [13].

- We extend and investigate the concept of complexity reduction via beamspace to three different scenarios:

    1. In a scenario where the number is of users is very high, compared to the number of base station antennas, while most interference is suppressed linearly by our proposed linear local LMMSE receiver, for each user, residual interference originating from a small number of strongly interfering users persists. In such cases, we propose to layer nonlinear interference cancellation on top of the local LMMSE receiver. The added nonlinear interference cancellation handles the residual interference after suitably whitening the local LMMSE output. This method provides reliable demodulation at higher load factor (defined as number of users divided by the number of antennas) than enabled by linear interference suppression alone, at order of magnitude lower complexity than

standard interference cancellation. These results have been published in [14].

2. We extend the linear LMMSE receiver idea to the dual downlink problem. Specifically, we propose a near-optimal linear precoding algorithm that exploits the sparsity of mmWave channels, employing a beamspace decomposition which limits the spatial channel seen by each user to a small window which does not scale with the number of antennas. This drastically reduces the complexity of computing the precoder, with complexity per iteration scaling linearly with the number of users, and makes it feasible to scale the system up to hundreds of antennas. These results have been published in [15].

3. In single-carrier mmWave systems, wideband operation results in multi-symbol delay spread across the receiver array, consequently, raising the issue of Inter-Symbol Interference (ISI). For such scenarios, we propose a novel technique that combines spatial domain FFT and time-domain FFT, together with an interpolation technique for limiting the spread of each user's spatial frequency across the band. These results have been published in [16]. As compared to the previous attempts on reducing the complexity of beamformer weight acquisition in such scenarios, we provide a framework for multiuser detection in single-carrier wideband mmWave systems, while previous papers have tackled the problem in single-user operation mode only [17, 18].

Chapter 2

A Design Framework For All-Digital mmWave Massive MIMO with Per-antenna Nonlinearities

In this chapter, we address the problem of desgining mmWave massive MIMO systems, taking into account the cost and power consumption of radio frequency (RF) frontends at high carrier frequencies and analog-to-digital conversion at large bandwidths. Due to the higher power consumption of high-speed RF chains, mmWave prototypes have thus far opted against fully digital arrays [19], and much of the recent research and development has focused on analog RF beamforming [20, 21, 22, 23, 24, 25], supporting a single user at a time, or hybrid beamforming [26, 27, 28, 29, 30, 31], where the number of supported users equals the number of RF chains, typically set to be much smaller than the number of antennas. However, recent advances in mmWave silicon hardware imply that scaling the number of RF chains with the number of antennas is on the cusp of feasibility, opening up the possibility of digital beamforming, where the number of supported users scales linearly with the number of antennas. By reducing dynamic range requirements

and increasing amplifier loading, we can boost the power efficiency of each RF chain enough to allow scaling to fully digital arrays with hundreds of elements, but at the cost of increasing nonlinearity in the RF chain. Similarly, drastically reduced precision can be used to control the cost and power consumption of analog-to-digital conversion, as well as that of communication and computation on the digital backend. Robust system design in the presence of such nonlinearities therefore plays a critical role in scaling digital beamforming to mmWave massive MIMO. Our goal in this chapter is to provide an analytical framework for quantifying the *system-level* impact of such nonlinearities, and to demonstrate how the increased degrees of freedom help relax linearity requirements and hardware specifications.

We consider, as a running example, a 256-element linear array at 140 GHz carrier frequency, with a symbol rate of 5 Gbaud with QPSK modulation for each supported user. For a load factor (defined as the ratio of the number of simultaneous users to the number of antennas) ranging from $\frac{1}{16}$ (16 users) to $\frac{1}{2}$ (128 users), the aggregate data rate ranges from 160 Gbps to 1.28 Tbps! However, scaling to this regime is challenging: wideband RF and baseband circuits scaled via relatively low-end silicon semiconductor processes (e.g., CMOS) exhibit significant nonlinearities, while the analog-to-digital converters (ADCs) available at multi-GHz sampling rates have relatively low precision. We provide in this chapter an analytical framework that enables designers to determine the permissible levels of such nonlinearities for their desired system-level performance guarantees. As we shall show, the RF linearity and ADC precision requirements for load factor $\frac{1}{2}$ are quite stringent, while reducing the load factor to $\frac{1}{16}$ results in significantly relaxed hardware specifications. For a given number of users to be supported, therefore, a massive MIMO architecture can be leveraged to overcome severe nonlinearities, by increasing the number of antenna elements in order to reduce the load factor. We note that, besides the enormous aggregate throughput from supporting multiple simultaneous

users, recent work also indicates that an all-digital solution can be more efficient in terms of hardware power consumption and area compared to a hybrid architecture [32].

Contributions:

We provide design guidelines based on linear MMSE reception, with an analytical framework based on two core concepts:

(a) We use a matched filter bound to show that the impact of per-antenna nonlinearities is effectively summarized by a quantity that we term the *intrinsic SNR,* corresponding to a normalized version of the nonlinearity. Key elements of this characterization are the well-known Bussgang decomposition, an overview of which can be found in [33], and the observation that, even for a moderate number of simultaneous users and without rich scattering, the antenna input is well modeled as zero-mean complex Gaussian. We show that the matched filter bound on the effective SNR for a given user, which captures the effect of the self-noise generated by per-antenna nonlinearities, depends only on four parameters: the user's SNR, the intrinsic SNR, the load factor and a power control factor which summarizes the variations in received signal power across users.

(b) We show that a pessimistic estimate of the degradation in performance due to multiuser interference can be obtained by analyzing (theoretically and/or numerically) an *ideal* system without nonlinearities. This enables us to provide a lower bound on the output signal-to-interference-plus-noise ratio (SINR) of a linear MMSE receiver, accounting for both nonlinearities and multiuser interference.

Combining these two concepts, averaging over the spatial distribution of users, and specializing to an edge user in the cell, allows us to provide analytical guidelines for maximum permissible levels of nonlinearities in order to provide the desired system-level performance guarantees (e.g., on outage probabilities). Consequently, our analysis utilizes the analytical capabilities of the Bussgang decomposition and LMMSE framework to provide a cross-layer design tool that links hardware level specifications to the system level

performance metrics of a multi-user massive MIMO system. This enables exploration of fundamental tradeoffs between power consumption and cost of RF frontends and system performance. We consider third order RF and baseband nonlinearities that can be specified using the so-called 1 dB compression point [34], termed P_{1dB}. The per-antenna ADCs for the in-phase and quadrature components are modeled as overloaded uniform quantizers optimized (for a specified number of bits) to minimize the mean square error with zero mean Gaussian input. Using our framework, we are able to provide compact design prescriptions for P_{1dB} and the number of ADC bits. For example, for a load factor of 1/2, the system can work with 4-bit ADC and passband/baseband P_{1dB} of 8.4 dB / 5 dB. On the other hand, 2-bit ADC with passband/baseband P_{1dB} of 1.4 dB / -1 dB suffice to work properly with a load factor of 1/16. We present extensive simulations verifying our analytical predictions and prescriptions.

2.1 Related Work

While the focus in the present chapter is on mmWave massive MIMO, there is a significant body of closely related recent research on the effect of nonlinearities on multiuser massive MIMO at lower carrier frequencies. Most of this prior work also employs Bussgang's theorem [35] to model the effect of nonlinearities, both for uplink reception and downlink precoding. Our discussion here is limited to the literature on uplink massive MIMO, since that is the focus of the present chapter, but the design framework for modeling downlink nonlinearities such as power amplifiers and digital-to-analog converters (DACs) is well known to be entirely analogous.

The line of sight (LoS) channel model used in our performance evaluation is different from that in much of this prior work, which employs models that are better matched to the propagation environments at lower frequencies. However, our analytical framework is

quite general, and can be used to obtain design prescriptions for lower carrier frequencies as well. Conversely, many of the general observations emerging from prior work at lower carrier frequencies are consistent with the conclusions in the present chapter, given a common underlying mathematical framework that employs the Bussgang decomposition and exploits the relaxation of hardware constraints enabled by the increase in the number of antennas. In the following, we briefly review this prior work in order to place the contributions of the present chapter in perspective.

The potential for relaxing hardware constraints by increasing the number of antennas is clearly brought out by the theoretical results in [36], which show that the performance degradation due to hardware impairments vanishes asymptotically as the number of base station antennas gets large. The same trend holds for a finite but large number of antennas, as is clear from the results in [37, 38, 39], which study the spectral efficiency of quantized massive MIMO over frequency nonselective Rayleigh and Rician fading channels using maximum ratio combining. Another interesting conclusion from the simulations of [38] is that, for Rician fading, the system is more vulnerable to drastic quantization as the relative strength of the dominant component increases. Thus, the LoS model considered in this chapter may be a worst-case scenario for obtaining design prescriptions regarding nonlinearities.

The impact of imperfect power control for quantized massive MIMO over frequency nonselective channels is included in the analysis in [40, 41]. Using spectral efficiency as a performance measure, an example conclusion from [40] is that 3-bit ADC suffices for a system with 100 antennas serving 10 users at a spectral efficiency of 3.5 bits per channel use, with 4-bit ADCs recommended to handle imperfections in power control and automatic gain control. Similar conclusions are obtained in [41], which shows moderate drops in spectral efficiency due to imperfect power control.

The impact of quantization on multiuser OFDM MIMO over a frequency-selective

channel is studied in [7], with a focus on low-complexity channel estimation and data detection. The simulations in this chapter show that, for the models considered, 4-bit ADC is sufficient to achieve a near-optimal performance (in terms of packet error rate) for a load factor of 1/8 or lower. More recent work with a similar model [8] employs a Bussgang-based analysis for the joint distortion introduced by nonlinear low-noise amplifiers, phase noise, and finite-resolution ADCs, and demonstrates its accuracy by comparing analytical predictions with simulations.

In comparison with the existing literature, the key conceptual novelty in the present chapter is that we provide an analytical framework for mapping *system-level* performance goals to *hardware design* prescriptions for per-antenna nonlinearities. The theoretical foundation for this mapping is our observation (Theorem V.2) that an ideal system without nonlinearities provides a means of obtaining pessimistic performance estimates, together with our abstraction of self-noise via intrinsic SNR and the associated matched filter bound (Theorem V.1). Thus, while prior work such as [7, 8] demonstrates the accuracy of Bussgang modeling and assesses design tradeoffs in particular scenarios, we are able to provide a general framework which provides *compact* prescriptions that hardware designers can apply to design RF chains jointly with ADCs, by considering the cascade of passband amplifiers, baseband amplifiers and ADCs as the nonlinearities employed in our performance evaluation. Finally, unlike prior work on fading channels, we employ a LoS model which is a more suitable abstraction for mmWave channels [9, 10, 11, 12].

2.2 System Model

Fig. 2.1 shows the system model. The base station performs horizontal scanning with a 1D half-wavelength spaced N-element array. Let K denotes the number of simultaneous users, and $\beta = \frac{K}{N}$ the *load factor.*

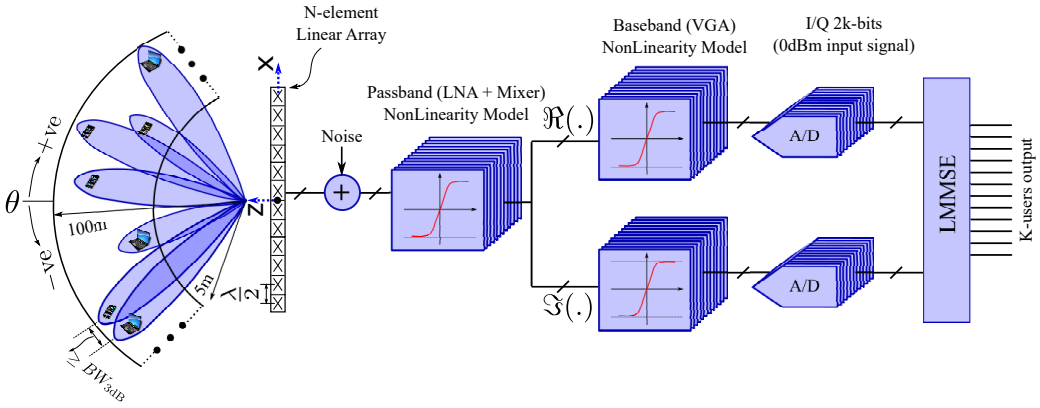


Figure 2.1: The cell size is constrained between $R_{\min}$ and $R_{\max}$ in link range and between $-\pi/3 \leq \theta \leq \pi/3$ in angle. $\mathrm{BW}_{3\mathrm{dB}}$ stands for the 3 dB beamwidth. The passband and baseband nonlinearities are modeled by saturated third order polynomials. An overloaded uniform ADC with b bits per dimension, optimized for a zero-mean standard Gaussian random variable, is used. Linear MMSE reception is employed after digitization.

We assume a line-of-sight (LoS) channel between the base station and each mobile. The direction of arrival (DoA) from the k^{th} mobile is denoted by θ_k, and corresponds to spatial frequency $\Omega_k = 2\pi \frac{d_x}{\lambda} \sin\theta_k$, where λ denotes the carrier wavelength and d_x denotes the inter-element spacing, set to $\frac{\lambda}{2}$ in our numerical results. The $N \times 1$ spatial channel for mobile k is given by

$$\mathbf{h}_k = A_k e^{j\phi_k} \left[1\, e^{j\Omega_k}\, e^{j2\Omega_k}\, \ldots\, e^{j(N-1)\Omega_k}\right]^{\intercal}, \tag{2.1}$$

where ϕ_k is an arbitrary phase shift and $A_k^2 = \left(\frac{\lambda}{4\pi R_k}\right)^2$ depends on the radial location R_k of mobile k, using the Friis formula for path loss.

The cascade of the nonlinearities described in Sections 2.2.2 and 2.2.3 is modeled as a complex baseband equivalent nonlinearity $g(\cdot)$. The complex baseband received signal

vector $\mathbf{z}$ at the base station is therefore given by

$$\mathbf{z} = g(\mathbf{y}) = g\left(\underbrace{\mathbf{H}}_{N\times K}\mathbf{x} + \mathbf{n}\right), \tag{2.2}$$

where $\mathbf{H} = [\mathbf{h}_1\mathbf{h}_2 \dots \mathbf{h}_K]$ is the channel matrix, $\mathbf{x} = [x_1, ..., x_K]^T$ is the vector of symbols (normalized to unit energy: $\mathbb{E}\left[|x_k|^2\right] = 1$) transmitted by the mobiles, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I})$ is the thermal AWGN vector, and $g(\cdot)$ is the effective per-antenna nonlinearity in complex baseband.

Running example: As mentioned in the introduction, for our running example, we set $N = 256$, with load factor β ranging from $\frac{1}{16}$ to $\frac{1}{2}$ (i.e., K ranging from 16 to 128). We provide the link budget analysis for the envisioned system in Appendix A to highlight the feasibility of low-cost silicon hardware realizations.

In the remainder of this section, we characterize the statistics of the received signal at each antenna and describe the nonlinearity models included in our numerical results.

2.2.1 Per-antenna Received Signal Statistics

The input to the effective complex baseband nonlinearity $g(\cdot)$ at, say, antenna m, is given by

$$y_m = \sum_{k=1}^{K} A_k e^{j\phi_k} x_k e^{jm\Omega_k}. \tag{2.3}$$

For a uniform spatial distribution of users over the region of interest, the amplitudes $\{A_k\}$ and spatial frequencies $\{\Omega_k\}$ are independent and identically distributed (i.i.d.). The phases $\{\phi_k\}$ are uniform over $[0, 2\pi]$, and x_k are i.i.d. QPSK symbols. By virtue of the central limit theorem (CLT), the received signal is well modeled as zero-mean complex Gaussian for large K, and jointly Gaussian across antennas. We have verified empirically, via histogram comparisons, quantile-quantile plots and KL divergence computations, that

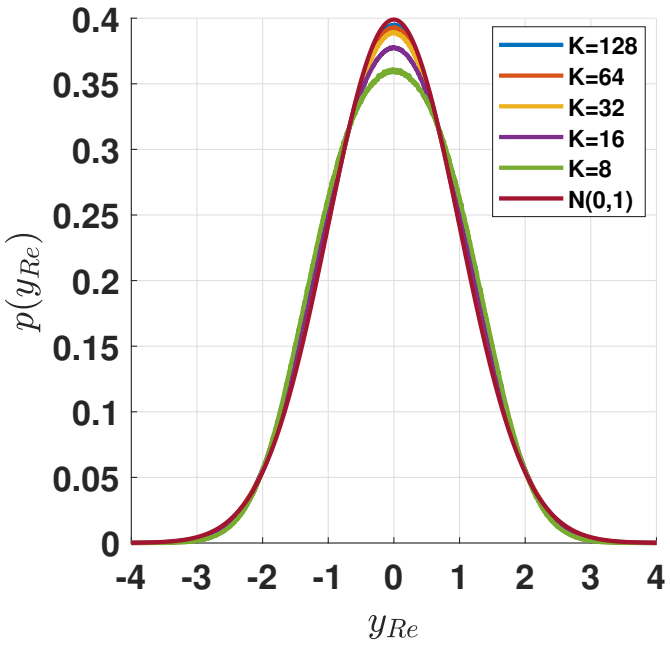


Figure 2.2: The pdf of the standard normal distribution and the histogram of the normalized real/imaginary part of the received signal at each antenna element when K users transmit.

this Gaussian approximation holds for even moderate number of mobiles (e.g., $K = 8$) in all settings that we have considered. Fig. 2.2 (a) illustrates a comparison between the histogram of the normalized real/imaginary component of the received signal and the standard normal distribution $\mathcal{N}(0, 1)$.

In terms of technical conditions for applying the CLT, we note that it holds for independent, non-identically distributed random variables if the variances are bounded [42], which is the case here: with power control, the amplitudes of $\{A_k\}$ are tightly clustered, whereas with no power control they lie within a range of values determined by the maximum and minimum link distance. The randomness in the terms of the sum in (2.3) results from randomness in channel phases and data modulations, and CLT is applied conditioned on $\{A_k\}$. We average over the realizations of $\{A_k\}$ to determine expected receiver performance and reliability bounds.

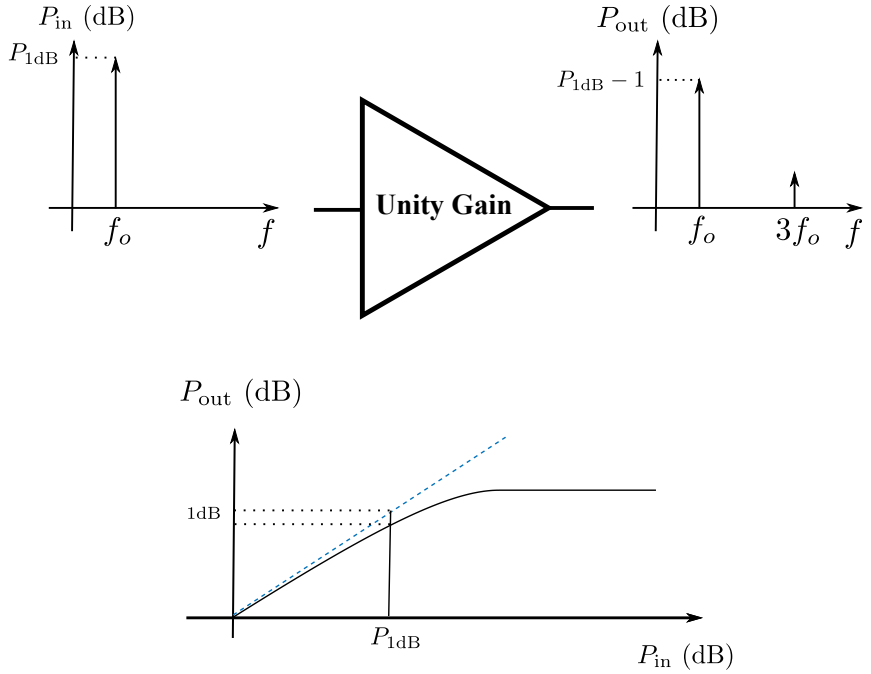


Figure 2.3: The 1 dB compression point (P_{1dB}) is defined as the input power at which the output power of the desired sinusoid (at f_o) is compressed by 1 dB.

2.2.2 Passband and Baseband Nonlinearity Model

The passband nonlinearity arises in the low noise amplifier and the mixer, while the baseband nonlinearity is in the variable gain amplifier. We model each nonlinearity as a saturated third-order polynomial function with a nominal gain of unity. The function is parametrized by the 1 dB compression point (P_{1dB}) [34], defined as the input power of a sinusoid of frequency f_o (taken to be the carrier frequency) at which the output power is reduced by 1 dB relative to the nominal. The concept is illustrated in Fig. 2.3. A commonly used model for gain saturation using third-order nonlinearity is the cubic soft clipper which can be parametrized by P_{1dB} as follows:

$$g(y(t)) = \begin{cases} y(t)(1 - \frac{0.44|y(t)|^2}{3P_{\text{1dB}}}) & \text{if } |y(t)|^2 \leq \frac{P_{\text{1dB}}}{0.44} \\ \frac{y(t)}{|y(t)|}\sqrt{P_{\text{1dB}}} & \text{if } |y(t)|^2 > \frac{P_{\text{1dB}}}{0.44} \end{cases}. \tag{2.4}$$

The gain compression for the passband nonlinearity depends on the absolute value of the complex baseband signal, while the gain compression depends on the absolute value of the I and Q components for the baseband nonlinearity. Fig. 2.4 (a) illustrates the distribution of the input powers of the passband and baseband nonlinearities, along with example input/output (I/O) characteristics. In this work, we consider the nonlinearities to be memoryless and free of phase distortion.

2.2.3 ADC Model

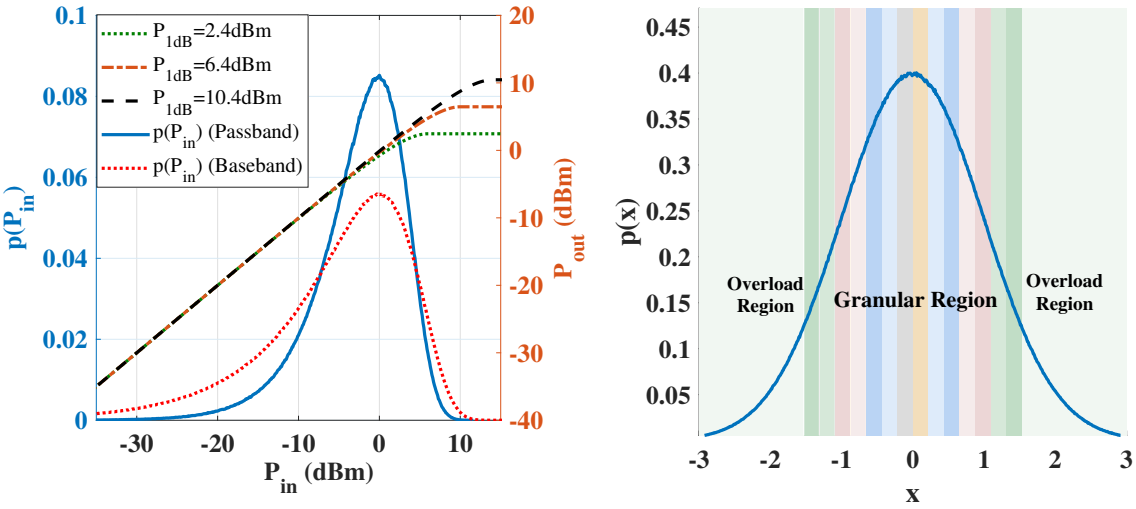
After down-conversion, each baseband component is quantized to b bits by an ADC. We design the quantizer to minimize the mean squared error (MSE) assuming that the incoming signal is Gaussian with zero mean and unit variance. An automatic gain control (AGC) precedes the ADC in order to normalize the average power of the input signal to unity, and ensure that the full dynamic range of the ADC is utilized. We employ an overloaded uniform ADC [43]: while the MSE could be improved slightly by designing a non-uniform quantizer for standard Gaussian input, the improvement is slight and has no discernible impact on system-level performance (see Appendix B for a quantitative discussion). Fig. 2.4 (b) depicts a 4-bit uniform overloaded quantizer.

2.2.4 Linear MMSE Detector

We show in a following section that the impact of a per-antenna nonlinearity $g(\cdot)$ can be modeled as additional noise, leading to an equivalent system model of the form

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \tilde{\mathbf{n}}, \tag{2.5}$$

where $\tilde{\mathbf{n}}$ is zero mean with variance $(\sigma_n^2 + \nu_g^2)\mathbf{I}$. The value of ν_g is specified in section 2.5. Thus, any adaptive implementation of the linear MMSE receiver automatically accounts



(a) Third-order nonlinearities (b) Overloaded uniform ADC

Figure 2.4: (a) Third-order nonlinearities characterized by P_{1dB}, and probability distribution function of instantaneous input power, $\text{p}(\text{P}_{\text{in}})$, for passband and baseband signals. (b) Histogram of I and Q baseband components along with ADC quantization bins.

for the nonlinearities. The linear MMSE receiver is specified as follows:

$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}, \tag{2.6}$$

where

$$\mathbf{W} = \left(\mathbf{H}^H\mathbf{H} + (\sigma_n^2 + \nu_g^2)\mathbf{I}\right)^{-1}\mathbf{H}^H. \tag{2.7}$$

The linear MMSE detector has a rich history with well-known properties [44, 45]. In order to provide a self-contained exposition, we state a few properties that are relevant for our present purpose, and sketch their proof in Appendix C.

2.3 Bussgang Linearization

In order to provide a self-contained exposition, we review Bussgang linearization in the context of our MIMO system.

2.3.1 Scalar Bussgang Linearization

For a zero mean complex-valued random variable y and a nonlinearity $g(\cdot)$, a linear MMSE approximation of $g(y)$ by ay satisfies the orthogonality principle [46]:

$$\mathbb{E}((g(y) - ay)y^*) = 0. \tag{2.8}$$

Standard computations for the linear gain a and the variance of the approximation error $e = g(y) - ay$ yield

$$a = \frac{\mathbb{E}(g(y)y^*)}{\mathbb{E}(|y|^2)}, \tag{2.9}$$

$$\sigma_g^2 = \mathbb{E}(|e|^2) = \mathbb{E}(|g(y)|^2) - |a|^2\mathbb{E}(|y|^2). \tag{2.10}$$

Hence, $g(y)$ can be written as

$$g(y) = ay + e, \tag{2.11}$$

where a and $\mathbb{E}(|e|^2) = \sigma_g^2$ can be computed analytically or empirically for any distribution of y and nonlinear function $g(\cdot)$. Bussgang evaluated a and σ_g^2 for different nonlinear functions when the input y is Gaussian random variable [35]. In our analysis, we consider the function $g(\cdot)$ described in (2.4) for the overall RF chain nonlinearity.

2.3.2 Vector Bussgang Linearization

The main part of Bussgang's theorem in [35], and its extension to the complex domain in [47], is the preservation of covariance structure under nonlinearities for jointly Gaussian random variables:
If y and z are jointly Gaussian random variables and $g(\cdot)$ is a nonlinear function, then $\mathbb{E}(g(y)z^*) = a\mathbb{E}(yz^*)$, where a is defined in (2.9).

This result allows us to characterize the linear MMSE fit for a Gaussian random vector in terms of the scalar linear MMSE fits for its components. It has been customized to MIMO in many recent papers [48, 8, 41, 49], hence we state the relevant result here without proof (see Appendix A in [49] for a derivation).

Theorem 2.3.1 Vector Bussgang Decomposition
Let $\mathbf{y}$ *denote the jointly Gaussian random vector input to the effective nonlinearity* $g(\cdot)$ *referred to complex baseband, so that the received signal* $\mathbf{z} = g(\mathbf{y})$. *Then the Bussgang decomposition of* $\mathbf{z}$ *is given by*

$$\mathbf{z} = \mathbf{g}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{e}, \tag{2.12}$$

where

$$\mathbf{A} = \mathcal{D}iag([a_1, \ldots, a_N]), \tag{2.13}$$

$$a_i = \frac{\mathbb{E}(g(y_i)y_i^*)}{\mathbb{E}(|y_i|^2)}, \tag{2.14}$$

and the variance of element e_i *of the approximation error vector* $\mathbf{e}$ *is given by*

$$\sigma_{gi}^2 = \mathbb{E}(|g(y_i)|^2) - |a_i|^2\mathbb{E}(|y_i|^2). \tag{2.15}$$

The Bussgang theorem on covariance preservation therefore leads to a linear MMSE fit with diagonal covariance structure. Moreover, the diagonal elements are equal if the statistics of $\{y_i\}$ are identical, as in the following straightforward corollary, stated without proof.

Corollary 1 *If the diagonal elements of the covariance of* $\mathbf{y}$ *are equal, i.e.,* $\mathbb{E}(|y_i|^2) = \mathbb{E}(|y_k|^2)$, $\forall\, i, k$, *then the Bussgang decomposition specializes to*

$$\mathbf{z} = \mathbf{g}(\mathbf{y}) = a\mathbf{y} + \mathbf{e}, \tag{2.16}$$

where a *and* $\mathbb{E}(|e_i|^2) = \sigma_g^2$ *are the scalar Bussgang parameters of* $g(\cdot)$.

It is worth noting that the self-noise $\mathbf{e}$ may be spatially correlated. However, recent work [48] indicates that this correlation becomes negligible when the number of users is large, and we ignore it in our analysis here.

2.4 Bussgang Normalization and Intrinsic SNR

In this section, we define a normalization such that the Bussgang parameters for a nonlinearity are independent of input power. We introduce the concept of *intrinsic SNR* to characterize the self-noise in this normalized setting. As we shall see, this is the summary specification that is provided by system-level design requirements to the hardware designer, based on the analytical framework described in the next section. Finally, we show, via the simple example of a limiter, how such a summary can be used to determine hardware specifications for a nonlinearity.

Normalized Nonlinearity

As shown in Fig. 2.5 (a) and (b), Bussgang decomposition characterizes a nonlinear function $g(\cdot)$ by parameters a and σ_g^2. These parameters depend on the input power by definition as shown in Eq. (2.9) and (2.10).

Fig. 2.5 (c) illustrates a normalized version of the nonlinearity in Fig. 2.5 (a): the input power is scaled to one before the nonlinearity, and the scaling is undone after the nonlinearity. The Bussgang linearization of the normalized nonlinearity, with parameters $\tilde{a}$ and $\tilde{\sigma}_g^2$, is depicted in Fig. 2.5 (d). The parameters $\tilde{a}$ and $\tilde{\sigma}_g^2$ represent the Bussgang decomposition of the *normalized* nonlinear function $\tilde{g}(\cdot)$, depicted in Fig. 2.5 (c). The equivalence of the nonlinear models (a) and (c) implies that the corresponding linear models (b) and (d) must satisfy $\tilde{a} = a$ and $\tilde{\sigma}_g^2 = \sigma_g^2/\mathbb{E}(|y|^2)$.

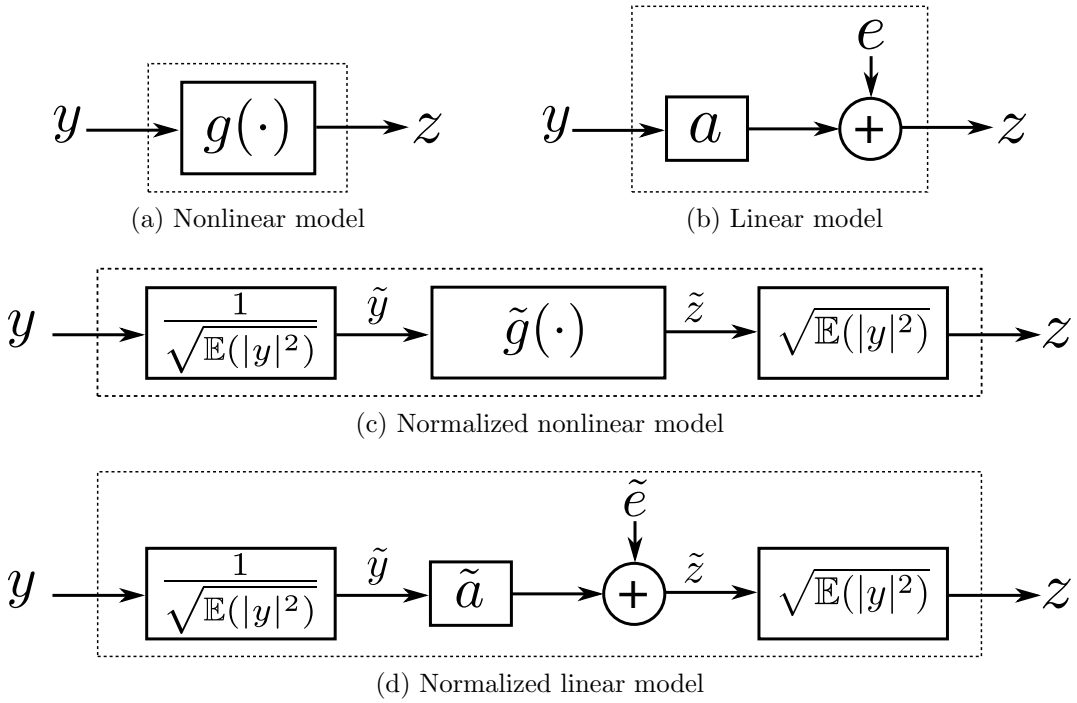
It is convenient to define hardware specifications for the normalized nonlinearity; in hardware design parlance, the specifications are "referred to the input power." We summarize these using the concept of *intrinsic SNR,* which plays a key role in our analytical framework.

Definition 2.4.1 Intrinsic SNR

We define the "intrinsic SNR" of a nonlinearity $g(\cdot)$ using the Bussgang parameters of its normalized version $\tilde{g}(\cdot)$ as follows:

$$\gamma_g = \frac{|\tilde{a}|^2}{\tilde{\sigma}_g^2}. \tag{2.17}$$

As a simple example, consider a memoryless limiter as depicted in Fig. 2.6 (a), which is specified by the gain G and the power threshold P_{th} at which the output signal is clipped. The normalized version of this function has unity gain, as shown in Fig. 2.6 (b), hence we only need to specify a single parameter to characterize it: the clipping threshold

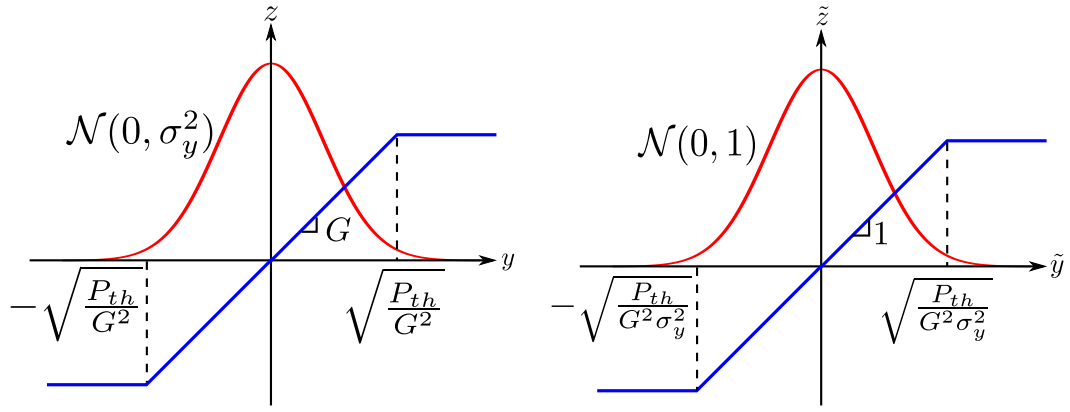


(a) Nonlinear model (b) Linear model

(c) Normalized nonlinear model

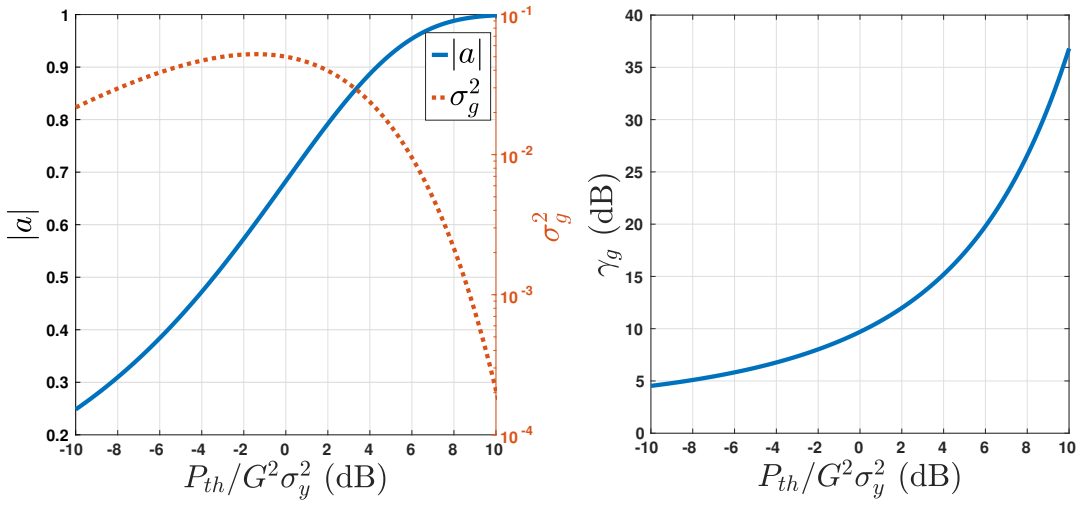
(d) Normalized linear model

Figure 2.5: The nonlinear function $g(\cdot)$ in (a) can be decomposed to the linear model in (b) whose parameters depend on the input power. We define a normalized version of the nonlinearity in (c), which allows us to provide design specifications independent of input power. The corresponding normalized linearization is depicted in (d).



(a) Limiter function (b) Normalized limiter function

Figure 2.6: (a) The conventional limiter function. (b) a unity-gain limiter function whose clipping threshold is normalized to the effective input power.



(a) Bussgang parameters (b) Intrinsic SNR

Figure 2.7: (a) Bussgang parameters and (b) the intrinsic SNR of the normalized limiter function.

$\tilde{P}_{th} = P_{th}/G^2\sigma_y^2$ normalized to the input power σ_y^2. The Bussgang parameters of the normalized limiter function are shown in Fig. 2.7 (a), and the intrinsic SNR is shown in Fig. 2.7 (b).

Henceforth, nonlinearities and their Bussgang parameters are normalized to the input power, and we drop the "tilde" notation to denote the normalized version. For example, the 1 dB compression point of a passband/baseband nonlinearity is normalized to the input power, and hence is measured in dB instead of dBm.

Design Approach

The analytical framework described in the next section leads the following design approach for going from system-level performance metrics to hardware design specifications:

• MIMO performance specifications lead to a requirement for the intrinsic SNR for the per-antenna nonlinearities, ignoring the specific nature of the nonlinearities. For exam-

ple, suppose that we require an intrinsic SNR of 20 dB at least 95% of the time.

• We map the intrinsic SNR requirement to a specification for the normalized nonlinearity. Taking the limiter in Fig. 2.6 as an example, we see from Fig. 2.7 (b), the clipping threshold normalized to the effective input power, $P_{th}/G^2\sigma_y^2$, must be at least 6 dB in order to attain an intrinsic SNR of 20 dB.

• In this step, the absolute value of the gain and clipping threshold is calculated. For example, suppose that the system in our running example is at load factor $\beta = 1/4$, i.e., 64 users. Then, according to the link budget presented in Appendix A, the input power to the receive chain is -60 dBm if power control is employed. We therefore obtain that $P_{th}/G^2 = -54$ dBm. The hardware designer now has to choose G and P_{th} in order to achieve this ratio or better.

2.5 Analytical Framework

Our analytical framework is developed as follows.

1. We derive a matched filter bound for each user in the MIMO system that accounts for the self-noise due to the per-antenna nonlinearities (which scales with the power summed across users) as well as thermal noise. To this end, we use Bussgang linearization and the intrinsic SNR discussed in the previous section.

2. We derive a lower bound for the output SINR of the LMMSE receiver for any given user. Defining the *efficiency* of the LMMSE receiver for a given user as the ratio of SINR to SNR, we show that the efficiency of a user in an ideal system without nonlinearities is a *lower bound* on that of the actual system. This, together with the matched filter bound, provides a lower bound on the LMMSE output SINR.

3. We obtain system-level design prescriptions by specializing the preceding lower

bound to an "edge" user whose performance is stochastically poorer than that of any other user.

2.5.1 Bussgang Linearized Model

As described in section 2.2, we denote by $\{A_k, k = 1, \ldots, K\}$ the amplitudes of the incoming waves for the K users, and by σ_n^2 the variance of the thermal noise at each antenna. We can therefore model the incoming signal at each receive antenna as $y_m \sim \mathcal{CN}(0, \sigma_y^2)$, where

$$\sigma_y^2 = \sum_{k=1}^{K} A_k^2 + \sigma_n^2 = \sigma_n^2 + K A_{\text{rms}}^2, \tag{2.18}$$

and

$$A_{\text{rms}} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} A_k^2} \tag{2.19}$$

is the root mean square (rms) amplitude, averaged across users.

As depicted in Fig. 2.5 (c), using the normalized Bussgang linearization requires scaling the incoming signal to unit variance as follows:

$$\tilde{y}_m = \frac{y_m}{\sigma_y}. \tag{2.20}$$

For a normalized nonlinearity $g(\cdot)$ as defined in the previous section, our per antenna linearized model is given by:

$$g(\tilde{y}_m) = a\tilde{y}_m + e_m. \tag{2.21}$$

For the received signal (2.2), the normalized signal prior to passing through the nonlinearity is given by

$$\tilde{\mathbf{y}} = \frac{\mathbf{y}}{\sigma_y}. \tag{2.22}$$

Using the Bussgang decomposition, we have

$$g(\tilde{\mathbf{y}}) = a\tilde{\mathbf{y}} + \mathbf{e} = \frac{a}{\sigma_y}\mathbf{y} + \mathbf{e}, \tag{2.23}$$

where $\mathbb{E}[\mathbf{e}\mathbf{e}^H] = \sigma_g^2\mathbf{I}$. We can now go back to the original signal scaling to obtain

$$\hat{\mathbf{y}} = \frac{\sigma_y}{a}g(\tilde{\mathbf{y}}) = \mathbf{y} + \frac{\sigma_y}{a}\mathbf{e} = \mathbf{H}\mathbf{x} + \mathbf{n} + \frac{\sigma_y}{a}\mathbf{e}. \tag{2.24}$$

This is the model (2.5), with effective noise

$$\tilde{\mathbf{n}} = \mathbf{n} + \frac{\sigma_y}{a}\mathbf{e} \tag{2.25}$$

of variance

$$\mathbb{E}(\tilde{\mathbf{n}}\tilde{\mathbf{n}}^H) = (\sigma_n^2 + \nu_g^2)\mathbf{I} \tag{2.26}$$

where

$$\nu_g^2 = \frac{\sigma_y^2}{|a|^2}\sigma_g^2 = \frac{\sigma_y^2}{\gamma_g}. \tag{2.27}$$

2.5.2 Matched Filter Bound

For the k^{th} user, the matched filter bound for the linearized model (2.5), with equivalent noise as in (2.24)-(2.25), is simply given by

$$\text{SNR}_k(g) = \frac{||\mathbf{h}_k||^2}{\sigma_n^2 + \nu_g^2}. \tag{2.28}$$

Our design framework is built around the dependence of this bound on key system parameters as stated in the following theorem. We first ignore thermal noise, in order to

clearly brings out the role of intrinsic SNR, γ_g, and load factor, β, and then include its effect.

Theorem 2.5.1 Matched filter bound

(a) **Self-noise only:** *Ignoring thermal noise, the matched filter bound for user k is given by*

$$SNR_k(g) = \gamma_g \frac{A_k^2}{\beta A_{rms}^2}. \tag{2.29}$$

(b) **Self-noise and thermal noise:** *The matched filter bound for user k, considering both self-noise and thermal noise, is given by*

$$SNR_k(g, \sigma_n^2) = \frac{1}{\frac{1}{SNR_k(g)} + \frac{1+\gamma_g}{\gamma_g}\frac{1}{SNR_k}}, \tag{2.30}$$

where $SNR_k = NA_k^2/\sigma_n^2$ is the SNR for user k accounting for thermal noise alone.

Proof: The proof involves algebraic manipulations based on the linearized model (2.24)-(2.25).

(a) Using (2.1), the numerator in (2.28) is given by

$$||\mathbf{h}_k||^2 = NA_k^2. \tag{2.31}$$

Using (2.18) and (2.27), and setting $\sigma_n^2 = 0$, the denominator in (2.28) is given by

$$\nu_g^2 = \frac{KA_{\text{rms}}^2}{\gamma_g}. \tag{2.32}$$

Plugging (2.31) and (2.32) into (2.28), we obtain

$$\mathrm{SNR}_k(g) = \frac{NA_k^2\gamma_g}{KA_{\mathrm{rms}}^2} = \frac{\gamma_g A_k^2}{\beta A_{\mathrm{rms}}^2}, \tag{2.33}$$

which is the desired result (2.29).

(b) From (2.28) and (2.31), we have

$$\frac{1}{\mathrm{SNR}_k(g,\sigma_n^2)} = \frac{\sigma_n^2}{NA_k^2} + \frac{\nu_g^2}{NA_k^2}. \tag{2.34}$$

For non-zero thermal noise, we have, using (2.18) and (2.27), that

$$\nu_g^2 = \frac{KA_{\mathrm{rms}}^2 + \sigma_n^2}{\gamma_g}. \tag{2.35}$$

Plugging into (2.34), we obtain upon simplification the desired result (2.30). ■

Note that, if $\gamma_g \gg 1$, then the formula (2.30) reduces to

$$\mathrm{SNR}_k(g,\sigma_n^2) = \frac{1}{\frac{1}{\mathrm{SNR}_k(g)} + \frac{1}{\mathrm{SNR}_k}}. \tag{2.36}$$

In order to provide system-level performance guarantees, we focus on supporting users at the cell edge. We therefore now set A_k to the worst-case amplitude A_{edge} (at 100 m range for our running example), while computing A_{rms} by a statistical average $\sqrt{\mathbb{E}[A^2]}$ given the users distribution, assuming a large enough number of users. Let us term the ratio of the power of the edge user to the rms power as the *power control factor,* since it depends on the power control scheme used. The power control factor α_p is given by

$$\alpha_p = \frac{A_{\mathrm{edge}}^2}{A_{\mathrm{rms}}^2}. \tag{2.37}$$

Specializing (2.29) to the edge user, we now obtain that

$$\mathrm{SNR}_{\mathrm{edge}}(g) = \gamma_g \, \frac{1}{\beta} \, \alpha_p. \tag{2.38}$$

Power control factor with no power control: For users who are uniformly distributed over the area bounded by $[R_{\min}, R_{\max}]$ and a given angular range, we obtain upon straightforward computation that, for a system without power control,

$$\begin{aligned} \alpha_p =& \frac{\frac{1}{R_{\max}^2}}{\frac{1}{R_{\max}^2 - R_{\min}^2} \int_{R_{\min}^2}^{R_{\max}^2} \frac{1}{r} dr}, \\ =& \frac{1 - \frac{R_{\min}^2}{R_{\max}^2}}{2 \log \frac{R_{\max}}{R_{\min}}}. \end{aligned} \tag{2.39}$$

which evaluates to -7.8 dB for $R_{\max} = 100$ m, $R_{\min} = 5$ m.

2.5.3 Lower Bound on LMMSE Output SINR

We now provide a lower bound on the output SINR of any user via the ideal system.

Theorem 2.5.2 LMMSE Lower Bound

In the presence of nonlinearity, a lower bound on the output SINR of a linear MMSE receiver for any user is given as

$$SINR \geq SNR(g, \sigma_n^2) \; \eta_{ideal}, \tag{2.40}$$

where

$$\eta_{ideal} = \frac{SINR(ideal)}{SNR(ideal)}. \tag{2.41}$$

is the efficiency in an ideal system with the same user configuration and amplitudes, but without nonlinearity.

Proof: Since the system described in (2.5) is a pessimistic model for the system in (2.2), we have by Lemma C.1 in Appendix C that

$$\frac{\text{SINR}}{\text{SNR}(g, \sigma_n^2)} \geq \frac{\text{SINR(ideal)}}{\text{SNR(ideal)}}, \tag{2.42}$$

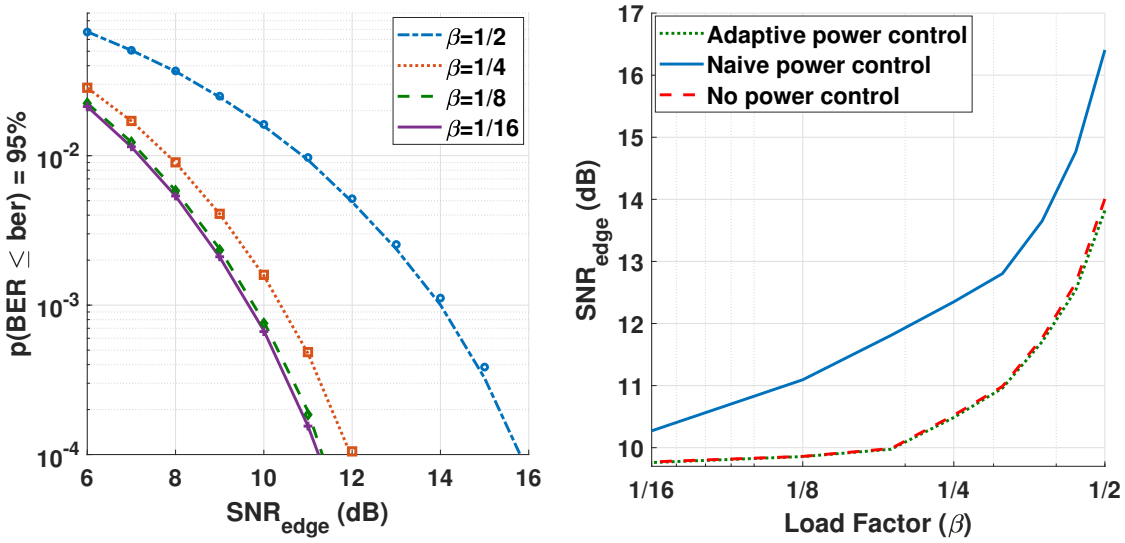
where SINR(ideal) is the target linear MMSE output SINR for a user in an ideal system (without nonlinearities). ■

We evaluate η_{ideal} through simulations of the ideal system for the edge user, as shown in Fig. 2.8, where $\text{SNR}_{\text{edge}} = \frac{NA_{\text{edge}}^2}{\sigma_n^2}$, and A_{edge} is the received amplitude of the user at 100 m. The target output SINR of the linear MMSE, i.e., $\text{SINR}_{\text{edge}}$(ideal) is 9.7 dB. This number corresponds to the SNR_{edge} in a single user case. Hence, in a single user case $\eta_{\text{ideal}} = 1$. As the load factor increases, there is noise enhancement due to interference suppression: $\eta_{\text{ideal}} = 9.7 - \text{SNR}_{\text{edge}}|_{\text{dB}}$ can be inferred from Fig. 2.8 (b).

2.5.4 From System-Level Performance to Intrinsic SNR

The chosen quality of service measure maps to an SINR requirement at the LMMSE output. We compute this for the ideal system. For example, simulating the ideal system, a target BER of 10^{-3} with 95% availability is obtained for $\text{SINR}_{\text{edge}}$(ideal) = 9.7 dB. Since the SNR for an edge user is 14 dB, we see from Fig. 2.8 (b) that the efficiency for the ideal system is given by $9.7 - 14 = -4.3$ dB for no power control and $\beta = 1/2$. This is an upper bound on the efficiency of the actual system.

We can now compute the minimum $\text{SNR}_{\text{edge}}(g, \sigma_n^2)$ from Eq. (2.40) to achieve the required SINR in the presence of nonlinearities. Finally, we can infer the intrinsic SNR



(a) BER in ideal system

(b) SNR_{edge} required in ideal system

Figure 2.8: (a) BER for 5% outage in an ideal system (no nonlinearities) for different load factors. (b) SNR for an edge user (100 m from base station) to guarantee that 95% of the mobiles have raw BER of 10^{-3} for different load factors.

γ_g required from Eq. (2.30) and Eq. (2.36). This is now mapped to detailed hardware specifications, as illustrated by examples in the next section.

2.6 Design Examples and Performance Evaluation

The system parameters are as described in Section 2.2. We illustrate our design for a target uncoded BER of 10^{-3}, which is low enough for reliable performance using a high-rate channel code with relatively low decoding complexity. For QPSK, the corresponding required SNR over a SISO AWGN link is 9.7 dB. This becomes our target SINR at the output of the LMMSE receiver for an edge user. This setting is simply for illustration: our analytical framework applies for any QoS measure that can be approximated in terms of SINR (e.g., outage capacity or spectral efficiency using Shannon's formula).

In the following, we first describe the user distribution and power control schemes

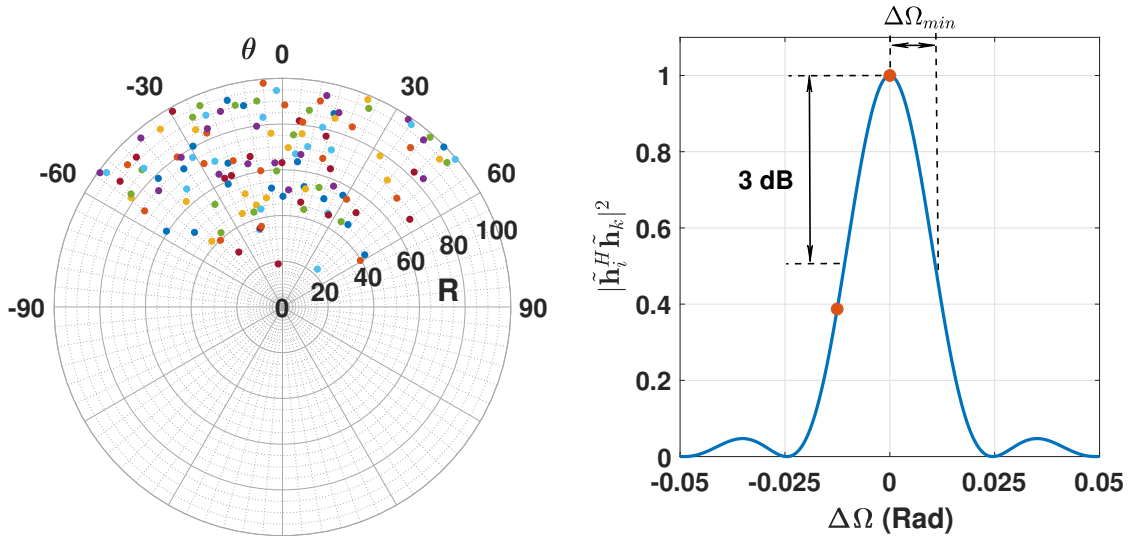
deployed in the cell. Then, we apply the analytical design framework to define the specification on the receive chain: the passband/baseband nonlinearity and the ADC resolution. We then evaluate the efficacy of the framework in attaining the desired system-level performance by simulations for selected scenarios. Finally, we provide design guidelines on the receive chain requirements in a more comprehensive set of scenarios.

2.6.1 User Distribution

The mobiles are uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, $R_{\min}$ and $R_{\max}$, respectively. Since $\frac{d\Omega}{d\theta} \sim \cos\theta$, the spatial frequency is less responsive to changes in DoA for θ near $\pm\frac{\pi}{2}$, which makes it more difficult to separate mobiles towards the edge of the angular field of view. We therefore confine the field of view for the antenna array to $-\pi/3 \leq \theta \leq \pi/3$. While the mobiles are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two mobiles in order not to incur excessive interference, choosing it as half the 3 dB beamwidth: $\Delta\Omega_{\min} = \frac{2.783}{N}$ [4] (mobiles closer in spatial frequency could be served in different time slots, for example). An example distribution of mobiles is depicted in Fig. 2.9.

2.6.2 Power Control Schemes

Our analysis in Section 2.5 first considers a system with no power control, in which each transmitter transmits at equal power. We then consider two power control schemes: a naive scheme in which transmitters adjust their powers to be roughly equal at the receiver, to within a tolerance, and an adaptive power control scheme designed for the linear MMSE receiver [50] aimed at meeting an SINR target for each mobile at the receiver. Power control is a very well-studied area, hence our goal is to provide quick insight on



(a) Example distribution of mobiles

(b) Normalized spatial cross-correlation

Figure 2.9: (a) An instantiation of 128 mobiles on a polar chart. (b) Normalized correlation between two users with spatial frequency difference of $\Delta\Omega$. Note that the closest users, depicted by red points, are separated by larger or equal to half the 3 dB beamwidth.

its implications for our system, rather than performing a comprehensive evaluation.

Naive power control

In this scheme, the base station asks all the users to decrease their power to make their received power at the base station equal the received power of the farthest mobile, i.e., at $R_{\max}$. A disadvantage of this scheme, illustrated by our performance results in subsequent subsections, is that nearby users are no longer able to use their larger signal strength to overcome the impact of interference from other users who are nearby (in terms of spatial frequency separation). The power factor α_p of the naive power control scheme is equal to 0 dB because all the users have the same received signal strength.

Adaptive power control

In order to avoid the pitfalls of naive power control, we consider an adaptive power control scheme aimed at meeting an SINR target SINR_{th} at the output of the linear MMSE receiver. This method was proposed and shown to converge in [50]. We restate it with all values represented in the dB scale in Algorithm 1. Starting from no power control and all users transmitting at maximum power, the algorithm seeks to enforce a threshold SINR, termed SINR_{th}, iteratively as follows: every mobile with SINR greater than SINR_{th} reduces its power by $\text{SINR} - \text{SINR}_{\text{th}}$. This process is repeated until a convergence criterion is met. The power factor, α_p, that results from this adaptive power control scheme is found via simulation to equal about -2 dB for our running example.

Algorithm 1: Adaptive power control

Input: $\mathbf{H}$, $\{P_k^{(0)}\}_{k=1,\ldots,K}$
parameter: SINR_{th}, ϵ
Output: $\{P_k\}_{k=1,\ldots,K}$

```
Margin = ε + 1;
while Margin > ε do
    {SINR_k}_{1:K} ← calculate LMMSE output SINR;
    for k ← 1 to K do
        ΔSINR_k ← max{SINR_k − SINR_th, 0};
        P_k ← P_k − ΔSINR_k;
    Margin ← max{ΔSINR_k}_{1:K};
```

2.6.3 Applying the Design Framework

For illustration, we consider $\beta = \frac{1}{2}$ and $\beta = \frac{1}{16}$, with no power control, naive power control and adaptive power control.

The design steps are as follows:

1. System-level design: We require $\mathrm{SINR}_{\mathrm{edge}}(\mathrm{ideal}) \approx 10$ dB for our target QoS. Using simulations for the ideal system, we compute the LMMSE efficiency η_{ideal} as shown in Fig. 2.8. For our four scenarios, the LMMSE efficiency η_{ideal} is found to be (a) 4.5 dB, (b) 0 dB, (c) 4.5 dB, and (d) 0 dB.

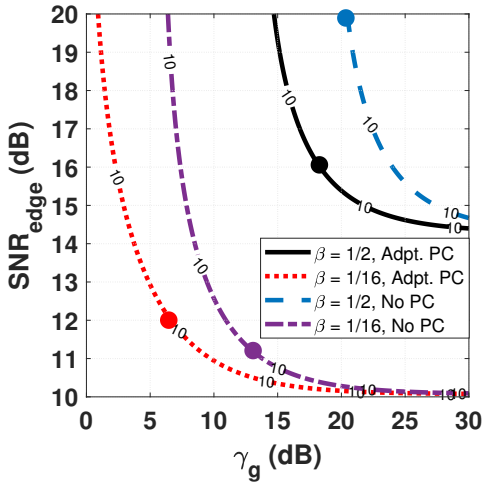
   After that, we determine the SNR of the edge mobile and the intrinsic SNR jointly to achieve the LMMSE lower bound. Specifically, the contours in Fig. 2.10 (a) illustrates the following equation for each scenario:

$$\mathrm{SNR}(g, \sigma_n^2) = \frac{\mathrm{SINR}_{\mathrm{edge}}}{\eta_{\mathrm{ideal}}},$$
$$\frac{1}{\frac{\beta}{\gamma_g \alpha_p} + \frac{1+\gamma_g}{\gamma_g} \frac{1}{\mathrm{SNR}_{\mathrm{edge}}}} = \frac{10}{\eta_{\mathrm{ideal}}}.$$

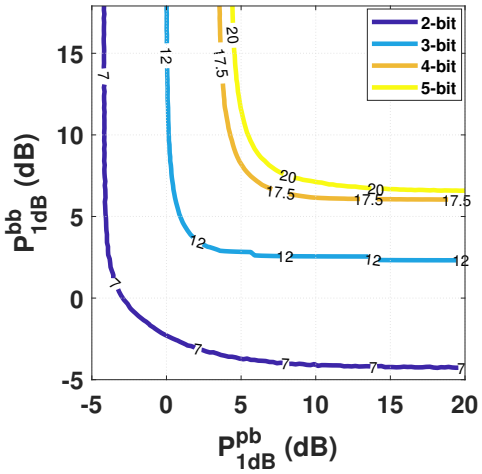
   We pick the following combinations of ($\mathrm{SNR}_{\mathrm{edge}}$,$\gamma_g$): (a) (20,20) dB, (b) (11,12) dB, (c) (16,17.5) dB, and (d) (12,7) dB.

2. Hardware-level design: This step determines the specifications of the passband/baseband nonlinearity and the ADC to achieve the required intrinsic SNR. Fig. 2.10 (b) shows the tradeoff between the number of ADC bits and the 1 dB compression point of the baseband nonlinearity $P_{\mathrm{1dB}}^{\mathrm{bb}}$ and the passband nonlinearity $P_{\mathrm{1dB}}^{\mathrm{pb}}$. The 1 dB compression points computed are normalized to the input power. The absolute compression points in dBm are computed by determining the average received input power at each base station antenna.

Here we have taken the link budget, or attainable $\mathrm{SNR}_{\mathrm{edge}}$, as our constraint, and have designed the nonlinearity specifications accordingly, We can, of course, utilize the same framework to determine the link budget required for a given set of nonlinearities.

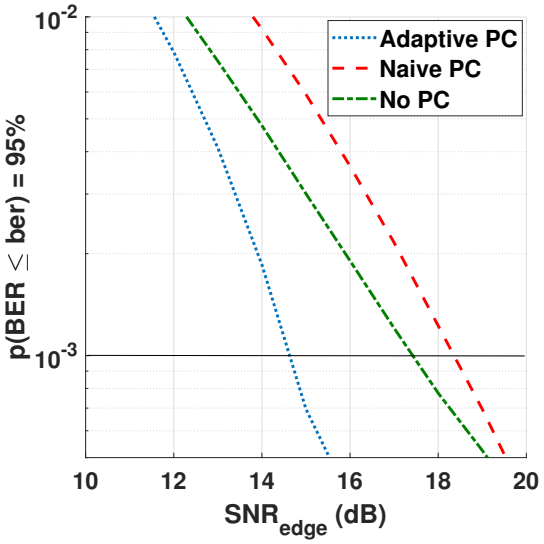


(a) Contours of $\text{SINR}_{\text{edge}}$

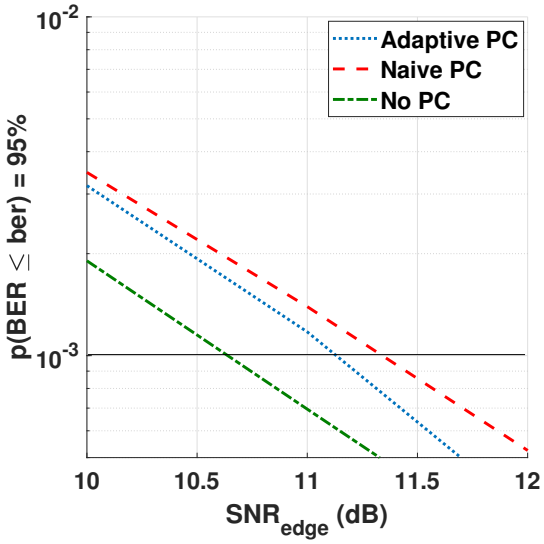


(b) Contours of intrinsic SNR γ_g

Figure 2.10: (a) Lower bound on the linear MMSE output SINR as a function in the intrinsic SNR, γ_g, and the SNR required for the edge user, SNR_{edge}, for different scenarios. The contours depicted are for constant $\text{SINR}_{\text{edge}} = 10$ dB. The solid circles in Fig. (a) show the operating points we choose to work at. (b) Intrinsic SNR of a receive chain comprising passband and baseband nonlinearities and ADC.



(a) BER with load factor of 1/2



(b) BER with load factor of 1/16

Figure 2.11: (a) and (b) show the BER attained by 95% of the users for load factor of 1/2 and 1/16, respectively. The SNR_{edge} is the SNR required by the user at 100 m away from the base station. The receive chain specifications for each curve are demonstrated in table 2.1.

β	PC	b	$P_{\text{1dB}}^{\text{bb}}$ (dB)	$P_{\text{1dB}}^{\text{pb}}$ (dB)	γ_g (dB)	SNR_{edge} (upper bound)	SNR_{edge} (sim.)
1/2	none	5	8.4	6.7	20	20	17.5
1/2	naive	4	8.4	4.9	17.5	18.7	18.4
1/2	adaptive	4	8.4	4.9	17.5	16	14.7
1/4	none	4	8.2	2.4	15	14	12.8
1/4	naive	3	3.7	0.7	10.5	15	14.8
1/4	adaptive	3	3.4	1.9	11.5	12.5	11.9
1/8	none	3	4.2	1.4	12	13	12.2
1/8	naive	3	2.2	-1.1	8.7	12.7	12.7
1/8	adaptive	3	3.2	1.9	11	10.9	10.8
1/16	none	3	4.2	1.4	12	11.2	10.8
1/16	naive	2	1.4	-1.1	7.6	11.5	11.5
1/16	adaptive	2	-1.1	-1.9	7	11.8	11.2

Table 2.1: This table presents the analytical predictions and simulation results for the SNR budget needed to meet the desired performance criterion (10^{-3} BER at 95% availability) for different scenarios. The intrinsic SNR γ_g corresponds to the cascade of the passband and baseband nonlinearities, specified by their 1 dB compression points ($P_{\text{1dB}}^{\text{pb}}$ and $P_{\text{1dB}}^{\text{bb}}$, respectively), together with b-bit ADCs for I and Q. PC and β denote the power control scheme used, and the load factor, respectively.

2.6.4 Simulation-based Verification

Here, we verify the designs produced by our analytical framework by numerical simulations. In Fig. 2.11, we plot the BER that 95% of the users attain for the cases we mention in the previous subsection. As shown, all the curves reach the 10^{-3} at slightly smaller SNR_{edge} than predicted by our analytical framework, which shows that our approach is both conservative and accurate. Table 2.1 summarizes our design prescriptions for different scenarios. As shown in the table, we examine the combination of four load factors with no power control and two power control strategies. We demonstrate the specification of the receive chain along with the resultant intrinsic SNR, γ_g. Then we compute an upper bound for the SNR needed for the edge user to achieve the performance metric. Finally, using simulations, we show the accuracy of the derived upper bound.

The results reported in Table 2.1 show that massive MIMO is key to relaxed hardware specifications: increasing the number of antennas for a given number of users reduces load factor, providing a "degrees of freedom" advantage that is used to level out the distortions caused by nonlinearities. For example, when serving 16 users with a 256 element array ($\beta = \frac{1}{16}$), only 2 bits of quantization per dimension is required and the 1-dB compression point of RF and baseband amplifiers is very low (both below 0 dB with adaptive power control). In practice, a lower compression point allows higher loading, i.e., the amplifier can support a larger input signal and produce a stronger signal in the output, which increases the power efficiency of an amplifier. This is very desirable trait for scaling to large arrays. The results also bring out the impact of power control. Higher disparities in user powers requires higher ADC granularity to allow effective interference suppression of strong users which might otherwise "drown out" weak users. The reduction in user power disparities and required dynamic range with better power control is reflected clearly in our results.

2.7 Conclusion

The analytical framework provided in this chapter is a conservative, yet accurate, approach for designing hardware specifications for nonlinear elements in all-digital mmWave massive MIMO. Scaling using a larger number of antennas with a smaller load factor is attractive, since the specifications for RF nonlinearities, baseband nonlinearities, and ADC precision can all be significantly relaxed by operating at lower load factors. The requirements can also be relaxed by use of appropriate power control, as illustrated by the simple adaptive power control scheme considered here.

While we have considered LoS channel models here, we note that our approach extends to sparse multipath channels. At high symbol rates, equalization over a large delay spread

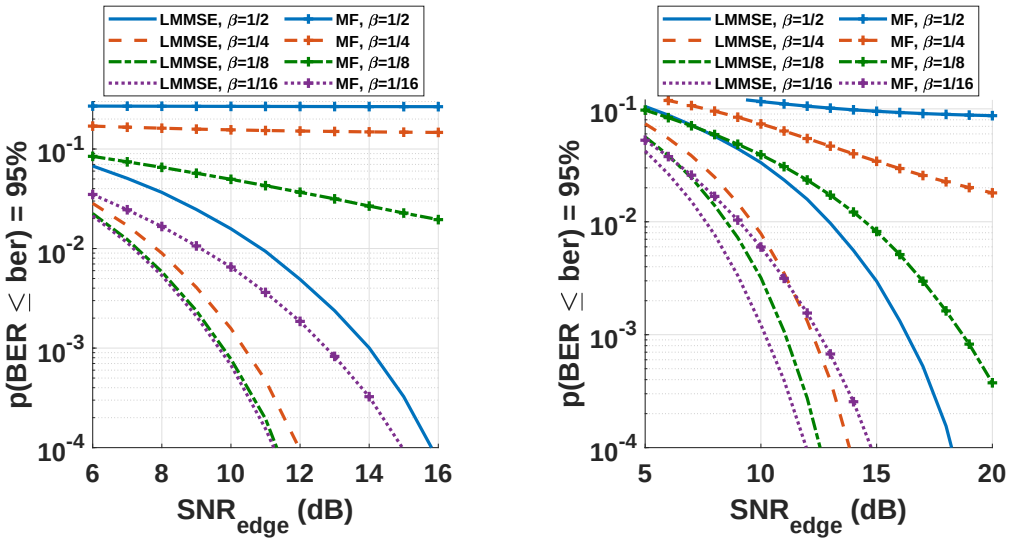
becomes computationally unattractive. In this case paths that differ significantly in delay and angular spread from the dominant path play the role of additional interference, and can be folded into our framework.

In addition to the extensive effort required to realize our design prescriptions in hardware, there are also important open issues related to the digital backend, given the challenges of both computation and data transport for the multiGigabaud, multiuser system considered here. Thus, despite the extensive prior research on multiuser detection, there are significant open issues on the design of strategies that are efficient enough (in terms of both computation and communication on the backend fabric) to scale with the number of antennas, number of users, and bandwidth. We also note the importance of exploring *nonlinear* reception techniques (such as interference cancellation) that could outperform LMMSE detection while maintaining low computational overhead. As discussed in Chapters 3 through 6, exploiting channel sparsity is a promising approach for developing such techniques.

Chapter 3

Beamspace Local LMMSE: An Efficient Digital Backend for mmWave Massive MIMO

All-digital architectures enable taking full advantage of the large number of antennas that can be integrated in mmWave transceivers, with fully flexible beamforming that enables the number of simultaneous users K sharing the band to scale with the number of antennas N, with scaling ratio, or *load factor*, $\beta = \frac{K}{N}$. Standard criteria for beamforming include spatial matched filtering (MF), as well as linear interference suppression using the zero forcing (ZF) or linear minimum mean square error (LMMSE) criteria. Fig. 3.1(a) depicts the raw bit error rate (BER) achieved by 95% of the mobiles for the picocellular uplink considered in this chapter. Clearly, interference suppression becomes necessary for moderate load factors (e.g., $\beta > 1/16$), where MF performance is far inferior to that of LMMSE, with the gap persisting even if power control is employed, as shown in Fig. 3.1(b). However, the computational complexity of LMMSE detection becomes prohibitive for large K and N.



(a) BER without power control

(b) BER with power control

Figure 3.1: Massive MIMO uplink performance using MF and LMMSE receivers for $\beta = K/N = \{1/16, 1/8, 1/4, 1/2\}$ and $N = 256$.

Recent efforts at complexity reduction, for both uplink and downlink, include two-stage beamforming strategies [51, 52, 53, 54]. In [51], a statistical outer beamformer based on grouping mobiles based on similar correlation matrices reduces the effective spatial dimension of the equivalent channels [51, 52]. This is followed by an inner beamformer that suppresses both intra- and inter-group interference, resulting in significant reduction in computation [53, 54].

In this chapter, we propose a Beamspace Local LMMSE algorithm which leverages the sparsity of the spatial channel in mmWave bands. A spatial discrete Fourier transform (DFT) is employed to concentrate the energy of each mobile into a smaller number of DFT bins, i.e., in "beamspace." We show that performance close to that of standard LMMSE can be obtained by a local LMMSE detector operating on a beamspace window of a size that does not scale with N. We provide analytical rules of thumb for choosing window size as a function of load factor β and target outage rate. We also show how our architecture provides a low-complexity solution for implicit channel estimation via

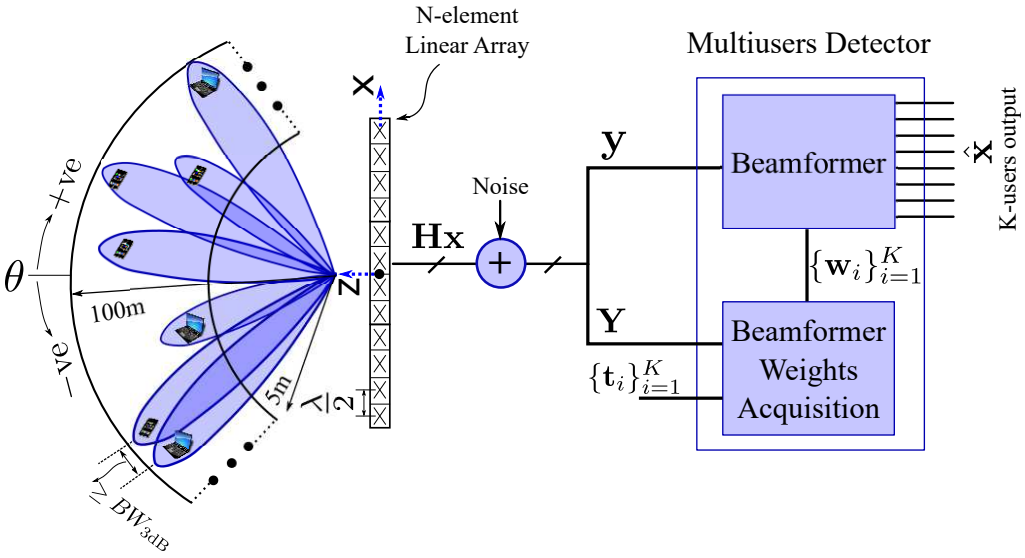


Figure 3.2: System model for the beamspace massive MIMO.

an efficient adaptive implementation.

3.1 System Model

The massive MIMO system model is depicted in Fig. 3.2: a sector covered by a base station which can scan horizontally with a 1D half-wavelength spaced N-element array, serving K mobiles (assumed to each have a single antenna for simplicity).

The linear multiuser detector comprises a beamformer weights acquisition module, and a beamformer module. Using the received training matrix $\mathbf{Y}$, defined later, and the training sequences for each user $\{\mathbf{t}_i\}_{i=1}^{K}$, the weight acquisition block generates the beamformer weights $\{\mathbf{w}_i\}_{i=1}^{K}$. These weights are used by the beamformer module to estimate the users' data vector $\mathbf{x}$ out of the received vector $\mathbf{y}$.

We assume a line-of-sight (LoS) channel from each mobile to the base station. The

channel vector for the k^{th} user can be written as follows:

$$\mathbf{h}_k = A_k \left[1\, e^{j\Omega_k}\, e^{j2\Omega_k}\, \ldots\, e^{j(N-1)\Omega_k}\right]^{\intercal}, \tag{3.1}$$

where $A_k^2 = \left(\frac{\lambda}{4\pi R_k}\right)^2$ depends on the radial location R_k of user k and wavelength λ, using the Friis formula for path loss. A spatial frequency $\Omega_k = 2\pi \frac{d_x}{\lambda} \sin\theta_k$ defines the angular location of the k^{th} user, where d_x is the inter-distance between antenna elements and chosen to be $\lambda/2$.

The received signal vector $\mathbf{y}$ in the base station is given by

$$\mathbf{y} = \underbrace{\mathbf{H}}_{N\times K}\mathbf{x} + \mathbf{n}, \tag{3.2}$$

where $\mathbf{H} = [\mathbf{h}_1\mathbf{h}_2 \ldots \mathbf{h}_K]$ is the channel matrix, $\mathbf{x}$ is the users symbols vector, $\mathbb{E}(x_k) = 1$, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the AWGN vector. As shown in Fig. 3.3, the LoS channel is sparse in beamspace.

3.1.1 Linear MMSE receiver

The LMMSE receiver is given by $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$, such that

$$\mathbf{W} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \sigma^2\mathbf{I})^{-1} = (\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I})^{-1}\mathbf{H}^H, \tag{3.3}$$

where the second equality is a direct result of Sherman–Morrison–Woodbury matrix identity.

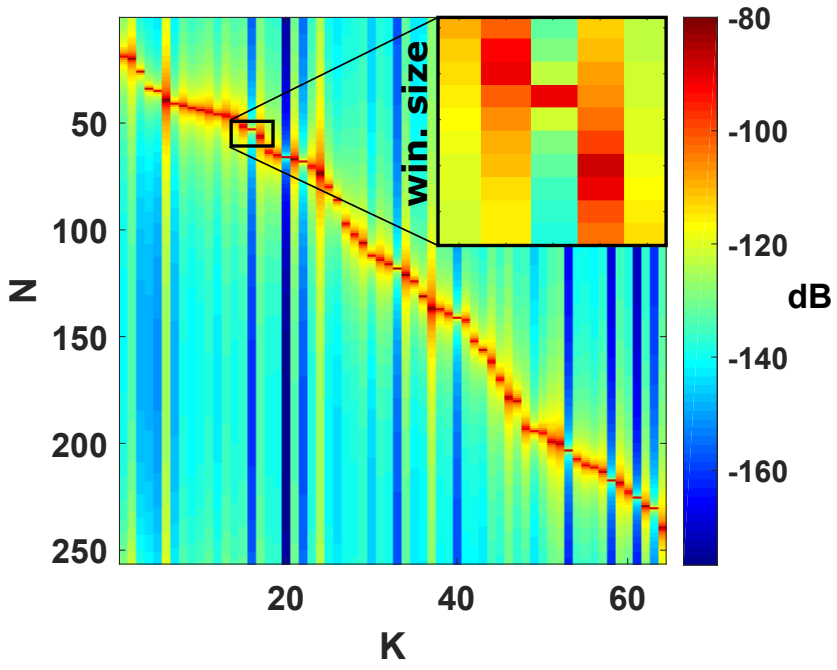


Figure 3.3: Sparse LoS channel in the beamspace.

3.1.2 Implicit channel estimation

Adaptive implementations of the LMMSE receiver implicitly estimate the channel using training sequences for the users of interest. Let $\mathbf{T} = [\mathbf{t}_1\mathbf{t}_2 \dots \mathbf{t}_K]^\intercal$ be a matrix that hold the training sequences $\mathbf{t}_k, k = 1, \dots, K$ of length L, such that $||\mathbf{t}_k||_2^2 = L$. We assume that sequences are orthogonal across users, i.e., $\mathbf{t}_i^H \mathbf{t}_j = 0 \forall i \neq j$. Hence the received training sequence can be written as,

$$\underbrace{\mathbf{Y}}_{N \times L} = \mathbf{HT} + \mathbf{N}, \tag{3.4}$$

where $\mathbf{N} = [\mathbf{n}_1\mathbf{n}_2 \dots \mathbf{n}_L]$ is the noise matrix. Therefore, the covariance of the received signal and the channel matrix can be computed empirically and approximated at high

SNR as (assuming $L > N$)

$$\mathbf{C} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H = \mathbf{H}\mathbf{H}^H + \frac{1}{L}\mathbf{N}\mathbf{N}^H \approx \mathbf{H}\mathbf{H}^H + \sigma^2\mathbf{I}, \tag{3.5}$$

and

$$\mathbf{U} = \frac{1}{L}\mathbf{T}\mathbf{Y}^H = \mathbf{H}^H + \frac{1}{L}\mathbf{T}\mathbf{N}^H \approx \mathbf{H}^H, \tag{3.6}$$

while the computational complexity of the previous two steps are $O(N^2L)$ and $O(\beta N^2 L)$, respectively. The LMMSE solution can be formed as,

$$\hat{\mathbf{x}} = \mathbf{U}\mathbf{C}^{-1}\mathbf{y}. \tag{3.7}$$

If we used the Cholesky decomposition [55], then the computation complexity of inverting $\mathbf{C}$ is $O(N^3)$, while the complexity of computing $\hat{\mathbf{x}}$ is $O(\beta N^2)$ per received signal vector $\mathbf{y}$.

3.2 Beamspace Local LMMSE

In general, the observation at each antenna is a superposition of signals received from all the users. For a specific user, the information about its symbol is distributed equally between all the antennas. Hence, the base station must engage all the antenna observations to extract the user's symbol with high precision. In beam domain processing, on the other hand, the base station compresses the information of a user into a smaller number of spatial observations. As a result, the computational complexity can be greatly reduced.

In this section, we first describe the method to attain the weights of beamspace local LMMSE in the absence of explicitly estimated CSI using the received training sequence

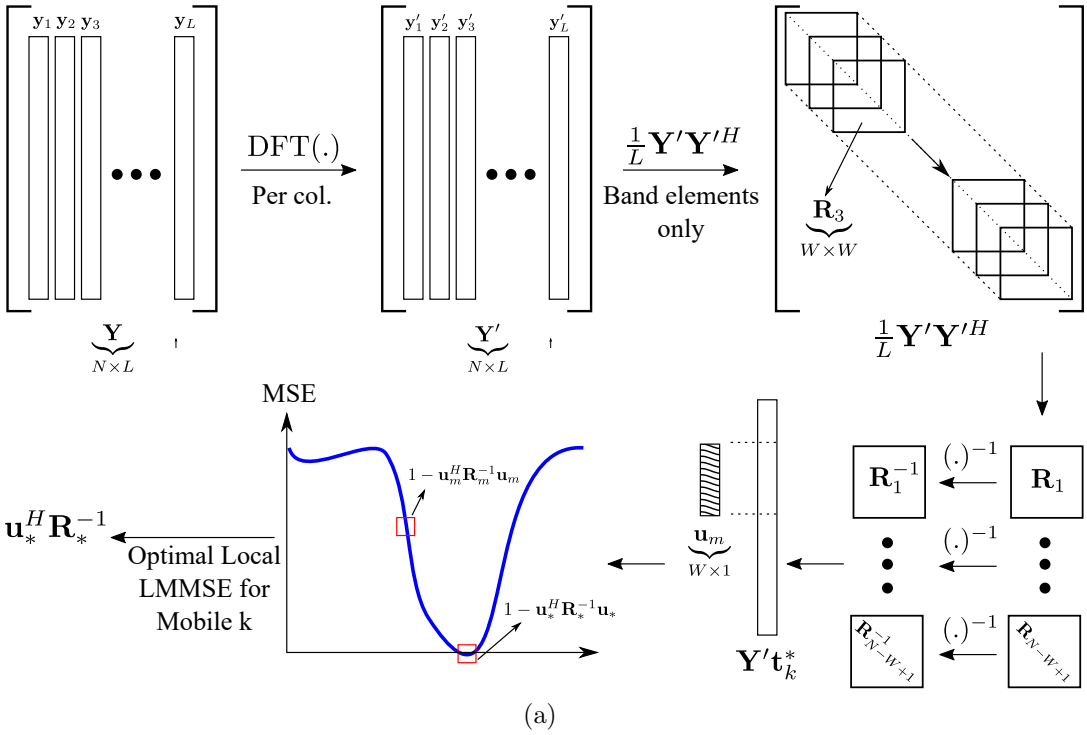


Figure 3.4: Local LMMSE weights acquisition in beamspace.

$\mathbf{Y}$. Then, we explain the beamforming process for the received data vector $\mathbf{y}$.

3.2.1 Beamspace Local LMMSE Weight Acquisition

Fig. 3.4 summarizes the steps of the local LMMSE weights acquisition. Starting by the received training sequence matrix $\mathbf{Y}$, the acquisition steps are given as follows.

Calculate the DFT of the received training sequence

The DFT is used to transform the antenna domain signal to beam domain by applying it on each column of the received training sequence matrix $\mathbf{Y}$ to get $\mathbf{Y}'$ as

$$y'_{p,\ell} = \sum_{n=1}^{N} y_{n,\ell} e^{-j2\pi(n-1)(p-1)/N}, \tag{3.8}$$

where $\mathbf{Y} = [y_{n,\ell}]$ and $\mathbf{Y}' = [y'_{p,\ell}]$, $\forall p = 1, \ldots, N$. The fast Fourier transform (FFT) algorithm [56] is applied to realize the DFT operation with a complexity of $O(LN \log N)$ for the whole training sequence matrix.

Generate the band matrix $\tilde{\mathbf{C}}$

The band matrix $\tilde{\mathbf{C}}$ is generated as

$$\tilde{c}_{n,p} = \begin{cases} \frac{1}{L}\mathbf{Y}'_{[n,*]}\mathbf{Y}'^{H}_{[p,*]}, & \text{if } |n-p| < W \\ 0, & \text{otherwise,} \end{cases} \tag{3.9}$$

where $\mathbf{Y}'_{[n,*]}$ is the n^{th} row in the matrix $\mathbf{Y}'$. Instead of computing the entire $N \times N$ beam domain sample covariance matrix ($\frac{1}{L}\mathbf{Y}'\mathbf{Y}'^{H}$), only the dominant elements around the diagonal within a window size W are computed. Hence, the complexity of this step is $O(WNL)$. We define the block matrices on the diagonal $\mathbf{R}_m$ as

$$\underbrace{\mathbf{R}_m}_{W \times W} = \tilde{\mathbf{C}}_{[m:m+W-1, m:m+W-1]}. \tag{3.10}$$

Invert each $\mathbf{R}_m$

One matrix inversion $\mathbf{R}_m^{-1}$ has a complexity of $O(W^3)$. Normally, the matrix inversion should be applied $N - W + 1$ times for all possible m, and the resulting total complexity would be $O(NW^3)$. In the following, we propose a method that greatly reduces the number of operations required. The beam domain sample covariance matrix $\tilde{\mathbf{C}}$ has the

following structure

$$\tilde{\mathbf{C}} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \dots & c_1 & \mathbf{b}_1^H & \cdot & \dots \\ \dots & \mathbf{b}_1 & \mathbf{A} & \mathbf{b}_2 & \dots \\ \dots & \cdot & \mathbf{b}_2^H & c_2 & \dots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.11}$$

Let us define

$$\mathbf{R}_m = \begin{bmatrix} c_1 & \mathbf{b}_1^H \\ \mathbf{b}_1 & \mathbf{A} \end{bmatrix} \text{ and } \mathbf{R}_{m+1} = \begin{bmatrix} \mathbf{A} & \mathbf{b}_2 \\ \mathbf{b}_2^H & c_2 \end{bmatrix}. \tag{3.12}$$

Using the elements in (3.12), $\mathbf{R}_m^{-1}$ can be defined as

$$\begin{aligned} \mathbf{R}_m^{-1} &= \begin{bmatrix} \frac{1}{s_1} & -\frac{1}{s_1}\mathbf{b}_1^H\mathbf{A}^{-1} \\ -\frac{1}{s_1}\mathbf{A}^{-1}\mathbf{b}_1 & \mathbf{A}^{-1} + \frac{1}{s_1}\mathbf{A}^{-1}\mathbf{b}_1\mathbf{b}_1^H\mathbf{A}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} x_{11} & \mathbf{x}_{12}^H \\ \mathbf{x}_{12} & \mathbf{X}_{22} \end{bmatrix}, \end{aligned} \tag{3.13}$$

where $s_1 = c_1 - \mathbf{b}_1^H\mathbf{A}^{-1}\mathbf{b}_1$ is the Schur complement of block $\mathbf{A}$ of matrix $\mathbf{R}_m$. Given $\mathbf{A}^{-1}$, the inverse $\mathbf{R}_{m+1}^{-1}$ for the block $m+1$ can be computed as

$$\mathbf{R}_{m+1}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \frac{1}{s_2}\mathbf{A}^{-1}\mathbf{b}_2\mathbf{b}_2^H\mathbf{A}^{-1} & -\frac{1}{s_2}\mathbf{A}^{-1}\mathbf{b}_2 \\ -\frac{1}{s_2}\mathbf{b}_2^H\mathbf{A}^{-1} & \frac{1}{s_2} \end{bmatrix}, \tag{3.14}$$

where $s_2 = c_2 - \mathbf{b}_2^H \mathbf{A}^{-1} \mathbf{b}_2$ is the Schur complement of block $\mathbf{A}$ of matrix $\mathbf{R}_{m+1}$. Finally, $\mathbf{A}^{-1}$ can be computed from the entries of $\mathbf{R}_m^{-1}$ as

$$\mathbf{A}^{-1} = \mathbf{X}_{22} - \frac{1}{x_{11}} \mathbf{x}_{12} \mathbf{x}_{12}^H. \tag{3.15}$$

The matrix inversion is computed only once for a particular m, while the rest of the matrix inverses are generated using the above recursive steps. Therefore, the total complexity is reduced to $O(NW^2)$.

Compute $\mathbf{h}'_k$ for each user k

The beam domain channel $\mathbf{h}'_k$ of user k can be computed empirically as

$$\mathbf{h}'_k \approx \frac{1}{L} \mathbf{Y}' \mathbf{t}_k^*, \tag{3.16}$$

while the approximation error vanishes for large L and SNR. The complexity of computing $\mathbf{h}'_k$ for all users is $O(NLK)$. Finally, the beam domain channels of size W for each sliding window m are constructed as

$$\mathbf{u}_{k,m} = \mathbf{h}'_k[m : m + W - 1]. \tag{3.17}$$

Search the index m with Minimum MSE

In this step, the window m providing the minimum mean square error (MSE) is found for each user k. The MSE criterion for each index pair (k, m) is given as [57]

$$MSE_{k,m} = 1 - \mathbf{u}_{k,m}^H \mathbf{R}_m^{-1} \mathbf{u}_{k,m}, \tag{3.18}$$

and the optimum window index m for each user k is found as

$$m_k^* = \arg\min_m MSE_{k,m}. \tag{3.19}$$

The complexity of the entire search process is $O(\beta N^2 W^2)$. Finally, the optimum local LMMSE weights for each user k in the beamspace is given as

$$\mathbf{w}_k = \mathbf{u}_{k,m_k^*}^H \mathbf{R}_{m_k^*}^{-1}. \tag{3.20}$$

3.2.2 Beamspace Local LMMSE Beamforming process

Given the computed beamspace weights computed in the previous subsection, the detection of data symbols from (3.2) simply consists of two steps:

Calculate the DFT $\mathbf{y}'$ for the received vector $\mathbf{y}$

The complexity of this step is $O(N \log N)$.

Calculate the estimate $\hat{x}_k = \mathbf{w}_k \mathbf{y}'[m_k^* : m_k^* + W - 1] \ \forall \ k$

The complexity of this step is $O(\beta W N)$.

3.3 Window Size W Does Not Scale with N

In this section, we sketch an argument showing that the required window size W does not scale with the number of antenna elements N. Under a simplified model of user's spatial distribution as being uniform across the N FFT bins, the number of users falling into a typical window is a binomial random variable $X \sim Bin(K, W/N)$. For $\beta < 1$, the mean number of users falling in a window, which is $\nu = KW/N = \beta W$, is smaller than

the available dimension W, which implies that linear interference suppression is expected to be successful for a window size W, where the choice of W depends only on β, and does not scale with N.

We can now obtain rules of thumb on the choice of W as a function of β and a target outage probability. For large N, K and fixed β, X tends to a Poisson with mean ν, with Chernoff bound on tail probabilities[58]

$$P(X \geq x) \leq \frac{e^{-\nu}(e\nu)^x}{x^x}, \tag{3.21}$$

Assuming that outage in a given window occurs if and only if the number of users $X > W$, the probability of outage in a window is given by $P[X > W]$. Assuming that all users falling in the window are in outage when this happens, the expected number of users in outage, after direct mathematical manipulation, is $E[XI_{X>W}] = \nu P[X \geq W] = \beta W P[X \geq W]$. Summing over the $N - W + 1$ distinct windows and dividing by the number of users K, we obtain

$$P(outage) = \frac{N - W + 1}{K} \beta W P[X \geq W] \leq W P[X \geq W]$$

Plugging in (3.21), we obtain upon simplification that

$$P(outage) \leq W \left(\beta e^{1-\beta}\right)^W \tag{3.22}$$

Given a 5% outage target, the window size computed using the above formula would be 2, 4, 8, 34 for load factors 1/16, 1/8, 1/4, 1/2, respectively, which are remarkably close to those obtained by simulations in Section 3.4.

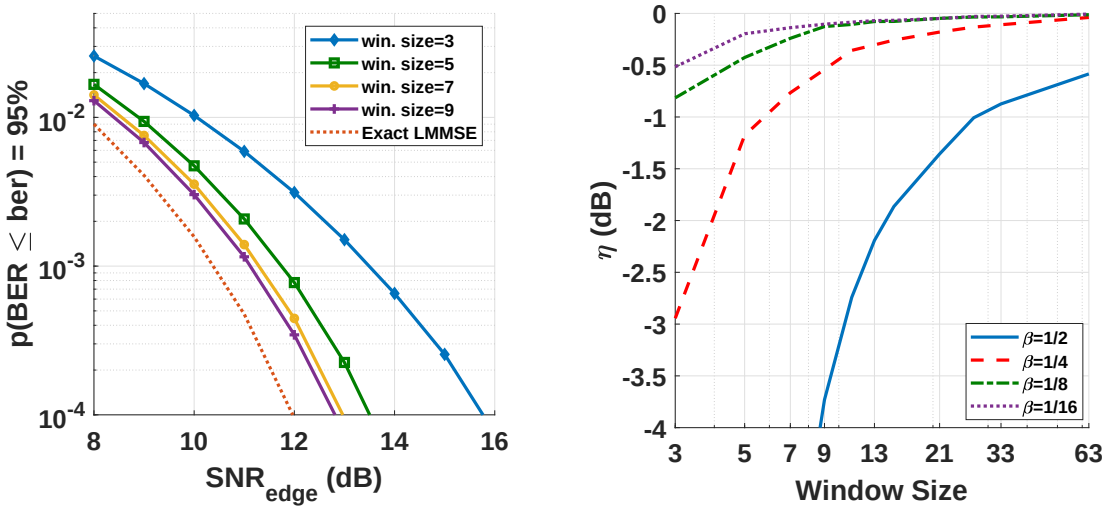
3.4 Numerical Results

We consider the system setup in Fig. 3.2 with number of antennas fixed at $N = 256$ for all numerical experiments. The field of view for the sector is restricted to $-\pi/3 \leq \theta \leq \pi/3$. The users are uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, $R_{\min} = 5$ m and $R_{\max} = 100$ m, respectively. While the user terminals are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two users in order not to incur excessive interference, arbitrarily choosing it as half the 3 dB beamwidth: $\Delta\Omega_{\min} = \frac{2.783}{N}$ [4]. BW_{3dB} in Fig. 3.2 stands for the 3 dB beamwidth. We assume that users with similar spatial frequency can be served in different time or frequency resource blocks.

We measure link quality by the outage probability at a target uncoded BER of 10^{-3} for QPSK which requires SNR of 9.7 dB for a SISO AWGN link. This becomes the target SINR at the output of the multiuser detector for an edge user. No power control is deployed. The efficiency of the proposed beam space local LMMSE is defined as the ratio between the SNR_{edge} required to attain the link quality using standard LMMSE, relative to that required with local LMMSE:

$$\eta = \frac{SNR_{edge}(LMMSE)}{SNR_{edge}(local)} \tag{3.23}$$

Assuming perfect CSI, Fig. 3.5(a) shows the BER achieved by at least 95% of the users for different window sizes W and with load factor $\beta = 1/4$. Fig. 3.5(b) illustrates the efficiency η of the beamspace local LMMSE where the edge user SNR is adjusted such that at least 95% of the users achieve BER of 10^{-3}. In order to incur loss of only 1 dB or less in performance, a window size of 2, 3, 7, and 31 should be applied for load factors 1/16, 1/8, 1/4, and 1/2, respectively.

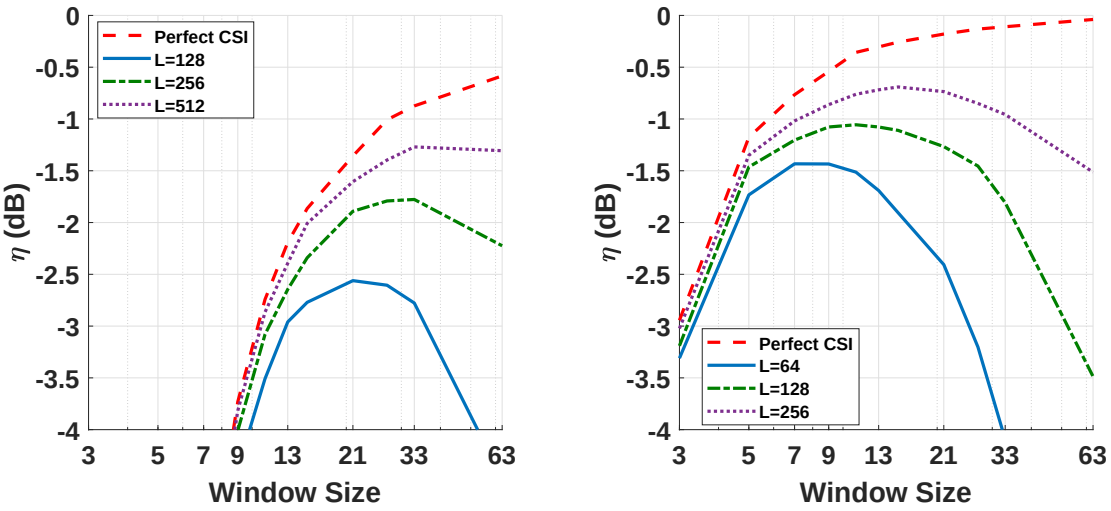


(a) $\beta = 1/4$ (b) Target BER 10^{-3}

Figure 3.5: (a) BER achieved by at least 95% of the users for different W. (b) Edge user η with $\beta = \{1/2, 1/4, 1/8, 1/16\}$

Figs. 3.6(a) and 3.6(b) plot the performance of the local LMMSE with implicit channel estimation as a function of training samples L, for $\beta = 1/2$ and $\beta = 1/4$, respectively. The training sequences are constructed from the Hadamard matrix such that user sequences are mutually orthogonal. There is a clear trade-off between the window size W and performance of the beamspace local LMMSE receiver (3.20). The optimum window size depends on both L and β. A larger window W provides an advantage in terms of the degrees of freedom available for suppressing inter-user interference in $\mathbf{R}_{m_k^*}$; see analysis in Section 3.3. However, at the same time, the estimation quality of outermost elements in $\mathbf{u}_{k,m_k^*}$ quickly decays as W is increased due to the sinc-like shape of the DFT beams. Thus, for finite L, the performance of (3.20) begins to deteriorate if W is made too large.

For a single user, local LMMSE approximates spatial matched filtering. Fig. 3.7 shows the worst-case loss in the output SNR of local LMMSE compared to the MF for a single user, which arises due to off-grid effects: with probability one, a user's location in beamspace is not aligned with the DFT bins. As shown, $W = 3$ is enough to collect



(a) $\beta = 1/2$ (b) $\beta = 1/4$

Figure 3.6: Local LMMSE with implicit channel estimation.

most of the energy.

Fig. 3.8 compares, for $\beta = 1/4$, the complexity of conventional LMMSE and beamspace local LMMSE, resulting in 10-fold complexity reduction for $N = 256$ with local LMMSE for both weights acquisition (Fig. 3.8 (a)) and beamforming (Fig. 3.8 (b)).

3.5 Conclusion

We have shown that, for the sparse spatial channels typical of mmWave bands, we can scale up the number of antennas and users without scaling the dimension of the signal subspace required to demodulate a given user. Thus, once we incur the $O(N \log_2 N)$ complexity of performing a spatial DFT, we can significantly reduce the complexity of multiuser detection: for example, the beamspace local LMMSE approach studied here achieves a ten-fold reduction in complexity compared to conventional LMMSE for the range of system parameters considered here. We have also shown how an adaptive implementation of this approach can be extended to provide implicit channel estimation.

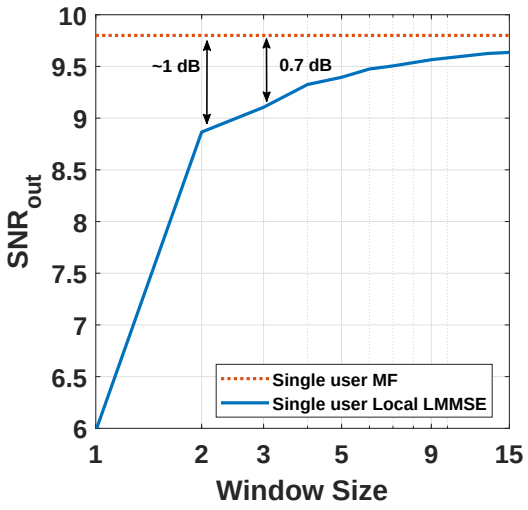
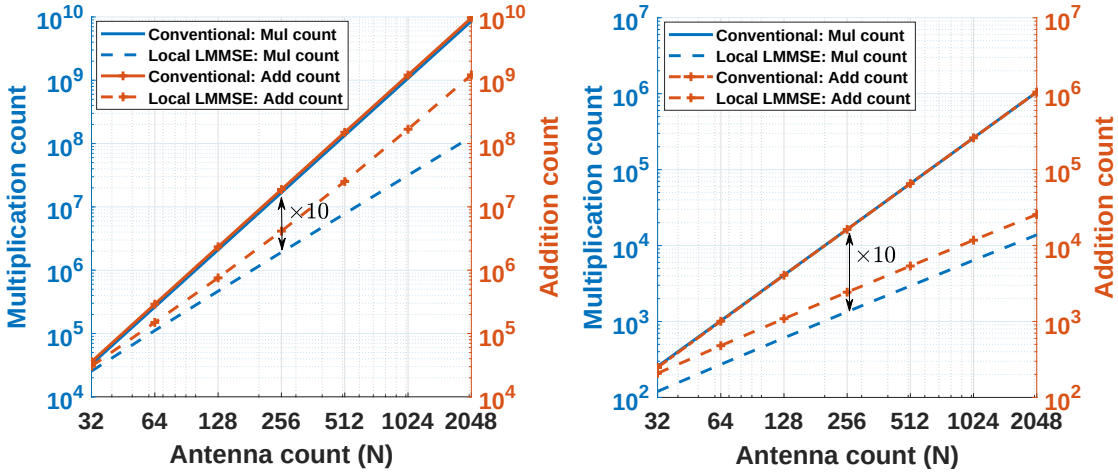


Figure 3.7: Local LMMSE vs. spatial MF for a single user.



(a) Acquisition complexity

(b) Beamforming complexity

Figure 3.8: Complexity comparison of beamspace local LMMSE and conventional LMMSE.

Chapter 4

Scalable Nonlinear Multiuser Detection for mmWave Massive MIMO

In Chapter 3, we show that effective linear interference suppression is provided by the "local LMMSE" receiver which demodulates each user from a small window of the signal vector in beamspace. The required size of the window depends on the load factor and the minimum user separation, but does not scale with N. In this chapter, we show that low-complexity nonlinear interference cancellation can be layered on top of such a local LMMSE receiver, enabling reliable demodulation at higher load factors.

The key idea is as follows. For any given "desired" user, the local LMMSE receiver effectively suppresses most of the interference except for a small number of users which are "nearby" (in beamspace). Therefore, the output of the local LMMSE receiver can be treated as a virtual MIMO system with a small number of users and colored noise. After noise whitening, we apply interference cancellation to this smaller system to enhance the reliability of demodulation for the desired user. We apply this second stage of interference

cancellation for each user separately. The parameters of our approach are the size of the window used for local LMMSE reception in beamspace, and the maximum number of interferers cancelled for each user in the second stage. Our approach retains the gains in computational efficiency obtained from the sparsity of the mmWave channel, while allowing us to push the system to higher load factors than is possible with linear interference suppression.

Related Work: Multiuser detection (MUD) for MIMO has a rich history [59], with many recent works focussing on complexity reduction motivated by massive MIMO. Matrix inversion in high dimensions is a bottleneck for linear interference suppression, and proposed complexity reduction techniques include Newton iteration [60], Neumann series expansion (NSE) [61], the Gauss-Seidel method [62, 63], and Cholesky decomposition [64]. The correlation structure of the matrix can be further exploited to reduce complexity. For example, [65] exploits a tridiagonal structure for the Wishart matrix in a VLSI implementation. Complexity reduction techniques for nonlinear MUD include [66], wherein a sphere decoder is selectively applied by leveraging a linear detector, and [67], where approximate message passing is applied to reduce sphere decoding complexity. A survey of massive MIMO detection techniques can be found in [68]. Unfortunately, these and other existing techniques are not easily scaled up to the system sizes that we consider here.

The results in Chapter 3 imply that, as long as we pay the $\mathcal{O}(N \log N)$ price of a spatial fast Fourier transform (FFT), linear interference suppression for each user can be accomplished at complexity that does not scale with system size, assuming that the mmWave channel is concentrated in beamspace. In this chapter, we show that similar conclusions hold for nonlinear multiuser detection as well, enabling us to push the system load factor further up without sactificing link reliability.

Notation: We use lowercase bold letters for vectors, and uppercase bold letters for

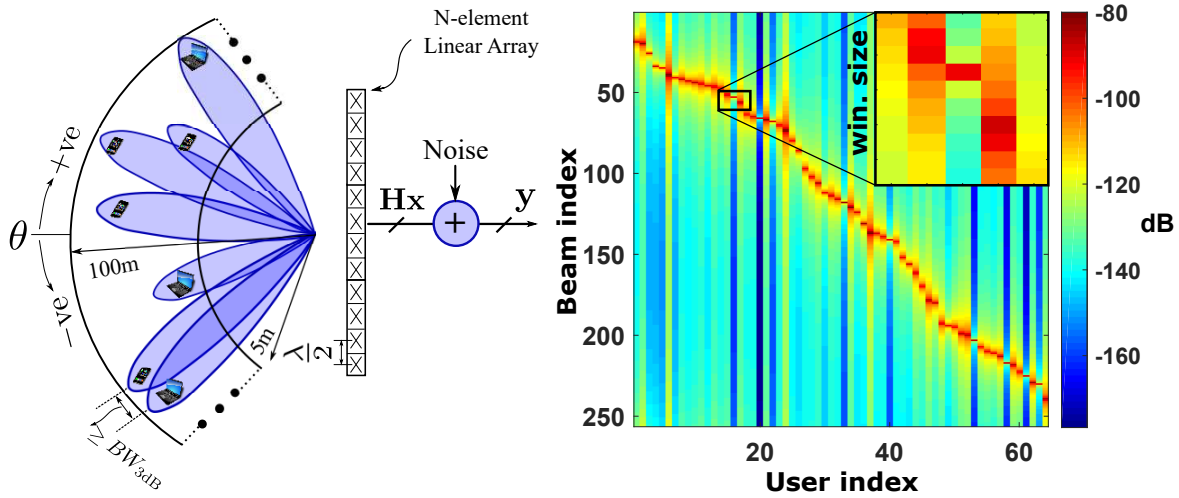
matrices. The notation $\mathbf{x} = [x_i]_{i=1}^{I}$ represents column vector $\mathbf{x}$ of length I and its elements are denoted by x_i. For a matrix, we use $\mathbf{X} = [x_{i,j}]_{i=1,j=1}^{I,J}$. $\{.\}_{k=1}^{K}$ denotes a list of K scalars, vectors or matrices. The identity matrix is denoted by $\mathbf{I}$ and $\mathbf{0}_M$ is a column vector which consists of M zeros.

4.1 System Model

We consider the uplink MIMO system depicted in Fig. 4.1a. The base station employs a linear array with N elements to simultaneously serve $K = \beta N$ mobile users, where β is the system load factor. Each mobile transmits a single data stream and uses an antenna array to perform ideal transmit beamforming towards the base station.

Channel Model: We assume that a single path dominates the channel between the base station and any mobile. Such a model is well suited for mmWave channels as it has been experimentally validated in a typical university campus at 60 GHz [69]. Therefore, the channel matrix is of the form $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \dots \mathbf{h}_K]$, where $\mathbf{h}_k$ is the $N \times 1$ spatial channel for the k^{th} mobile, $\mathbf{h}_k = \alpha_k \, [1 \, e^{j\Omega_k} \, e^{j2\Omega_k} \, \dots \, e^{j(N-1)\Omega_k}]^\intercal$. Here, Ω_k is the spatial frequency (corresponding to the angle of arrival) and α_k is the complex channel amplitude of the path of user k.

The single path channel has a concentrated structure in the discrete spatial frequency domain, or "beamspace", as shown for a typical example in Fig. 4.1b. We define the beamspace channel matrix as $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1 \dots \tilde{\mathbf{h}}_K]$, where $\tilde{\mathbf{h}}_k = \mathcal{DFT}(\mathbf{h}_k)$ and $\mathcal{DFT}(.)$ is the discrete Fourier transform (DFT) operator. The received signal vector in the original spatial domain is given by $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{x}$ is the $K \times 1$ vector of users' symbols, $\mathbb{E}(|x_k|^2) = 1$, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is additive white Gaussian noise (AWGN). By taking the DFT of this vector, we obtain the beamspace signal model, $\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{n}}$, where $\tilde{\mathbf{n}} = \mathcal{DFT}(\mathbf{n}) \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$. Our goal is to perform MUD by estimating the transmitted



(a) MIMO system model. (b) mmWave beamspace channel matrix.

Figure 4.1: (a) Uplink massive MIMO system model. (b) The sparsity of single-path channel in beamspace.

symbol vector, $\mathbf{x}$, from this beamspace signal, which is a sufficient statistic for estimating $\mathbf{x}$. We describe in the following section some conventional approaches for MUD.

4.2 Conventional MIMO Detectors

In this chapter, we assume that channel estimation is error-free and focus our discussion on the MUD, which consists of two blocks: a preprocessor and a demodulator. Based on the estimated channel matrix and noise variance, the preprocessor provides the demodulator with filters and parameters required to decode users' symbols from the received signal vector. Conventional multiuser reception techniques include the following.

LMMSE reception is the optimal *linear* MUD method. It provides the best combination of zero-forcing interference suppression and matched filtering which are optimal at high and low SNR, respectively. The LMMSE beamformer estimates the vector of transmitted symbols as $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$, with the optimal receive matrix, W, which is computed by

the preprocessor based on channel state information as

$$\mathbf{W} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \sigma^2\mathbf{I})^{-1} = (\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I})^{-1}\mathbf{H}^H.$$

It should be noted that the second equality is more computationally efficient to calculate as it requires inverting a smaller matrix (since $K \leq N$). The computational complexity of LMMSE is $\mathcal{O}(\beta N^2)$ for beamforming and $\mathcal{O}(\beta^2 N^3)$ for computing $\mathbf{W}$, which does not scale well with the number of antennas. Furthermore, even though it provides near-optimal performance in low-loaded systems, its performance diminishes once the load factor, β, exceeds $\frac{1}{4}$, especially when the near-far power disparity is large or some channels are very close in spatial frequency. In these conditions, nonlinear techniques can provide significant performance gains.

Interference cancellation (IC) is the most intuitive and well-known nonlinear MUD method. After decoding a user's *digital* symbol, this receiver calculates the interference of that user on other channels and subtracts it from the original observations before a second demodulation. This can be done successively (SIC) [70], starting from the strongest user to the weakest, or in parallel (PIC) [71]. The former is advantageous in terms of performance, especially when variation in channel strength is large among users, but entails higher delay and is not easily parallelizable. An efficient implementation of SIC is the V-BLAST algorithm that admits a complexity of $\mathcal{O}(N^3)$ [72].

In the next section, we describe our proposed MUD which combines linear techniques in beamspace with nonlinear interference cancellation to provide a scalable receiver design.

4.3 Scalable Nonlinear Multiuser Detection

The beamspace local LMMSE receiver developed in our prior work [13] takes advantage of channel concentration in spatial frequency domain to perform linear estimation of each user's symbol using a small window of the beamspace signal vector, $\tilde{\mathbf{y}}$. This windowing approach significantly reduces the computational burden of the linear detector; however, the limited dimensionality of observations limits its interference suppression capability, and linear techniques, in general, are very suboptimal at high load factor or when trying to detect users that are close in spatial frequency and have highly correlated channels. Nonlinear techniques are effective in these cases, but their complexity can become a bottleneck for massive MIMO systems. To facilitate a scalable nonlinear detector, we augment the beamspace local LMMSE receiver with a user-centric *virtual MIMO system* that models the cross-interaction of nearby neighbors in beamspace. Nonlinear detection is possible on this smaller virtual system, especially as the number of significant interferers for any given user remains relatively constant as the system is scaled up in size. In this section, we describe the stages of this proposed approach and determine the computational complexity of each stage.

4.3.1 Local LMMSE

The initial local LMMSE stage (summarized in Algorithm 2) carries out a lightweight linear estimation of users' symbols by transferring the received signal vector to beamspace via an FFT operation, and then limiting the observation window used for the m^{th} user to a small number (W_1) of FFT bins around the m^{th} user's spatial frequency. This window is chosen for each user such that the resulting mean squared error (MSE) is minimized, as described in Algorithm 2. Using these limited dimensions, the local LMMSE receiver suppresses interference produced by other users via linear projection, and provides the

local LMMSE estimate,

$$\bar{x}_m = \mathbf{w}_m^H \tilde{\mathbf{y}}, \tag{4.1}$$

where $\mathbf{w}_m$ is the local LMMSE filter for the m^{th} user (obtained by Algorithm 2) which contains W_1 nonzero entries. It is worth noting that there are only W_1 nonzero complex multiplication operations in (4.1). The estimates $\{\bar{x}_k\}_{k=1}^K$ serve as observations for the next stage of processing.

4.3.2 User-centric whitened virtual MIMO

In the second stage, we create a small virtual MIMO system in order to obtain a better estimate of its symbol as follows. The virtual MIMO system for user m is obtained by taking the set of $W_2 - 1$ nearest users (in beamspace) and forming the set $\mathcal{J}_m$ of users, as shown in Fig. 4.2. The measurement vector for this system is denoted as

$$\begin{aligned} \mathbf{z}_m = [\bar{x}_k]_{k\in\mathcal{J}_m} &= \dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{n}}) \\ &= \underbrace{\dot{\mathbf{W}}_m^H \tilde{\mathbf{H}}_{\mathcal{J}_m}}_{\mathbf{B}_m} \mathbf{x}_{\mathcal{J}_m} + \underbrace{\dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}_{\mathcal{J}_m^c}\mathbf{x}_{\mathcal{J}_m^c} + \tilde{\mathbf{n}})}_{\mathbf{i}_m}, \end{aligned} \tag{4.2}$$

where $\dot{\mathbf{W}}_m = [\mathbf{w}_k]_{k\in\mathcal{J}_m}$, $\tilde{\mathbf{H}}_{\mathcal{J}_m} = [\tilde{\mathbf{h}}_\ell]_{\ell\in\mathcal{J}_m}$, $\tilde{\mathbf{H}}_{\mathcal{J}_m^c} = [\tilde{\mathbf{h}}_\ell]_{\ell\notin\mathcal{J}_m}$, $\mathbf{x}_{\mathcal{J}_m} = [x_\ell]_{\ell\in\mathcal{J}_m}$, and $\mathbf{x}_{\mathcal{J}_m^c} = [x_\ell]_{\ell\notin\mathcal{J}_m}$. We treat the interference from users that are not in $\mathcal{J}_m$ as noise, and compute the overall noise covariance matrix of (4.2) as

$$\mathbf{\Sigma}_m = \mathbb{E}[\mathbf{i}_m \mathbf{i}_m^H] = \dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}_{\mathcal{J}_m^c}\tilde{\mathbf{H}}_{\mathcal{J}_m^c}^H + \sigma^2\mathbf{I})\dot{\mathbf{W}}_m. \tag{4.3}$$

Since the FFT taps used for different users in $\mathcal{J}_m$ are likely to overlap, this distortion is "colored". We whiten each virtual MIMO system by computing

$$\bar{\mathbf{z}}_m = \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{z}_m.$$

The whitening filter can be computed efficiently using a Cholesky decomposition [55]. This yields the following model for the m^{th} whitened virtual MIMO system:

$$\bar{\mathbf{z}}_m = \mathbf{A}_m \mathbf{x}_{\mathcal{J}_m} + \mathbf{n}_m, \tag{4.4}$$

where the effective channel seen by the users in $\mathcal{J}_m$ becomes

$$\mathbf{A}_m = \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{B}_m,$$

and the "noise" (which includes interference due to users in $\mathcal{J}_m^c$) in the virtual MIMO system is white: $\mathbf{n}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

4.3.3 Nonlinear MUD for the virtual MIMO systems

We may now apply any nonlinear MUD to the W_2-dimensional virtual MIMO system centered at user m to get an estimate of its symbol, x_m. We report results for LMMSE-SIC operating on each virtual MIMO system. For simplicity, we describe it for a generic whitened virtual MIMO system of the form:

$$\bar{\mathbf{z}} = \mathbf{A}\mathbf{x} + \mathbf{n},$$

where we drop the subscripts from (4.4), and number the users from 1 to W_2.

The SIC demodulator consists of W_2 successive stages. In the first stage, it receives

Algorithm 2: Local LMMSE Preprocessing.

Input: $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_k]_{k=1}^{K} \in \mathcal{C}^{N\times K}$, σ^2, N, and m
Output: $\mathbf{w}_m$
Parameter: W_1 (window size)

1: set $\mathbf{G} = \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H + \sigma^2\mathbf{I}$ {Compute the covariance matrix}
2: **for** $i = 1$ **to** $N - W_1 + 1$ **do**
3: set $\mathbf{R}_i = [G_{\ell,n}]_{\ell=i,n=i}^{i+W_1-1,i+W_1-1}$
4: set $\dot{\mathbf{h}}_i = \left[\tilde{H}_{\ell,m}\right]_{\ell=i}^{i+W_1-1}$
5: set $MSE_i = 1 - \dot{\mathbf{h}}_i^H \mathbf{R}_i^{-1}\dot{\mathbf{h}}_i$
6: **end for**
7: set $i^* = \arg\min_i MSE_i$ {Optimal window location}
8: set $\mathbf{w}_m^H = [\mathbf{0}_{i^*-1}^\intercal, \dot{\mathbf{h}}_{i^*}^H \mathbf{R}_{i^*}^{-1}, \mathbf{0}_{N-i^*-W_1+1}^\intercal]$
9: set $\mathbf{w}_m \leftarrow \frac{\mathbf{w}_m}{\mathbf{w}_m^H \tilde{\mathbf{h}}_m}$ {Remove the estimation bias}

the observation vector $\bar{\mathbf{z}}$ and linearly projects it on $\mathbf{v}_1$ (see the description of the preprocessing in Algorithm 3) to decode user ℓ_1's symbol, i.e., $\dot{x}_{\ell_1} = \mathbf{v}_1^H \bar{\mathbf{z}}$. Then, using a constellation demapper, the demodulator retrieves the original constellation symbol $\hat{x}_{\ell_1}$ from the estimate $\dot{x}_{\ell_1}$. The SIC then subtracts its effect from the observation vector to get $\bar{\mathbf{z}}^{(1)} = \bar{\mathbf{z}} - [A_{n,\ell_1}]_{n=1}^{W_2}\hat{x}_{\ell_1}$. In the next step, the same process is applied on $\bar{\mathbf{z}}^{(1)}$ to decode user ℓ_2's symbol, and so on.

The nonlinear demodulator needs the order of users $\mathcal{L} = \{\ell_k\}_{k=1}^{W_2}$ and the projection vectors $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^{W_2}$ from the preprocessing step. As shown in algorithm 3, preprocessing starts by computing the MSE of each user's estimate (step 3), picks the user with the highest SINR (step 4) and computes its projection vector (step 5-6). It then removes that user's channel vector from the channel matrix $\mathbf{A}$ (step 8). The acquisition repeats this procedure W_2 times until it completely computes $\mathcal{L}$ and $\mathcal{V}$.

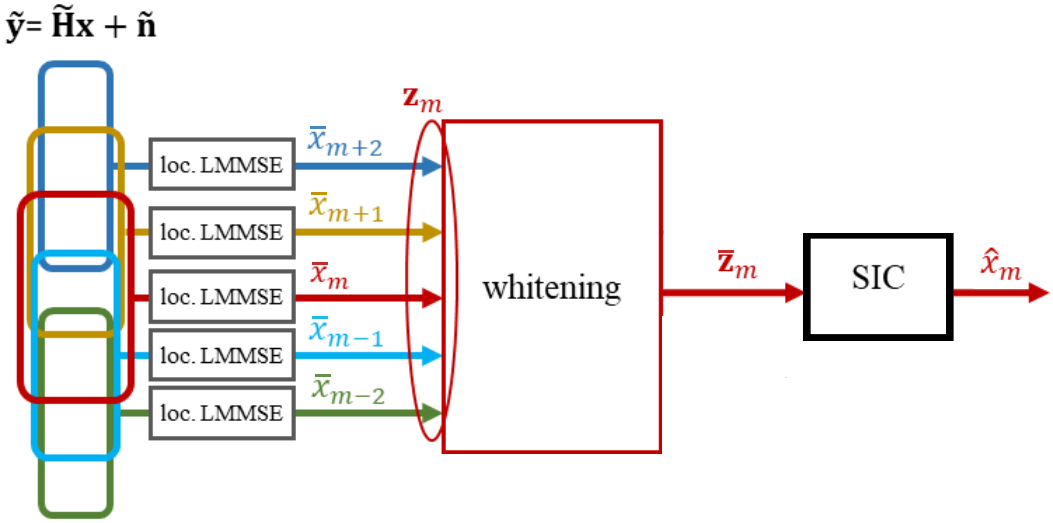


Figure 4.2: Proposed MUD scheme for one virtual MIMO system.

4.3.4 Computational complexity

Algorithm 2 describes the preprocessing stage of the local LMMSE block. The algorithm starts with computing the covariance matrix $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H+\sigma^2\mathbf{I}$ (step 1), and then searches for the location of the optimum observation window to minimize the MSE for each user (steps 2-7). The algorithm then forms the local LMMSE projection vector (steps 8-9).

The complexity of the local LMMSE beamformer is $\mathcal{O}(\beta W_1 N)$ for demodulation (performed on a per-symbol scale), and $\mathcal{O}(\beta W_1 N^2)$ for preprocessing (performed on scale of channel coherence time). The most computationally expensive part of this step is computing the Gram matrix $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H$. Initially, it has a computational complexity of $\mathcal{O}(\beta N^3)$, however, the algorithm uses the elements on the matrix diagonal band only, reducing the computational complexity to $\mathcal{O}(\beta W_1 N^2)$.

The complexity of the whitening process is $\mathcal{O}(W_2^2 \beta N)$. For preprocessing, computing the overall noise covariance matrix, described in (4.3), dominates the computational complexity. Notice that the matrix $\dot{\mathbf{W}}_m$ has only $W_1 \times W_2$ nonzero elements. Hence, computing $\dot{\mathbf{W}}_m^H \tilde{\mathbf{H}}_{\mathcal{J}_m^c}$ and the covariance matrix $\mathbf{\Sigma}_m$ incur a computational complexity

Algorithm 3: SIC Preprocessing.

Input: $\mathbf{A} \in \mathcal{C}^{W_2 \times W_2}$, and W_2
Output: $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^{W_2}$ and $\mathcal{L} = \{\ell_k\}_{k=1}^{W_2}$

set $\mathcal{M} = \{1, 2, \ldots, W_2\}$
for $i = 1$ **to** W_2 **do**
  set $\mathbf{B} = \left(\mathbf{A}^H\mathbf{A} + \mathbf{I}\right)^{-1}$
  set $n = \arg\min_k B_{kk}$ {Find the user with minimum MSE}
  set $\mathbf{v}_i^H = [B_{nk}]_{k=1}^{W_2}\, \mathbf{A}^H$
  set $\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i}{\mathbf{v}_i^H [A_{mn}]_{m=1}^{W_2}}$ {Remove the estimation bias}
  set $\ell_i = \mathcal{M}_n$
  set $[A_{mn}]_{m=1}^{W_2} = [\,]$ {Remove the n^{th} column}
  set $\mathcal{M}_n = [\,]$ {Remove the n^{th} entry}
end for

of $\mathcal{O}(W_2 W_1 \beta N)$ and $\mathcal{O}(W_2^2 \beta N)$ per virtual MIMO system, respectively. Since we have K virtual system, the total computational complexity of the whitening preprocessor becomes $\mathcal{O}(W_2(W_1 + W_2)\beta^2 N^2)$. The total computational complexity of the nonlinear MUD step is $\mathcal{O}(W_2^2 \beta N)$ for detection and $\mathcal{O}(W_2^3 \beta N)$ for preprocessing.

Thus, the overall complexity of our proposed algorithm is dominated by $\sim \mathcal{O}(N \log N)$ for demodulation (which is performed on a symbol by symbol basis) and $\sim \mathcal{O}(N^2)$ for preprocessing (which is repeated on a time scale of channel coherence time), assuming window sizes are small.

4.4 Results

We consider the MIMO system illustrated in Fig. 4.1a with a carrier frequency of 140 GHz. We select the number of antennas at the base station to be $N = 256$ according to the link budget calculation described in [5]. All numerical simulations are conducted at load factor $\beta = 1/2$, unless otherwise stated. The sector field of view is restricted to $-\pi/3 \leq \theta \leq \pi/3$ radians. The users are placed uniformly in the coverage area, at a

distance of at least 5 m and at most 100 m from the base station.

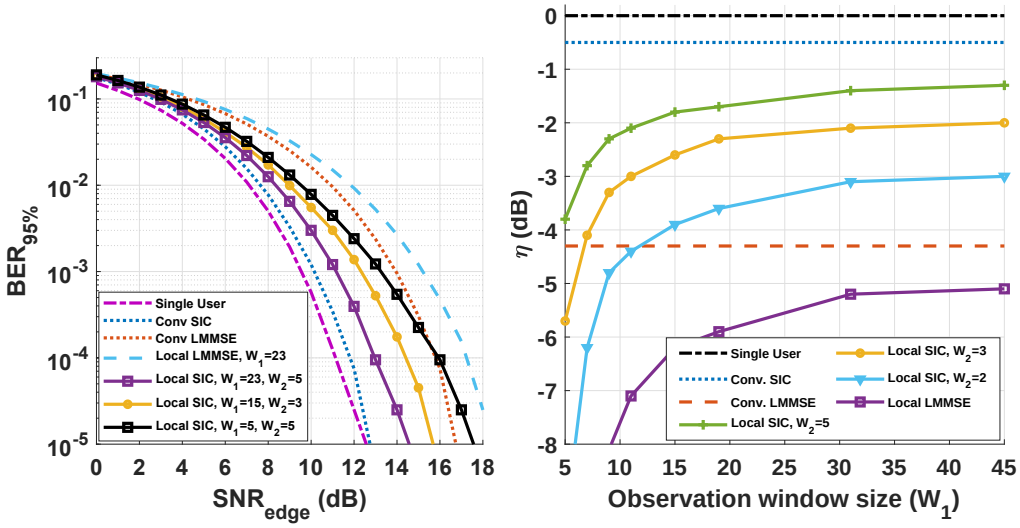
While the user terminals are placed randomly in our simulations, we enforce a minimum separation between them in spatial frequency to avoid irrecoverable excessive interference. As shown in Fig. 4.1a, the minimum spatial frequency between any two users $\Delta\Omega_{\min}$ is at least half the 3 dB beamwidth (BW_{3dB}), i.e., $\Delta\Omega_{\min} = \frac{2.783}{N}$ radians [4]. We assume that the base station can serve users that are closer than this threshold in different time or frequency resource blocks.

We do not deploy any power control scheme in our simulations, and hence the near-far effect between users prevails. We assume that the base station has perfect knowledge of the channel state information (CSI). We measure link quality by the outage probability at a target uncoded BER of 10^{-3} for QPSK. This BER requires SNR of about 10 dB for a single AWGN link, which becomes the target SINR at the output of the multiuser detector for an edge user. We use the single user scenario as a benchmark, and compare between four MUD schemes: conventional LMMSE, conventional SIC, local LMMSE, and the proposed scheme, which we refer to in figures by "Local SIC."

Performance and efficiency: Fig. 4.3 (a) depicts the bit error rate that 95% of users in the cell achieve as a function of the SNR of the edge user, which is defined as $\mathrm{SNR}_{\mathrm{edge}} = \frac{NP_{\mathrm{tx}}|\alpha_{100}|^2}{\sigma^2}$ where $|\alpha_{100}|^2$ is the free-space path loss at 100 m and P_{tx} is the transmitted power of user devices.

Fig. 4.3 (b) shows the efficiency of each MUD scheme compared to single user performance at target uncoded BER of 10^{-3}. We define the efficiency η as the ratio between the transmit power levels required for single user operation and multiuser operation (with a given MUD scheme) achieving the target BER, i.e.,

$$\eta = \frac{\mathrm{SNR}_{\mathrm{edge}}(\text{Single User})}{\mathrm{SNR}_{\mathrm{edge}}(\text{MUD})},$$



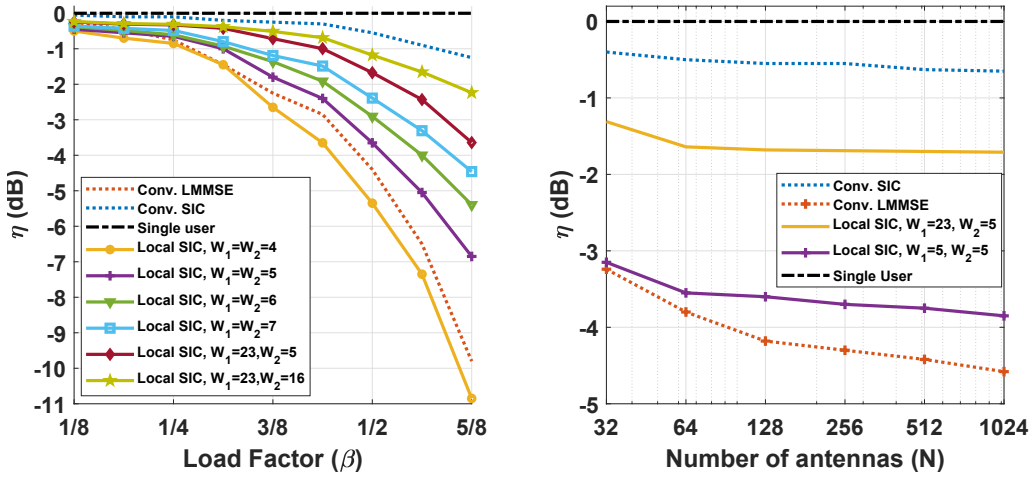
(a) $\beta = 1/2$. (b) Target BER 10^{-3}.

Figure 4.3: (a) The BER achieved by at least 95% of the users for different window sizes. (b) The efficiency of the proposed scheme relative to the conventional SIC.

where SNR_{edge}(Single User) and SNR_{edge}(MUD) are the SNR levels required for the edge user to achieve the target BER in single user and multiuser scenarios, respectively.

Scalability: Fig. 4.4 (a) depicts efficiency relative to single-user performance as a function of load factors for different MUD schemes and window sizes. The performance gap between the different MUDs and the single-user baseline increases as the load factor increases. Fig. 4.4 (b) reports these trends as a function of array size. It is clear that the efficiency of the proposed MUD is almost constant, regardless of the number of elements. Therefore, for maintaining the desired performance, window sizes do not need to be scaled with the number of antennas.

Computational complexity: We categorize the complexity of the MUD into beamforming and preprocessing complexities. Fig. 4.5 (a) and (b) demonstrate the number of complex multiplications required to carry out each MUD scheme. The FFT dominates the proposed scheme's beamforming complexity as $\mathcal{O}(N \log(N))$, whereas the preprocess-



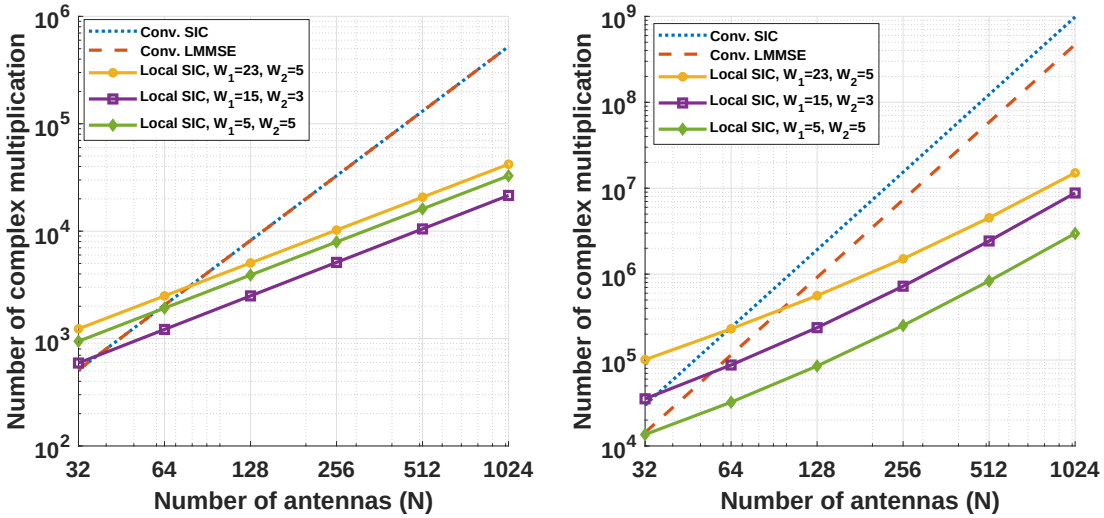
(a) MUD efficiency versus β. (b) MUD efficiency versus N.

Figure 4.4: The efficiency of different configurations of the proposed MUD versus (a) the load factor and (b) the number of antenna elements.

ing complexity is $\mathcal{O}(\beta W_1 N^2)$ which mostly pertains to computing the Gram matrix. At the cost of 1 dB performance degradation in efficiency, the proposed algorithm achieves savings in complexity by four and ten times compared to SIC in beamforming and preprocessing, respectively.

4.5 Conclusions

The proposed nonlinear multiuser detection strategy leverages the sparsity of the mmWave channel in beamspace to accomplish drastic reductions in the complexity of both computing the receiver parameters (which remain unchanged over a channel coherence time) in *preprocessing,* and of per-symbol *demodulation.* Preprocessing complexity scales quadratically instead of cubically (which is the complexity of standard linear multiuser detection) with system size. Per-symbol complexity is dominated by the spatial FFT, and is $\mathcal{O}(N \log N)$, instead of $\mathcal{O}(\beta N^2)$ as with linear multiuser detection. The performance is close to that of standard interference cancellation with an order of magnitude



(a) Beamformer complexity.

(b) Preprocessing complexity.

Figure 4.5: Complexity comparison of the proposed scheme with other MUD techniques.

lower complexity.

Chapter 5

Efficient Beamspace Downlink Precoding for mmWave Massive MIMO

We investigate linear transmit precoding for all-digital millimeter wave (mmWave) massive MIMO cellular *downlink* with a large number N of base station antennas, and with the number of simultaneously served users K scaling with N: we set $K = \beta N$, where β is termed the *load factor.* This complements the work in the earlier chapters, in which we explore the feasibility, efficacy, and challenges of the *uplink* in such a system. Specifically, we have shown in Chapters 3 and 4 that the signal processing for uplink receive beamforming could be vastly simplified with beamspace techniques that exploit the sparsity of the mmWave channel. In this chapter, we demonstrate that beamspace techniques may have an even greater impact in terms of accomplishing downlink precoding with reasonable complexity as K and N get large. The problem of linear downlink precoding involves two tasks, power allocation subject to a total budget at the base station, and beamforming for interference suppression across users. Such power allocation is not

possible on the uplink: a mobile might use power control and not use the entirety of its power budget, but it cannot transfer this power to another user. However, optimal linear downlink precoding can be mapped [73, 74] to a *virtual uplink* problem with analogous power control and beamforming steps.

Contribution: Optimal downlink precoding is typically accomplished by iterative optimization, with computational complexity scaling as $O(KN^2)$, or $O(N^3)$ in the scaling regime of interest. This is clearly infeasible for the regimes of interest to us: at mmWave frequencies, hundreds of base station antennas can be packed into compact form factors, which opens up the capability to support a correspondingly large number of simultaneous users in each base station sector using spatial multiplexing. In this chapter, we propose precoding in beamspace, exploiting the sparsity of the spatial channel from the base station to each mobile user. Under our model, the channel vector for each user in beamspace spans a few spatial frequency bins, and the optimal beamformer for a given user is well approximated over a window in beamspace whose size W does not scale with the number of base station antennas. The computational complexity of the resulting algorithm is $O(KW^2)$, which is linear in the number of users/antennas, and can therefore scale to the regimes of interest to us. Our numerical results illustrate the drastic reduction in complexity, and show that, for a computational budget which yields near-optimal performance with the proposed scheme, the performance of the standard approach to computing the precoder exhibits significantly poorer performance (e.g., 6 dB worse SINR) because of the small number of iterations that can be run within that computational budget.

Related Work: The transmit precoding problem can be posed as minimizing the total transmitted power at the base station, subject to each user attaining a desired SINR. The duality between this problem and that of receive beamforming problem was pointed out in [73, 74], and used to provide an iterative algorithm that converges to the optimal

solution, assuming that a feasible solution exists. Discussion of feasibility within this duality framework was included in [75]

An alternative formulation of transmit precoding is to maximize the minimum SINR across users. In this form, the problem is always feasible, and can be solved by considering fixed point iterations for normalized transmit beamforming vectors and power allocations [1]. This is the approach adopted in this chapter as we seek to exploit spatial channel sparsity in beamspace.

It is worth noting that the connections between various forms of the transmit precoding problem are discussed in [76], where the authors also provide fast algorithms to approach local optima which are globally optimum under sufficiently weak interference.

Notation: We use lowercase bold letters for vectors, and uppercase bold letters for matrices. The notation $\mathbf{x} = [x_i]_{i=0}^{I}$ represents column vector $\mathbf{x}$ of length I and its elements are denoted by x_i. For a matrix, we use $\mathbf{X} = [x_{i,j}]_{i=0,j=0}^{I,J}$. If the size of the vector or the matrix can be inferred from the context, we write $\mathbf{X} = [x_{i,j}]_{i,j}$ for simplicity. $\{.\}_{k=1}^{K}$ denotes a list of K scalars, vectors or matrices.

5.1 The Downlink Precoding Problem

Consider the downlink system depicted in Fig. 5.1. The base station employs a linear array with N elements to simultaneously serve $K = \beta N$ mobile users. We assume that each mobile can perform ideal receive beamforming towards the base station, and include the gain due to such spatial matched filtering into the spatial channel $\mathbf{h}_k$ from the base station to mobile k, $k = 1, ..., K$.

Linear Precoding: The linear precoder at the base station allocates power p_k to mobile k, and employs beamforming direction $\{\bar{\mathbf{w}}_k\}$ (normalized to unit norm), so that the

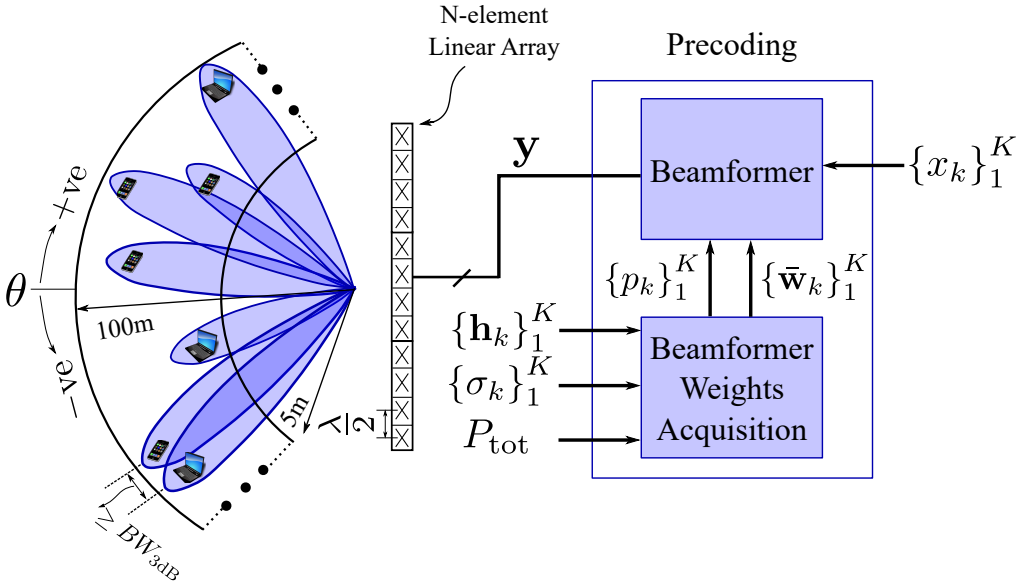


Figure 5.1: Downlink massive MIMO system model.

transmitted signal is given by

$$\mathbf{y} = \sum_{i=1}^{K} \bar{\mathbf{w}}_i \sqrt{p_i} x_i, \tag{5.1}$$

where x_k is the k^{th} user symbol. Thence, the k^{th} user's equipment receives

$$z_k = \mathbf{h}_k^H \bar{\mathbf{w}}_k \sqrt{p_k} x_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{h}_k^H \bar{\mathbf{w}}_i \sqrt{p_i} x_i + n_k, \tag{5.2}$$

where n_k is additive white Gaussian noise (AWGN) with variance σ_k^2.

In Fig. 5.1, the weights acquisition block computes the power allocation and beamforming directions, given the mobile users' channel vectors, $\{\mathbf{h}_k\}$, and receiver noise variances, $\{\sigma_k^2\}$, along with the total power budget, P_{tot}. The beamformer block performs the actual precoding (5.1) using the computed weights.

SINR: The signal-to-interference-plus-noise ratio (SINR) of the k^{th} user is given by

$$SINR_k = \frac{|\mathbf{h}_k^H \bar{\mathbf{w}}_k|^2 p_k}{\sigma_k^2 + \sum\limits_{\substack{i=1 \\ i \neq k}}^{K} |\mathbf{h}_k^H \bar{\mathbf{w}}_i|^2 p_i}. \tag{5.3}$$

The SINR is a widely used performance measure because, under a Gaussian approximation for the interference-plus-noise, it provides an excellent approximation for the bit error rate (BER) (e.g., see [77] for the closely related problem of uplink multiuser detection), as well as for the achievable data rate.

Channel Model: We assume that the channel between the base station and any mobile is dominated by a single path, so that, for a linear array, the $N \times 1$ spatial channel for the k^{th} mobile is given by

$$\mathbf{h}_k = A_k \, [1 \, e^{j\Omega_k} \, e^{j2\Omega_k} \, \ldots \, e^{j(N-1)\Omega_k}]^\intercal, \tag{5.4}$$

where Ω_k is the spatial frequency and $|A_k|$ the channel amplitude for the path.

Such a model is well suited for mmWave channels for several reasons:

- Typical surfaces (e.g., roads, concrete walls) look rougher at small carrier wavelengths. Hence a significant portion of the energy from a reflection is scattered. Thus, mmWave channels are typically comprised of a small number of dominant paths.

- The relative delay between different paths is large (relative to the symbol interval) for the large signaling bandwidths at mmWave bands. Gathering the energy across a large number of symbols using an appropriately designed space-time filter is computationally complex. Hence a reasonable design is to focus spatial beams along a single dominant path.

- For a large antenna array, beamforming along a given path significantly attenuates other paths, so that they can be safely neglected post-beamforming.

5.1.1 Problem Formulation

We consider here the max-min fair formulation of the precoding optimization problem. Thus, precoding weights acquisition block calculates the beamforming directions, $\{\bar{\mathbf{w}}_k\}$, and the power allocation, $\{p_k\}$, by solving the following optimization problem:

$$\gamma_o = \underset{\bar{\mathbf{w}}_k, p_k \forall k}{\text{maximize}} \quad \min_k \; SINR_k \tag{5.5a}$$

$$\text{subject to} \quad \sum_{i=1}^{K} p_i \leq P_{\text{tot}}, \tag{5.5b}$$

$$\|\bar{\mathbf{w}}_k\|_2 = 1 \;\; \forall k, \tag{5.5c}$$

$$p_k \geq 0 \;\; \forall k. \tag{5.5d}$$

This problem can be cast as a generalized eigenvalue problem and is always feasible [1].

After defining suitable Lagrange multipliers λ_k, the optimality conditions for problem (5.5) can be formulated as follows,

$$\mathbf{h}_k^H \left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{\sigma_i^2} \mathbf{h}_i \mathbf{h}_i^H \right)^{-1} \mathbf{h}_k \frac{\lambda_k}{\sigma_k^2} = \frac{\gamma_o}{1+\gamma_o} \;\; \forall k, \tag{5.6}$$

$$\sum_{i=1}^{K} \lambda_i = P_{\text{tot}}, \tag{5.7}$$

$$\lambda_k \geq 0 \;\; \forall k. \tag{5.8}$$

As a consequence, the beamforming directions can be written as follows,

$$\bar{\mathbf{w}}_k = \frac{\left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{\sigma_i^2} \mathbf{h}_i \mathbf{h}_i^H\right)^{-1} \mathbf{h}_k}{\left\| \left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{\sigma_i^2} \mathbf{h}_i \mathbf{h}_i^H\right)^{-1} \mathbf{h}_k \right\|_2}, \tag{5.9}$$

and the power vector, $\mathbf{p} = [p_1, \ldots, p_K]^\intercal$, can be evaluated by solving the following system of linear equations,

$$\left(\frac{1+\gamma_o}{\gamma_o} \mathbf{I} - \left[\frac{|\mathbf{h}_i^H \bar{\mathbf{w}}_j|^2}{|\mathbf{h}_i^H \bar{\mathbf{w}}_i|^2} \right]_{i=1,j=1}^{K,K} \right) \mathbf{p} = \left[\frac{\sigma_i^2}{|\mathbf{h}_i^H \bar{\mathbf{w}}_i|^2} \right]_{i=1}^{K}. \tag{5.10}$$

It is evident that the Lagrange multipliers, λ_k, play a critical role in solving the optimization problem posed in (5.5). Hence, all solution approaches revolve around finding optimal (or sub-optimal) values of the Lagrange multipliers λ_k.

5.1.2 Fixed Point Iterations for Optimal Precoding

We review the method proposed in [1] for tackling the optimization problem (5.5). This provides a benchmark for optimal precoding for general channel models, as well as a basis for our proposed beamspace approach tailored to sparse channels. The optimality condition (5.6) can be rewritten as follows:

$$\lambda_k = \frac{\gamma_o}{1+\gamma_o} \frac{\sigma_k^2}{\mathbf{h}_k^H \left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{\sigma_i^2} \mathbf{h}_i \mathbf{h}_i^H\right)^{-1} \mathbf{h}_k} \ \forall k, \tag{5.11}$$

which motivates using a fixed-point iteration method to find the optimal Lagrange multipliers, $\{\lambda_k\}$. The scaling of the fixed point depends on γ_o, which is the max-min SINR solution to the optimization problem, and is therefore unknown. Thus, fixed point iterations are interleaved with a scaling step based on the total power constraint (5.7). The

Algorithm 4: Fixed point iteration to find optimal λ_k [1]

Input: $\{\mathbf{h}_k\}$, $\{\sigma_k^2\}$, and P_{tot}
Output: $\{\lambda_k\}$

1: initialize $\lambda_k = P_{\text{tot}}/K$
2: **repeat**
3: set $\mathbf{B} = \left(\mathbf{I} + \sum_i \mathbf{h}_i \mathbf{h}_i^H \lambda_i/\sigma_i^2\right)$ (III)
4: set $\mathbf{G} = \mathbf{B}^{-1}$ (IV)
5: set $q_k = \mathbf{h}_k^H \mathbf{G} \mathbf{h}_k \lambda_k/\sigma_k^2$ (V)
6: set $\bar{\lambda}_k = \lambda_k/q_k$
7: set $\lambda_k = P_{\text{tot}} \bar{\lambda}_k / \sum_i \bar{\lambda}_i$
8: **until** q_k are all equal $\forall k$.

resulting algorithm, whose convergence is proved in [1], is summarized as Algorithm 4: one fixed point iteration (steps 5 and 6) is followed by imposing the total power constraint (step 7), repeated until convergence to within some tolerance of the optimality condition (5.6).

Most prior evaluations of optimal precoding focus on a relatively small number of antennas. As we increase the number of antennas, the computational complexity for attaining convergence becomes excessive. In order to compare our low-complexity beamspace technique with the state of the art, we consider terminating Algorithm 4 after a fixed number of iterations based on a computational budget. The resulting Lagrange multipliers are suboptimal, and the optimality condition (5.6) is not necessarily satisfied. We can still compute the normalized beamforming directions (5.9) using these suboptimal Lagrange multipliers, but the power allocation (5.10) cannot be used, since we do not know γ_o. Instead, we fix the suboptimal beamforming directions $\bar{\mathbf{w}}_k$, and solve an optimal

power allocation problem as follows to obtain a benchmark for comparison:

$$\underset{p_k \forall k}{\text{maximize}} \quad \min_k \ \frac{|\mathbf{h}_k^H \bar{\mathbf{w}}_k|^2 p_k}{\sigma_k^2 + \sum\limits_{\substack{i=1 \\ i \neq k}}^{K} |\mathbf{h}_k^H \bar{\mathbf{w}}_i|^2 p_i} \tag{5.12a}$$

$$\text{subject to} \quad \sum_{i=1}^{K} p_i \leq P_{\text{tot}}, \tag{5.12b}$$

$$p_k \geq 0 \ \ \forall k. \tag{5.12c}$$

Once again, the optimization problem (5.12) is always feasible and admits a fixed point solution that satisfies

$$\tilde{\mathbf{p}} = \left(\left[\frac{|\mathbf{h}_i^H \bar{\mathbf{w}}_j|^2}{|\mathbf{h}_i^H \bar{\mathbf{w}}_i|^2} \right]_{i=1,j=1}^{K,K} - \mathbf{I} \right) \mathbf{p} + \left[\frac{\sigma_i^2}{|\mathbf{h}_i^H \bar{\mathbf{w}}_i|^2} \right]_{i=1}^{K}, \tag{5.13}$$

$$p_k = \tilde{p}_k \frac{P_{\text{tot}}}{\sum_i \tilde{p}_i}. \tag{5.14}$$

Computational Complexity: The complexity of Algorithm 4 is dominated by the steps labeled (III), (IV), and (V). The computational complexity *per iteration* for these steps is calculated as follows.

- (III): The complexity of computing matrix $\mathbf{B} \in \mathcal{C}^{N \times N}$ is $\mathcal{O}(KN^2)$.
- (IV): The matrix inversion can be carried out efficiently using Cholesky decomposition [55], whose complexity is $\mathcal{O}(N^3)$.
- (V): The complexity of this step is $\mathcal{O}(KN^2)$.

5.2 Proposed Beamspace Solution

We define the beamspace representation of the channel matrix as

$$\bar{\mathbf{H}} = [\mathcal{DFT}(\mathbf{h}_1), \ldots, \mathcal{DFT}(\mathbf{h}_K)]$$

where $\mathcal{DFT}(.)$ is the discrete Fourier transform (DFT) operator. We plot the magnitude of $\bar{\mathbf{H}}$ in Fig. 5.2, which makes evident the sparsity of single-path channels in beamspace. As shown in our prior work [13], for such channel models, operating in beamspace can significantly reduce the complexity of uplink multiuser detection. Given downlink-uplink duality and the iterative nature of optimization for downlink precoding, we expect even greater savings in complexity in our present setting.

We describe the proposed beamspace optimization algorithm, depicted in Algorithm 5, as follows. We assume here that we have access to estimates of the $N \times 1$ channel vectors, $\{\mathbf{h}_k\}$, and hence account for the complexity of taking DFT to go to beamspace. This process could potentially be avoided by use of channel estimation techniques that utilize beamspace techniques up front (e.g., the use of reciprocity, and uplink techniques such as those in [13]).

1) **Computing the DFT of the channel vectors:** The DFT is used to transform each channel vector $\mathbf{h}_k$ from the antenna domain to the beam domain to get $\bar{\mathbf{h}}_k$ evaluated as follows,

$$\bar{h}_{ki} = \sum_{n=1}^{N} h_{kn} e^{-j2\pi(n-1)(i-1)/N}. \tag{5.15}$$

Using the fast Fourier transform (FFT) algorithm [56], the complexity of this step becomes $\mathcal{O}(KN\log(N))$.

2) **Energy detection:** The energy distribution of the channel vector in beamspace is concentrated around its spatial frequency. Because we do not know the spatial frequency beforehand, we search for a window of size W that contains most of the channel energy. The use of a sliding window for this purpose incurs $\mathcal{O}(N)$ complexity per user.

For a given user, after finding the window that holds most of its channel energy, it is convenient to define two "synthetic" channels in beamspace: a truncated $W \times 1$ channel $\tilde{\mathbf{h}}_k$ centered on the chosen window for user k, and an approximated $N \times 1$ channel $\hat{\mathbf{h}}_k$ obtained by filling in zeros around the window.

3) **Computing Lagrange multipliers:** We use steps similar to Algorithm 4 to calculate Lagrange multipliers, but with a drastic reduction of complexity by using synthetic channels in beamspace.

• We use the approximated channel vectors, each containing only W nonzero elements, to compute the matrix $\mathbf{B}$. As a consequence, the complexity of this step decreases to $\mathcal{O}(KW^2)$ instead of $\mathcal{O}(KN^2)$ per iteration, where $W \ll N$.

• In step (V) of Algorithm 4, we replace the original channel vector with the approximated ones. For each user, only the inverse of a small $W \times W$ block inside $\mathbf{B}$, denoted by $\mathbf{G}_k$, needs to be computed in step (IV): compare step (IV) in Algorithm 4, where we invert the entire matrix $\mathbf{B}$, with that in Algorithm 5, where we invert K blocks of size $W \times W$. Thus, the complexity of step (IV) is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(KW^3)$.

• Finally, the complexity of step (V) is automatically reduced from $\mathcal{O}(KN^2)$ to $\mathcal{O}(KW^2)$.

5.3 Results

We consider the system depicted in Fig. 5.1, with number of antennas fixed at $N = 256$. The field of view for the sector is restricted to $-\pi/3 \leq \theta \leq \pi/3$. The users are

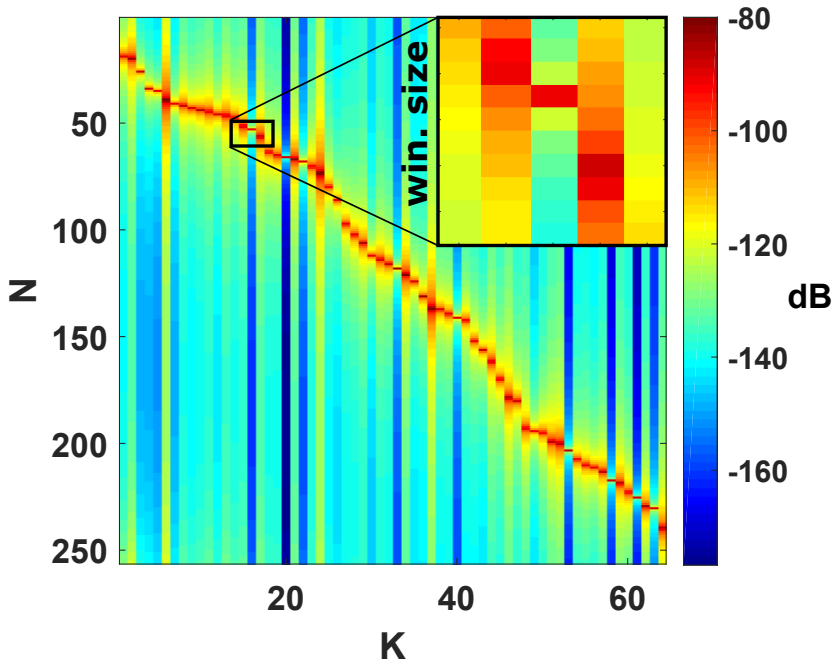


Figure 5.2: Sparsity of single-path channel in beamspace.

uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, $R_{\min} = 5$ m and $R_{\max} = 100$ m, respectively. While the user terminals are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two users in order not to incur excessive interference, arbitrarily choosing it as half the 3 dB beamwidth: $\Delta\Omega_{\min} = \frac{2.783}{N}$ [4]. BW_{3dB} in Fig. 5.1 stands for the 3 dB beamwidth. We assume that users with similar spatial frequency can be served in different time or frequency resource blocks.

We measure link quality by the outage probability at a target uncoded BER of 10^{-3} for QPSK, which corresponds to a target SINR of 9.8 dB for each downlink user.

We define the SNR_{edge} as the SNR that would be attained by a single user at the cell edge (100 m away from the base station) if the entire power budget of the base station were directed at that user. For free space propagation and ideal beamforming at both

Algorithm 5: Proposed beamspace approach to find near-optimal λ_k

Input: $\{\mathbf{h}_k\}$, $\{\sigma_k^2\}$, W and P_{tot}
Output: $\{\lambda_k\}$

set $\bar{\mathbf{h}}_k = \mathcal{FFT}(\mathbf{h}_k)$ (I)
set $\ell_k = \arg\max_{\ell} \sum_{i=\ell}^{\ell+W-1} |\bar{h}_{ki}|^2$ (II)
set $\tilde{\mathbf{h}}_k = [\bar{h}_{ki}]_{i=\ell_k}^{\ell_k+W-1}$
set $\hat{\mathbf{h}}_k = \left[\mathbf{0}_{1\times(\ell_k-1)}\ \tilde{\mathbf{h}}_k^{\intercal}\ \mathbf{0}_{1\times(N-\ell_k-W+1)}\right]^{\intercal}$
initialize $\lambda_k = P_{\text{tot}}/K$
repeat
  set $\mathbf{B} = [b_{ij}]_{i,j} = \left(\mathbf{I} + \sum_i \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^H \lambda_i/\sigma_i^2\right)$ (III)
  set $\mathbf{G}_k = \left([b_{ij}]_{i=\ell_k,j=\ell_k}^{\ell_k+W-1,\ell_k+W-1}\right)^{-1}$ (IV)
  set $q_k = \tilde{\mathbf{h}}_k^H \mathbf{G}_k \tilde{\mathbf{h}}_k \lambda_k/\sigma_k^2$ (V)
  set $\bar{\lambda}_k = \lambda_k/q_k$
  set $\lambda_k = P_{\text{tot}}\bar{\lambda}_k/\sum_i \bar{\lambda}_i$
until q_k are all equal $\forall k$.

ends, we have

$$SNR_{\text{edge}} = \frac{NMG_{\text{t}}G_{\text{r}}}{L_{100m}\sigma^2}P_{\text{tot}}, \tag{5.16}$$

where L_{100m} is the free space path loss incurred at 100 m away from the base station, M is the number of elements in the mobile's array, σ^2 is the noise variance in the mobile (which is identical in all mobiles), and G_{t} and G_{r} are the transmit and receive element gain, respectively.

Precoding Efficiency: Fig. 5.3 (a) shows the 5^{th} percentile of the minimum SINR across different channel realization, namely $SINR_{\min}$, versus the power budget represented in SNR_{edge}. That is, $SINR_{\min}$ is defined such that $\mathbb{P}(\min(SINR) \leq SINR_{\min}) = 5\%$.

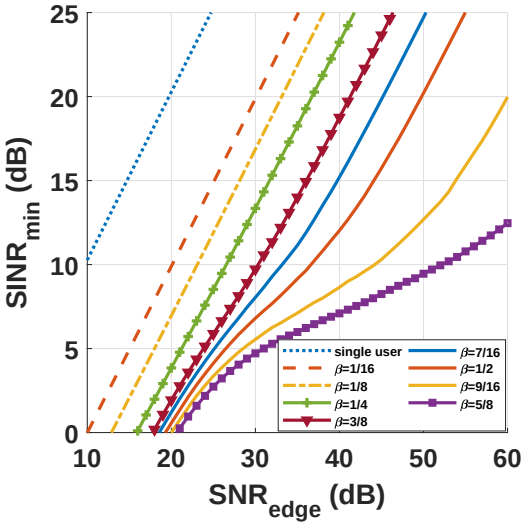
Assuming no interference between the users, if the base station power budget is allo-

cated equally between K edge users, then each user would attain an SINR of SNR_{edge}/K. Using this as the benchmark against which we compare the minimum SINR attained by our precoding scheme, the precoding efficiency η is defined as

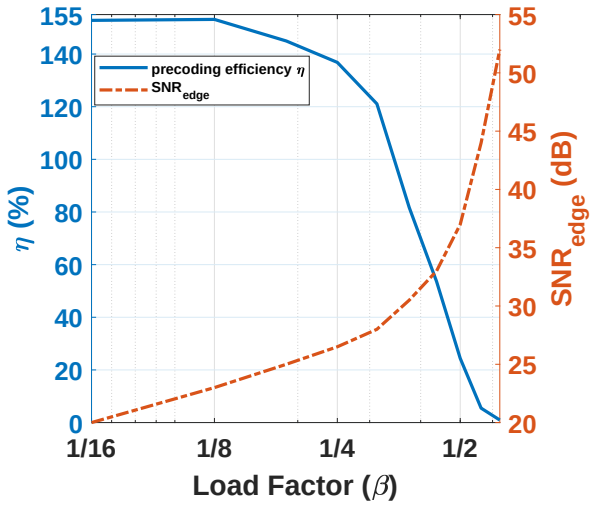
$$\eta = \frac{SINR_{\min}}{SNR_{\text{edge}}/K}. \tag{5.17}$$

As shown in Fig. 5.3 (b), the efficiency can exceed 100% at low load factor β, since the base station can transfer power from nearby users to edge users to enhance the minimum SINR, and the noise enhancement due to interference suppression on the virtual uplink is small. As the load factor increases, the loss in SINR due to interference suppression becomes more significant, and efficiency drops below 100%.

Feasibility of Target SINR: We evaluate this using the same system settings as in our prior work on uplink design [5]: $M = 16$, $G_t = G_r = 3$ dBi, $L_{100m} = 115$ dB and $\sigma^2 = -70$ dBm. For a given SNR_{edge}, the resulting link budget requires a total transmitted power of $P_{\text{tot}} = SNR_{\text{edge}}(\text{dB}) + 3$ dBm. The required emitted power for the power amplifier (PA) driving each antenna is a factor of N smaller, or 24 dB smaller for $N = 256$, and is therefore given by $P_{\text{PA}} = SNR_{\text{edge}}(\text{dB}) - 21$ dBm. The required SNR_{edge} corresponding to attaining the target SINR of 9.8 dB with 5% outage is obtained by simulations and shown in Fig. 5.3 (b). For $\beta = 1/2$, $SNR_{\text{edge}} = 37$ dB, corresponding to $P_{\text{tot}} = 40$ dBm and $P_{\text{PA}} = 16$ dBm. Such a PA specification is difficult to obtain with low-cost CMOS technologies (CMOS PA designs of up to 11 dBm have been reported in [78]), and may require more expensive alternatives such as InP technology [79]. On the other hand, if we reduce the load factor to $\beta = 1/4$, we obtain $P_{\text{tot}} = 30$ dBm and $P_{\text{PA}} = 6$ dBm, which can be comfortably attained in CMOS.



(a) The 5$^{\text{th}}$ percentile of the minimum SINR.



(b) Feasibility and Efficiency.

Figure 5.3: (a) The solution to the optimization problem (5.5) for different power budgets and system load factors. (b) The power budget required to achieve minimum SINR of ~ 10 dB along with the precoding efficiency at various system load factors.

Complexity and Performance: Table 5.1 lists the computational complexity, in terms of number of multiplication and addition operations, of the computationally expensive steps, labeled by Roman numerals, in algorithms 4 and 5. The table clearly brings out the big savings in complexity due to the proposed beamspace algorithm. Of course, the proposed algorithm incurs the additional cost of going to beamspace (steps I and II). However, these steps are required only once per channel realization, whereas the other steps (III, IV, V) are invoked on every iteration. Furthermore, as noted earlier, we may be able to fold steps I and II into channel estimation algorithms operating in beamspace.

Fig. 5.4 (a) depicts, for different load factors, the multiplication operations count for both the conventional and the proposed algorithm to achieve the same performance versus the number of elements in the base station array. It is evident that the difference in complexity is at least one order of magnitude, even for a relatively small $N = 16$.

Fig. 5.4 (b) illustrates the performance gap between the conventional and the proposed

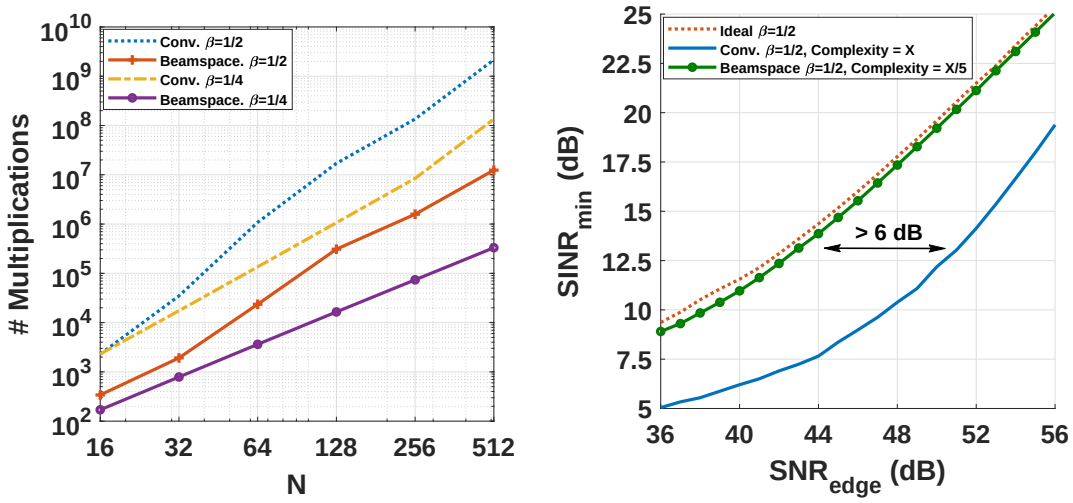
Step	# Multiplications		# Additions	
	Conventional	**Beamspace**	**Conventional**	**Beamspace**
I	0	$\frac{KN}{2}(log_2(N)-1)$	0	$KNlog_2(N)$
II	0	KN	0	$2KN$
III	KN^2J	KW^2J	KN^2J	KW^2J
IV	$\frac{N^3}{2}J$	$K\frac{W^3}{2}J$	$\frac{N^3}{2}J$	$K\frac{W^3}{2}J$
V	KN^2J	KW^2J	KN^2J	KW^2J

Table 5.1: The approximate number of multiplications and additions in the conventional [1] and the proposed beamspace algorithm to find nearly-optimal values of Lagrange multipliers λ_k. W and J denote the window size and the number of iterations.

algorithm if the computational budget is limited to that of a single iteration of the conventional algorithm. As shown, the beamspace algorithm achieves higher SINR (by 6 dB) while using only one-fifth of hardware resources.

5.4 Conclusion

We have demonstrated the drastic complexity reduction in computing optimal downlink linear precoding weights via beamspace techniques exploiting spatial channel sparsity. Conventional iterative techniques, which are required for general channel models, require a complexity per iteration which is cubic in the number of antennas, while the proposed beamspace algorithm requires linear complexity per iteration. Coupled with the work in Chapters 3 and 4 showing the efficacy of beamspace techniques for uplink multiuser detection, it is clear that beamspace techniques are a powerful tool for supporting truly massive MIMO in the mmWave and THz bands, since they are naturally matched to the channel sparsity characteristic of these bands.



(a) Computational Complexity. (b) Performance.

Figure 5.4: (a) Comparison of the number of multiplication operations in the conventional and the beamspace algorithm as the number of elements in base station increases. (b) The beamspace algorithm needs less than one-fourth of the budget power to achieve the same minimum SINR.

Chapter 6

An Efficient Digital Backend for Wideband Single-Carrier mmWave Massive MIMO

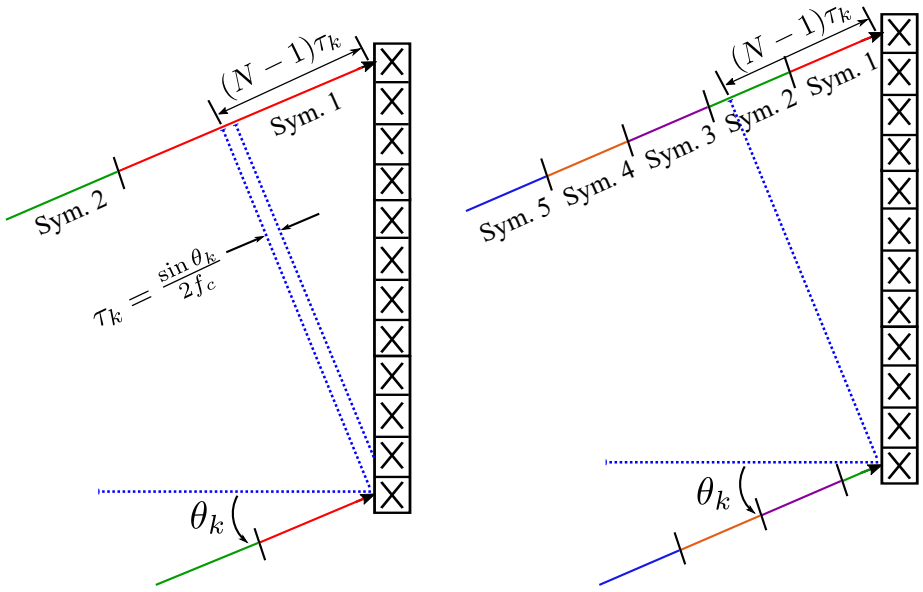
In this chapter, we consider the design of all-digital mmWave massive MIMO systems that take advantage of the massive available bandwidth. The fundamental bottleneck that we address is that, as we increase both the number of antenna elements and the bandwidth, the "narrowband assumption" for modeling the array response for a given user no longer applies, and the spatial channel for the user varies across the band.

Specifically, the worst-case delay spread for a user across the array, normalized to the symbol period, is given by

$$\tau_D = BW\tau_{\max} \tag{6.1}$$

$$= \frac{BW}{2f_c} \times (N-1) \times \sin\theta_{\max}, \tag{6.2}$$

where BW denotes bandwidth (and hence symbol rate, ignoring excess bandwidth), f_c



(a) Narrowband scenario. (b) Wideband scenario.

Figure 6.1: (a) Narrowband assumption holds, (b) Wideband modeling is required.

is the carrier frequency. The nominal system parameters we consider to illustrate our ideas are 140 GHz carrier frequency, 14 GHz bandwidth (10% of the carrier frequency), and $N = 256$, which yields a worst-case delay spread of about 11 symbols. On the other hand, if we reduced the bandwidth to as low as 200 MHz, the delay spread becomes 16% of a symbol, and the narrowband assumption is a good approximation. Fig. 6.1 provides a geometric illustration on when the narrowband assumption holds, and when wideband modeling must be used.

If we employ a standard multiuser detection strategy such as linear minimum mean square error (LMMSE) for spatial interference suppression based on nominal array responses at the center of the band, namely *narrowband LMMSE*, we can expect good performance in the scenario of Fig. 6.1 (a), and poor performance in that of Fig. 6.1 (b). This is illustrated by Fig. 6.2, which plots the BER attained with 5% outage in a picocell using narrowband LMMSE as a function of bandwidth. The performance is adequate for a small bandwidth of 200 MHz (worst-case delay spread of about 0.16 symbol), but

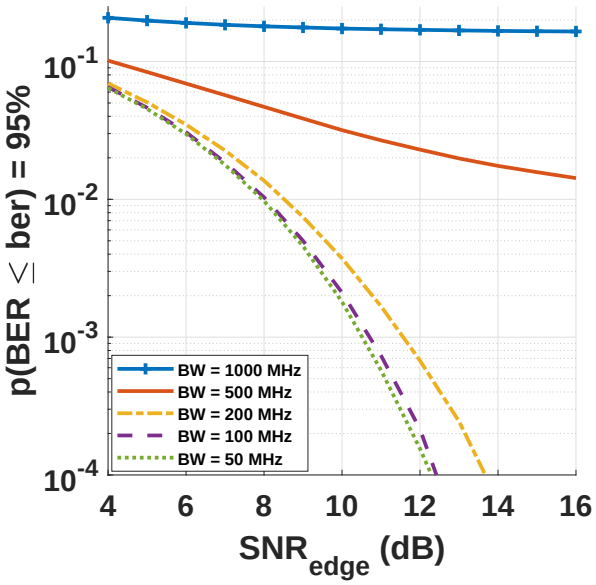


Figure 6.2: The figure shows the BER attained by 95% of the users if the narrowband LMMSE is used versus the SNR of the edge user at 100 m. The carrier frequency is 140 GHz and the base station is equipped with 256-element linear array.

deteriorates drastically by the time the bandwidth is increased to 1 GHz (worst-case delay spread of 0.8 symbol).

A natural approach to address this problem is to employ MIMO-OFDM, a common strategy for lower carrier frequencies. For a large enough number of subcarriers, the narrowband assumption applies to each subcarrier, and we can perform per-subcarrier multiuser detection. While this has also been proposed for mmWave systems in a number of papers [80, 81, 7, 82], there are two drawbacks to this approach. The first is the high Peak-to-Average Power Ratio (PAPR) of OFDM, which impairs power amplifier efficiency, already low in mm-wave systems with carrier frequencies above 100GHz. The second is that, since the spatial channels for the users are different over different subcarriers, one must employ a different spatial multiuser detection receiver for each subcarrier. This is potentially wasteful of computation, since it does not take advantage of the sparsity of the mmWave channel. In this chapter, therefore, we consider single-carrier

modulation, and exploit channel sparsity for the design of an efficient digital backend for wideband massive MIMO.

Related Work: For wideband single-carrier multiuser systems, there is general agreement in the literature on the strategy of dividing the spectrum of the received signal into smaller chunks and performing multiuser detection on each chunk separately, but there are variations in detail. In [83], the authors investigate a single-carrier frequency-domain equalization (SC-FDE) solution that employs LMMSE for MIMO processing. The authors of [84] combine SC-FDE with a time-domain decision feedback equalizer, along with interference cancellation. Since there is no channel structure assumed in [83] and [84] (the channel between each transmit and receive antenna element is modeled as multi-tap Rayleigh-fading), the LMMSE weights acquisition is complex.

In [17], the authors study the use of long arrays in LoS single-input multiple-output (SIMO) systems. They illustrate coupled signal dispersion in time and spatial frequency in the channel model, and design space-time receiver processing only over the dominant beams to reduce complexity. However, the problem of multiuser interference is not considered in [17]. In [18], the authors exploit the sparsity of the channel in both angular and delay domains to come up with channel estimation techniques which require less training overhead and have no pilot contamination. However, the paper does not address efficient multiuser detection.

Contributions: In this work, we address the problem of efficient multiuser detection in single-carrier wideband massive MIMO. Our benchmark is a signal-carrier frequency domain strategy, which may be interpreted simply as moving the inverse FFT in an OFDM transmitter to the receiver. Our main contribution is an alternative approach that exploits the sparsity of the received signal in the beamspace-frequency domain, which results

from the sparsity of the spatial channel. Intuitively, transmitting over a wide band to a long array amounts to transmitting from a spread of spatial frequencies. In our proposed algorithm, we correct this spread by mapping it to fewer spatial frequencies. While the mapping does not completely remove the dependence of the channels on frequency, it dramatically reduces the variations, which in turn reduces the number of weights to be acquired for multiuser detection (e.g., by 16-fold compared to the benchmark approach for the parameters considered here).

Notation: We use lowercase bold letters for vectors, and uppercase bold letters for matrices. The notation $\mathbf{x} = [x_i]_{i=0}^{I}$ represents column vector $\mathbf{x}$ of length I and its elements are denoted by x_i. For a matrix, we use $\mathbf{X} = [x_{i,j}]_{i=0,j=0}^{I,J}$. If the size of the vector or the matrix can be inferred from the context, we write $\mathbf{X} = [x_{i,j}]_{i,j}$ for simplicity. $\{.\}_{k=1}^{K}$ denotes a list of K scalars, vectors or matrices.

6.1 System Model

The massive MIMO system comprises one base station and K single-antenna user terminals, as depicted in Fig. 6.3. The linear multiuser detector comprises two main blocks, the beamformer weights acquisition and the beamformer. The weights acquisition takes the spatial frequency of each mobile $\Omega_k = \pi \sin\theta_k$ and the noise variance σ^2 as inputs and generate the beamformer weights $\{\mathbf{w}\}_{i=1}^{K}$. After that, the beamformer uses these weights to estimate the mobiles' data vector $\mathbf{x}(t)$ out of the received vector $\mathbf{y}(t)$.

The time-domain received signal at the n^{th} antenna element in the base station, $y_n(t)$, can be expressed as

$$y_n(t) = \sum_{k=0}^{K-1} A_k x_k(t - n\tau_k) e^{-j2\pi n f_c \tau_k} + n_n(t), \tag{6.3}$$

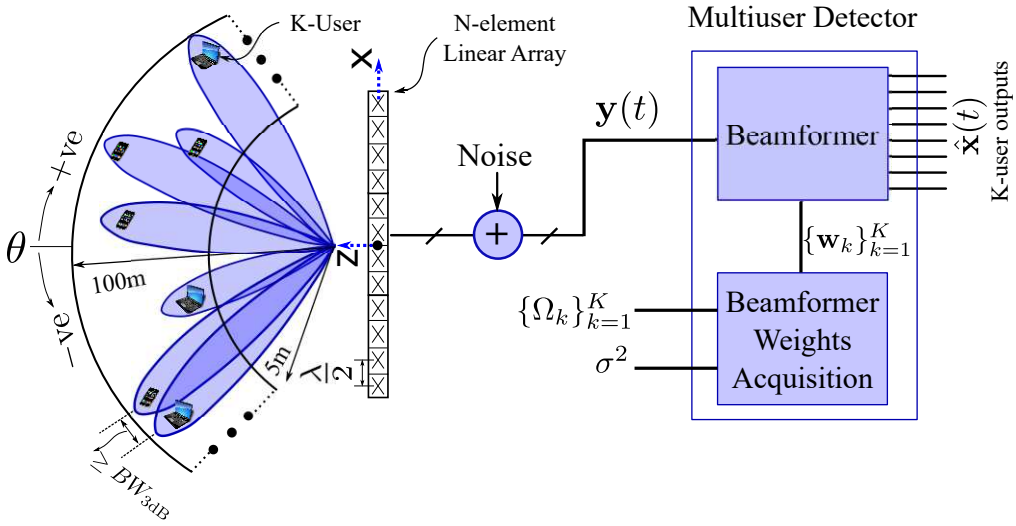


Figure 6.3: The cell size is constrained radially between 5 m and 100 m, and angularly between $-\pi/3 \leq \theta \leq \pi/3$. BW_{3dB} and $\hat{\mathbf{x}}(t)$ stand for the 3dB beamwidth the estimated data symbols vector. λ denotes the carrier's wavelength.

where $x_k(t)$ is a unit-variance time-domain symbols stream of the k^{th} user, $n_n(t)$ is band-limited complex white Gaussian noise process with variance σ^2, and $A_k^2 = \lambda^2/(4\pi R_k)^2$ depends on the radial location R_k of mobile k, using the Friis formula for path loss. f_c and λ are the carrier frequency and wavelength, and τ_k is the delay experienced by the k^{th} user's symbols stream between successive antenna elements. The delay $\tau_k = (\sin\theta_k)/(2f_c)$, $n \in [0, N)$, and $k \in [0, K)$. We assume that the pulse shape deployed in $x_k(t)$ is the sinc function.

6.1.1 Narrowband system

Here the delays τ_k are small compared to the symbol time, so that $x_k(t - n\tau_k) \approx x_k(t)$, and (6.3) can be written as

$$\mathbf{y}(t) = \mathbf{H}_{NB}\mathbf{x}(t) + \mathbf{n}(t), \tag{6.4}$$

where the received vector $\mathbf{y}(t) = [y_0(t), \ldots, y_{N-1}(t)]^\intercal$, the data symbols vector $\mathbf{x}(t) = [x_0(t), \ldots, x_{K-1}(t)]^\intercal$, the noise vector $\mathbf{n}(t) = [n_0(t), \ldots, n_{N-1}(t)]^\intercal$, the narrow-band chan-

nel matrix $\mathbf{H}_{NB} = [e^{-jn\Omega_k}]_{n,k}\mathcal{D}([A_k]_k)$, and $\mathcal{D}(.)$ is the diagonlization operator.

The LMMSE receiver is given by

$$\hat{\mathbf{x}} = \mathbf{W}_{NB}\mathbf{y}, \tag{6.5}$$

such that

$$\mathbf{W}_{NB} = (\mathbf{H}_{NB}^H\mathbf{H}_{NB} + \sigma^2\mathbf{I})^{-1}\mathbf{H}_{NB}^H. \tag{6.6}$$

6.1.2 Wideband system

Here, the delays τ_k can be larger than a symbol period. Taking the temporal Fourier transform $\mathcal{F}(.)$ for equation (6.3) yields

$$\begin{aligned}\tilde{y}_n(f) &= \sum_{k=0}^{K-1} A_k\tilde{x}_k(f)e^{-j2\pi n(f+f_c)\tau_k} + \tilde{n}_n(f) \\ &= \sum_{k=0}^{K-1} A_k\tilde{x}_k(f)e^{-jn(1+f/f_c)\Omega_k} + \tilde{n}_n(f),\end{aligned} \tag{6.7}$$

where $\tilde{y}_n(f) = \mathcal{F}(y_n(t))$, $\tilde{x}_k(f) = \mathcal{F}(x_k(t))$, and $\tilde{n}_n(f) = \mathcal{F}(n_n(t))$. This system of equations can be written in vector form as follows:

$$\tilde{\mathbf{y}}(f) = \tilde{\mathbf{H}}_{WB}(f)\tilde{\mathbf{x}}(f) + \tilde{\mathbf{n}}(f), \tag{6.8}$$

where the received vector in the frequency domain at the base station

$$\tilde{\mathbf{y}}(f) = [\tilde{y}_0(f), \ldots, \tilde{y}_{N-1}(f)]^\intercal$$

, the transmitted symbols stream in the frequency domain by the K users

$$\tilde{\mathbf{x}}(f) = [\tilde{x}_0(f), \ldots, \tilde{x}_{K-1}(f)]^\intercal$$

, the noise vector $\tilde{\mathbf{n}}(f) = [\tilde{n}_0(f), \ldots, \tilde{n}_{N-1}(f)]^\intercal$, and the wide-band channel matrix $\tilde{\mathbf{H}}_{WB}(f) = [e^{-jn(1+f/f_c)\Omega_k}]_{n,k}\mathcal{D}([A_k]_k)$.

The LMMSE solution at each frequency f is given by

$$\hat{\mathbf{x}}(t) = \mathcal{F}^{-1}(\tilde{\mathbf{W}}(f)\tilde{\mathbf{y}}(f)), \tag{6.9}$$

where

$$\tilde{\mathbf{W}}_{WB}(f) = \left(\tilde{\mathbf{H}}_{WB}(f)^H\tilde{\mathbf{H}}_{WB}(f) + \sigma^2\mathbf{I}\right)^{-1}\tilde{\mathbf{H}}_{WB}(f)^H \tag{6.10}$$

Practical digital signal processing methods must work with discretized system models, as discussed in the next subsection.

6.1.3 Signaling structure

Each mobile sends data in blocks of the length of M symbols. There is a guard interval of length L symbols between successive blocks to prevent Inter-Block Interference (IBI). Each guard interval is filled with a cyclic prefix to emulate a circular convolution. At the receiver, the base station discards the cyclic prefix and handles each data block separately. The sampled received signal can be written as

$$y_{n,m} = \sum_{k=0}^{K-1} A_k x_k(mT_s - n\tau_k)e^{-jn\Omega_k} + n_{n,m}, \tag{6.11}$$

where $T_s = 1/BW$ is the sampling time and BW is the bandwidth of the received signal, and $m \in [0, M)$.

6.2 Benchmark Wideband LMMSE

Fig. 6.4 illustrates the block diagram of the benchmark wideband LMMSE algorithm, which involves the following steps.

Temporal FFT

The discrete Fourier transform of equation (6.11) is given as follows,

$$\tilde{y}_{n,\ell} = \sum_{k=0}^{K-1} A_k \tilde{x}_{k,\ell} e^{-jn(1+\ell\frac{BW}{Mf_c})\Omega_k} + \tilde{n}_{n,\ell}, \tag{6.12}$$

where $[\tilde{x}_{k,-M/2}, \ldots, \tilde{x}_{k,M/2-1}] = \text{DFT}([x_k(\frac{m}{BW})]_m)$, DFT(.) is the discrete Fourier transform operation, and $\ell \in [-M/2, M/2)$. Given that $\tilde{\mathbf{y}}_\ell = [\tilde{y}_{n,\ell}]_n$, $\tilde{\mathbf{x}}_\ell = [\tilde{x}_{k,\ell}]_k$, and $\tilde{\mathbf{n}}_\ell = [\tilde{n}_{n,\ell}]_n$, the previous equation can be written in vector form as follows,

$$\tilde{\mathbf{y}}_\ell = \mathbf{H}(\ell)\tilde{\mathbf{x}}_\ell + \tilde{\mathbf{n}}_\ell, \tag{6.13}$$

where the channel matrix is given as

$$\mathbf{H}(\ell) = \left[e^{-jn(1+\ell\frac{BW}{Mf_c})\Omega_k}\right]_{n,k} \mathcal{D}([A_k]_k). \tag{6.14}$$

The temporal FFT divides the frequency domain of the received signal to M sub-band assuming that the channel is almost constant on the sub-band.

LMMSE Detection

In each sub-band, we can recall the narrowband assumption and detect $\tilde{\mathbf{x}}_\ell$ using LMMSE as follows,

$$\hat{\tilde{\mathbf{x}}}_\ell = \mathbf{W}(\ell)\tilde{\mathbf{y}}_\ell, \tag{6.15}$$

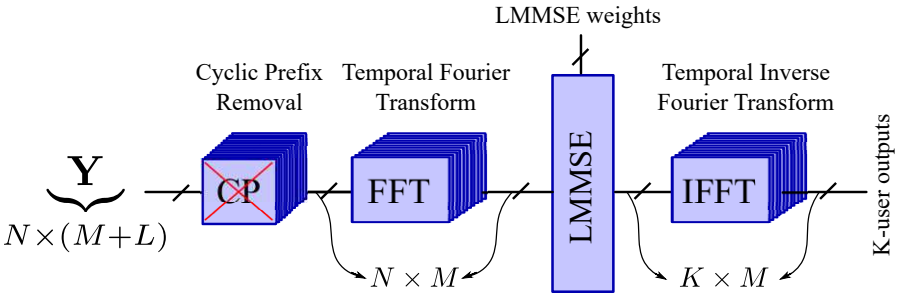


Figure 6.4: Block diagram for benchmark wideband LMMSE, where $\mathbf{Y}$ denotes the grid of received samples in antenna-space and time domain.

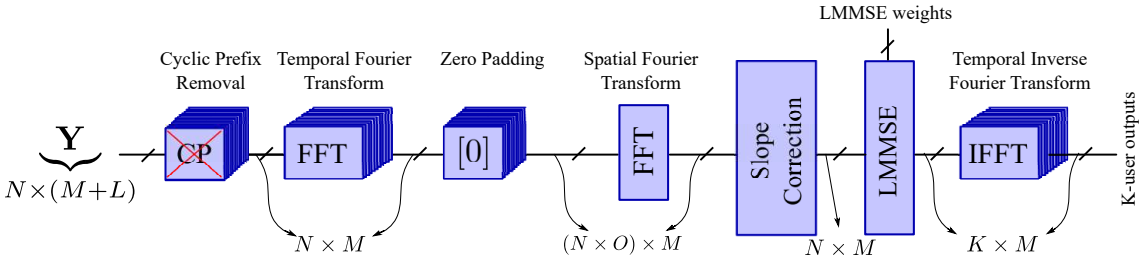


Figure 6.5: Block diagram of the proposed wideband LMMSE approach, where $\mathbf{Y}$ denotes the grid of received samples in antenna-space and time domain.

where

$$\mathbf{W}(\ell) = \left(\mathbf{H}(\ell)^H \mathbf{H}(\ell) + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}(\ell)^H. \tag{6.16}$$

Temporal IFFT

Finally, the receiver retrieves the data of the k^{th}user from the frequency-domain using temporal IFFT as $[\hat{x}_{k,0}, \ldots, \hat{x}_{k,M-1}] = \text{IDFT}([\hat{\tilde{x}}_{k,\ell}]_\ell)$, where IDFT(.) is the inverse discrete Fourier transform operator. If the mobiles used OFDM instead of single carrier signaling, this step would have been already on the transmitter side.

6.3 Proposed Wideband LMMSE

Fig. 6.5 delineates the block diagram of the proposed wideband LMMSE algorithm. The proposed algorithm is described in the following steps.

Temporal FFT

Similar to equation (6.13), the temporal FFT divide the bandwidth to M equally sub-bands. After this step, the conventional and proposed schemes start to differ.

Zero Padding and Spatial FFT

In this step, we aim to transform the received samples from the antenna-space to the beamspace with oversampling ratio O. So first, we pad the received samples vector per frequency bin $\tilde{\mathbf{y}}_\ell$ by a vector of zeros of size $(N(O-1)) \times 1$. Then, we perform upon the spatial FFT as follows,

$$\breve{\mathbf{y}}_\ell = \mathrm{DFT}([\tilde{\mathbf{y}}_\ell, \mathbf{0}_{N(O-1)}]^\intercal), \tag{6.17}$$

where $\mathbf{0}_M$ is vector of zeros of size $M \times 1$. Fig. 6.6 portrays the grid of the received samples power in beamspace-frequency-domain, i.e, $[|\breve{y}_{p,\ell}|^2]_{p,\ell}$, where $p \in [-\frac{NO}{2}, \frac{NO}{2})$. This figure delineates the key idea of the proposed algorithm. As shown, each user is represented by a line that is governed by the equation $\Omega = \Omega_k(1 + f/f_c)$, where Ω is the spatial frequency, and intersect the point $f = 0$ and $\Omega = \Omega_k$. This can be inferred by the explicit expression of equation (6.17) which is given as follows,

$$\breve{y}_{p,\ell} = \sum_{k=0}^{K-1} A_k \tilde{x}_{k,\ell} c_{k,p,\ell} \frac{\sin(\frac{N}{2}(s_\ell \Omega_k + \frac{2\pi p}{NO}))}{\sin(\frac{1}{2}(s_\ell \Omega_k + \frac{2\pi p}{NO}))} + \breve{n}_{p,\ell}, \tag{6.18}$$

where $s_\ell = (1 + \ell \frac{BW}{M f_c})$, $c_{k,p,\ell} = \frac{1}{\sqrt{NO}} e^{-j\frac{(N-1)}{2}(s_\ell \Omega_k + \frac{2\pi p}{NO})}$, $\ell \frac{BW}{M}$ is the discretized version of f, $\frac{2\pi p}{NO}$ is the discretized version of Ω, and $\breve{n}_{p,\ell}$ is the noise element in the beamspace. The main idea of the proposed algorithm is to correct these tilted lines to make them horizontal. This is done by dividing the grid into multiple segments and correcting the slope of each segment, as shown in Fig. 6.7. Consequently, one LMMSE beamformer is adequate to retrieve all the data of a user per segment.

Slope Correction

We use linear interpolation to acquire the in-between points to correct the slope. We correct the slope by scaling p in equation (6.18), in that $\breve{y}_{p,\ell}$ becomes $\breve{y}_{p\frac{s_\ell}{s_x},\ell}$ where s_x corresponds to the frequency of the mid of the segment. We choose to work with linear interpolation due to its feasibility to be built in hardware.

Beamspace LMMSE Detection

In a segment, we use the channel corresponding to the mid of the segment to calculate the LMMSE solution for that segment. One can use efficient algorithm to do LMMSE detection in the beamspace as shown in [13].

Temporal IFFT

Finally, similar to the benchmark algorithm, the receiver retrieves the users' data from the frequency-domain using temporal IFFT.

Why the need for partitioning the data grid into multiple segments?

In slope correction process, the beam shapes get compressed or expanded. Fig. 6.8 depicts the beam shape of a single user after the data grid correction. Because the beam shape changes across the band, then one LMMSE beamformer is not adequate to retrieve the data from the entire band.

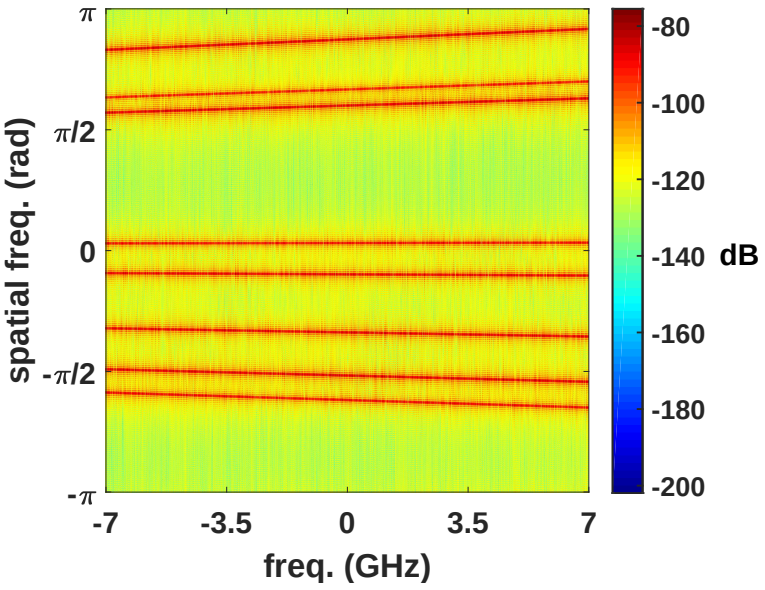


Figure 6.6: The beamspace-frequency-domain data grid before correction.

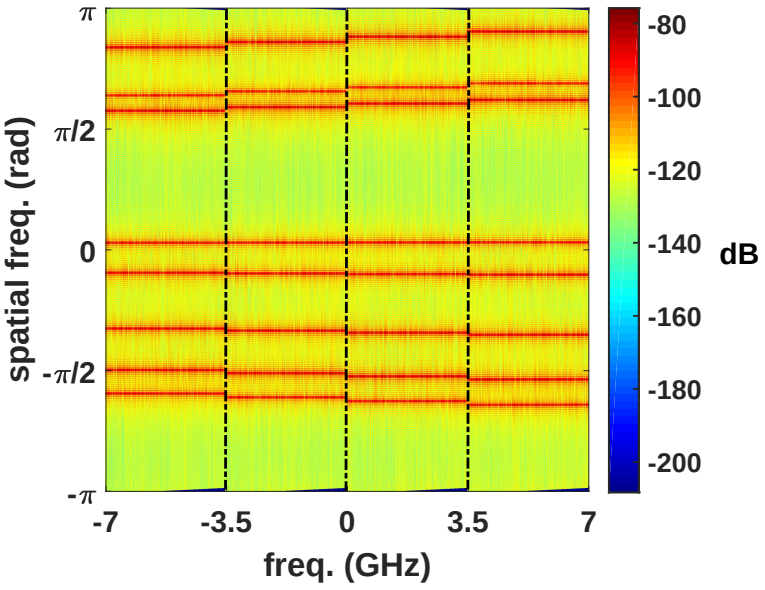


Figure 6.7: The beamspace-frequency-domain data grid after correction.

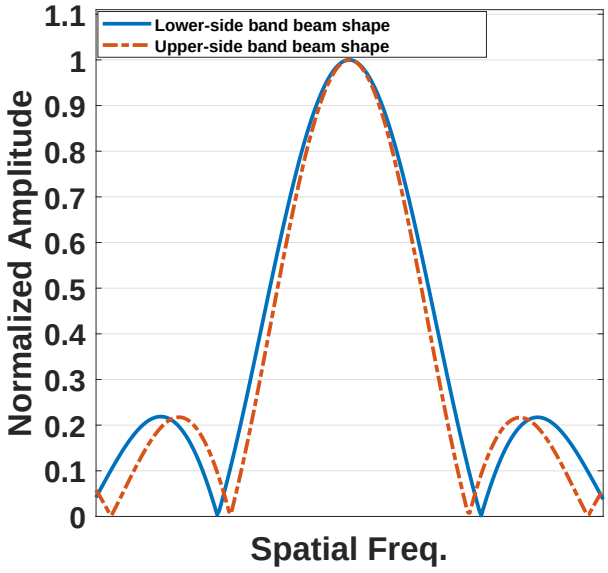


Figure 6.8: The beam shape of a single user after the grid correction.

6.4 Results

The system setup is as illustrated in Fig. 6.3. The number of antennas is fixed at $N = 256$ for all numerical experiments. In the simulations, the field of view is restricted to $-\pi/3 \leq \theta \leq \pi/3$. The users are uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, $R_{\min} = 5$ m and $R_{\max} = 100$ m, respectively. While the user terminals are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two users in order not to incur excessive interference, arbitrarily choosing it as half the 3dB beamwidth: $\Delta\Omega_{\min} = \frac{2.783}{N}$ [4]. We assume that users with similar spatial frequency can be served at different time. The number of users served simultaneously in the cell is 64 users, each is using QPSK modulation.

The carrier frequency of the system is $f_c = 140$ GHz and we choose to operate at bandwidth equals to 10% of f_c, i.e., 14 GHz. As can be inferred from Fig. 6.2,

the narrowband assumption is inappropriate to use when dealing with such wide bandwidth and long array. The length of the cyclic prefix used $L = 2 \times \lceil BW \times \tau_{\max} \rceil = 24$ samples. We assume perfect CSI and no power control is deployed. The presented BER curves is plotted versus the SNR of the edge user at 100 m which is defined as $\text{SNR}_{\text{edge}} = NA_{100m}^2/\sigma^2$, where $N = 256$ elements, and the channel strength at 100 m $A_{100}^2 = \lambda^2/(4\pi R_k)^2 = 0.002^2/(4\pi 100)^2$.

6.4.1 Benchmark Wideband LMMSE

Fig. 6.9 shows the BER at 95% availability when the benchmark wideband LMMSE is employed. The benchmark algorithm is parameterized with the block length M, which determines the number of chunks into which the bandwidth is divided. As shown in the figure, a block length of 128 is enough to be within 1 dB of the baseline, defined for a system in which the narrowband assumption holds. Generally, what matters most is the maximum delay spread experienced by each sub-band not the block length itself. Using a block length of 128 symbols and bandwidth of 14 GHz, the maximum delay spread can be computed from equation (6.2) to be 9% of the symbol time.

6.4.2 Proposed Wideband LMMSE

Fig. 6.10 shows the BER at 95% availability for the proposed wideband LMMSE scheme. We choose the block length to be 256 to compensate for any additional errors introduced by the proposed algorithm. O refers to the oversampling ratio in beamspace, and S denotes the number of segments used for beamspace partitioning. We use linear interpolation for simplicity. As shown in the figure, the proposed algorithm that uses an oversampling ratio of $O = 8$ and $S = 8$ yields a BER that is within 1 dB of the baseline. In contrast to the benchmark algorithm, which must learn a set of LMMSE weights per

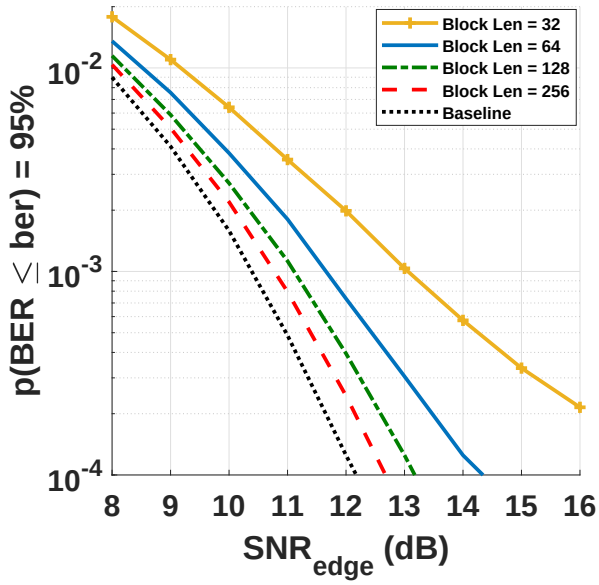


Figure 6.9: BER at 95% availability for the benchmark scheme with different block sizes.

sub-band per user (i.e., 128 weight vectors per user), the proposed scheme only needs to learn 8 weight vectors per user. The price paid for simplifying the acquisition process is the increase in beamformer complexity due to an O times larger spatial FFT (the additional complexity due to linear interpolation is negligible).

6.5 Conclusion

Our work shows that, as we push the limits of all-digital processing in scaling both bandwidth and spatial degrees of freedom, it is critical to exploit the characteristics of the mmWave channel to simplify processing. Specifically, we exploit the sparsity of the mmWave channel in this chapter to drastically simplify the process of weights acquisition for frequency domain LMMSE (by 16-fold for the parameters considered here).

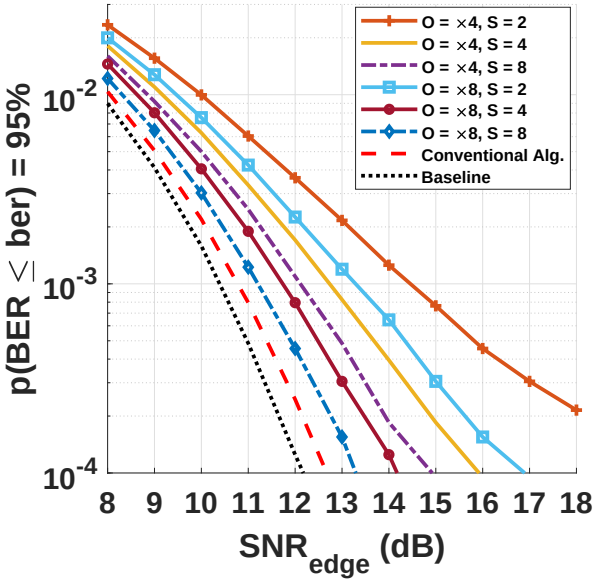


Figure 6.10: BER at 95% availability for the proposed wideband LMMSE for different settings and a block size of 256.

Chapter 7

Conclusions and Future Work

The contributions presented in this thesis demonstrate the critical role of co-design of signal processing and hardware, and of accounting for the unique characteristics of the mmWave band, for progress towards the realization of next generation all-digital massive MIMO systems. The small carrier wavelength allows realization of antenna arrays with a large number of elements, and leads to spatially sparse channels because of loss due to reflection (reflecting surfaces look rougher at small wavelengths) and blockage (obstacles look larger at small wavelengths). The large available bandwidth allows us to consider relatively small signaling constellations even when targeting multiGbps data rates per user. Key technical insights from the work presented in this thesis are summarized as follows:

- The results of Chapter 2 imply that if we wish to support a given number of simultaneous users, the linearity requirements on RF and baseband analog processing, as well as ADC precision, can be relaxed by increasing the number of antenna elements in order to reduce the load factor. It is worth mentioning complementary results demonstrating that hardware specifications regarding phase noise also benefit from scaling [85, 86].

- The results of Chapters 3 through 6 take advantage of the sparsity of the mmWave spatial channel to concentrate the power for each user into a small number of signaling dimensions via a spatial FFT across antennas. This observation can be employed to devise beamspace signal processing schemes that scale far better (as the number of users and antennas increase) than conventional multiuser detection or precoding strategies.

- The narrowband assumption for array modeling and processing must be revisited as we scale up both the bandwidth and the number of antenna elements. The work reported in Chapter 6 is a first step in this direction.

There are many important directions for future research and development. A natural extension of our system design framework is to explore the impact of analog nonlinearities and low-precision ADCs on beamspace multiuser detection and precoding. The development of efficient channel estimation and tracking for uplink multiuser detection and downlink precoding for all-digital massive MIMO, and the design of link layer and medium access control protocols, are important steps towards a complete system design. The design of efficient uplink and downlink signal processing architectures for *wideband* massive MIMO, building on the preliminary results in Chapter 6, remains an open problem.

Hardware realizations of our proposed approaches require continuing research, including exploration of RFIC design and packaging [87] for tiled architectures [86] for scaling up the number of antennas, and development of efficient FFT front ends [88, 89] for beamspace signal processing.

Appendix A

Uplink Link Budget

We provide here example parameters that demonstrate that the link budget for all-digital massive multiuser MIMO uplink system is realizable with low-cost silicon:

- antenna element gain covering a hemisphere is 3 dBi,
- 16-element array at the mobile gives 12 dBi transmit beamforming gain, plus 12 dB power pooling gain,
- 256-element array in the base station gives 24 dBi receive beamforming gain,
- noise figure for each RF chain in the base station of 7 dB,
- thermal noise power over 5 GHz bandwidth is about -77 dBm,
- and free space path loss of an edge user at 100 m using a carrier frequency of 140 GHz is about 115 dB.

The transmit power required from each power amplifier (PA) at the mobile to achieve a target SNR (in dB) for an edge-user, namely $\mathrm{SNR}_{\mathrm{edge}}|_{\mathrm{dB}}$, can now be computed as

$$P_{\mathrm{PA}} = \mathrm{SNR}_{\mathrm{edge}}|_{\mathrm{dB}} - 9 \ \ \mathrm{dBm}. \tag{A.1}$$

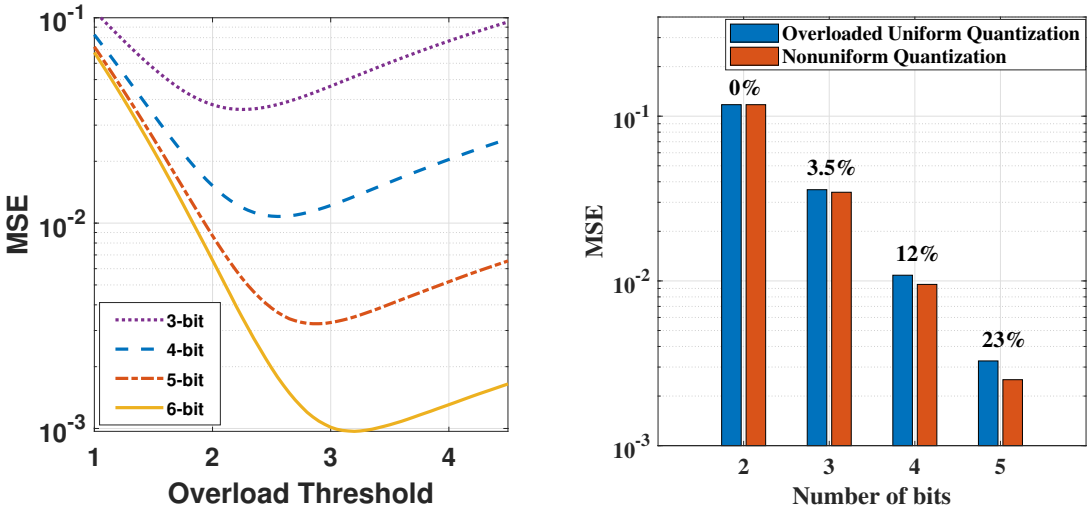
For example, $\mathrm{SNR}_{\mathrm{edge}}|_{\mathrm{dB}}$ of about 16 dB (shown to suffice for our case study) requires 7 dBm PA output, which is realizable in CMOS (CMOS designs of up to 11 dBm have been reported in [90]).

Appendix B

Uniform VS Nonuniform Quantization

Our simulation results are for an overloaded ADC. The overloaded uniform ADC comprises two regions in its I/O characteristic, the granular and overload regions. The granular region is quantized uniformly, with bounded quantization noise. While quantization noise in the overload region, represented by the quantizer levels at the edges, is unbounded, the contribution to the MSE is kept comparable to that of the granular region by minimizing the MSE for the given input distribution; see Fig. B.1 (a), where MSE is plotted against overload threshold.

An alternative is to employ an MSE-optimal quantizer using Lloyd's algorithm [91], with quantization bins as listed in [92]. The MSE comparison between these two options is shown in Fig. B.1 (b). The advantage of nonuniform MSE-optimal quantization is barely noticeable for the small number of quantization bits of interest here, hence we choose to work with the simpler overloaded uniform quantizer.



(a) (b)

Figure B.1: (a) MSE versus overload threshold. (b) MSE comparison of overload uniform quantizer versus MSE-optimal nonuniform quantizer. The percentages represent the relative reduction in MSE from using MSE-optimal nonuniform quantization

Appendix C

Linear MMSE Properties

From the point of view of a given user (the *desired* user) with channel $\mathbf{h}$, we may write the received signal corresponding to a single symbol as

$$\mathbf{r} = s\mathbf{h} + \mathbf{w}_I + \mathbf{w}_N, \tag{C.1}$$

where s denotes the transmitted symbol, $\mathbf{w}_I$ denotes the interference vector and $\mathbf{w}_N$ is the zero mean noise vector with covariance matrix $\sigma_n^2\mathbf{I}$. Standard assumptions necessary for effective interference suppression are that the desired symbol is uncorrelated with the interference and noise: $\mathbb{E}[s^*\mathbf{w}_I] = \mathbb{E}[s^*\mathbf{w}_N] = \mathbf{0}$. We also assume that the interference and noise are uncorrelated.

A linear correlator $\mathbf{c}$ produces a decision statistic $\mathbf{c}^H\mathbf{r}$ for the desired symbol, and its SINR is given by

$$\begin{aligned}\mathrm{SINR}(\mathbf{c}) =& \frac{\mathbb{E}[|s\mathbf{c}^H\mathbf{h}|^2]}{\mathbb{E}[|\mathbf{c}^H(\mathbf{w}_I + \mathbf{w}_N)|^2]} \\ =& \frac{\sigma_s^2|\mathbf{c}^H\mathbf{h}|^2}{\mathbf{c}^H\mathbf{R}_I\mathbf{c} + \sigma_n^2||\mathbf{c}||^2},\end{aligned} \tag{C.2}$$

where $\mathbf{R}_I = \mathbb{E}[\mathbf{w}_I \mathbf{w}_I^H]$ is the interference covariance matrix, and $\mathbf{R}_N = \mathbb{E}[\mathbf{w}_N \mathbf{w}_N^H] = \sigma_n^2 \mathbf{I}$ is the noise covariance matrix.

The LMMSE correlator minimizes MSE $= \mathbb{E}[|\mathbf{c}^H \mathbf{r} - s|^2]$ and maximizes SINR [44]. For the additive noise-plus-interference model (C.1), it is known to be proportional to a whitened matched filter (i.e., it suppresses interference by whitening it):

$$\mathbf{c}_{\text{MMSE}} = \alpha(\mathbf{R}_I + \mathbf{R}_N)^{-1}\mathbf{h} = \alpha(\mathbf{R}_I + \sigma_n^2 \mathbf{I})^{-1}\mathbf{h}, \tag{C.3}$$

where α is a scale factor that can be solved for easily (e.g., see [44]). Since SINR does not depend on scale factor, it is easy to show, plugging into (C.2), that

$$\begin{aligned} \text{SINR} =& \sigma_s^2 \mathbf{h}^H \left(\mathbf{R}_I + \mathbf{R}_N\right)^{-1} \mathbf{h} \\ =& \sigma_s^2 \mathbf{h}^H \left(\mathbf{R}_I + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{h}. \end{aligned} \tag{C.4}$$

Let us also for reference define the SNR:

$$\text{SNR} = \sigma_s^2 \mathbf{h}^H \left(\mathbf{R}_N\right)^{-1} \mathbf{h} = \sigma_s^2 ||\mathbf{h}||^2 / \sigma_n^2. \tag{C.5}$$

Remark 1 *A positive definite matrix $\mathbf{A}(\theta)$ increases with θ if $\mathbf{A}(\theta) - \mathbf{A}(\theta') \geq \mathbf{0}$ for any $\theta > \theta'$. That is, for any vector $\mathbf{u}$, $\mathbf{u}^H \mathbf{A}(\theta)\mathbf{u} \geq \mathbf{u}^H \mathbf{A}(\theta')\mathbf{u}$.*

We can now infer the following properties relevant for our approach to performance analysis, stated as a lemma.

Lemma C.0.1 *If the noise level σ_n^2 increases, with the signal and interference characteristics unchanged, then*

(a) Absolute performance gets worse, with SINR and SNR both decreasing.

(b) The noise enhancement gets better: $\frac{SNR}{SINR}$ decreases.

Proof: For (a), we note that the positive definite matrix $\mathbf{R}_I + \sigma_n^2\mathbf{I}$ increases with σ_n^2, hence its inverse decreases with σ_n^2. For (b), note that

$$\begin{aligned}\frac{\text{SNR}}{\text{SINR}} =& \frac{||\mathbf{h}||^2/\sigma_n^2}{\mathbf{h}^H\left(\mathbf{R}_I + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{h}}, \\ =& \frac{||\mathbf{h}||^2}{\mathbf{h}^H\left(\mathbf{R}_I/\sigma_n^2 + \mathbf{I}\right)^{-1}\mathbf{h}}. \end{aligned} \tag{C.6}$$

The positive definite matrix $\mathbf{R}_I/\sigma_n^2 + \mathbf{I}$ decreases with σ_n^2, hence its inverse increases with σ_n^2. Thus, the denominator on the right-hand side of equation (C.6) increases with σ_n^2, while the numerator is independent of it, proving the desired result.

■

Bibliography

[1] A. Wiesel, Y. C. Eldar, and S. Shamai, *Linear precoding via conic optimization for fixed MIMO receivers*, *IEEE Transactions on Signal Processing* **54** (Jan, 2006) 161–176.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, *Massive mimo for next generation wireless systems*, *IEEE communications magazine* **52** (2014), no. 2 186–195.

[3] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson, *The role of small cells, coordinated multipoint, and massive mimo in 5g*, *IEEE communications magazine* **52** (2014), no. 5 44–51.

[4] C. A. Balanis, *Antenna Theory: Analysis and Design*. Wiley-Interscience, New York, NY, USA, 2005.

[5] M. Abdelghany, A. A. Farid, U. Madhow, and M. J. Rodwell, *Towards all-digital mmWave massive MIMO: Designing around nonlinearities*, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 1552–1557, IEEE, 2018.

[6] M. Abdelghany, A. A. Farid, M. E. Rasekh, U. Madhow, and M. J. Rodwell, *A design framework for all-digital mmWave massive MIMO with per-antenna nonlinearities*, *IEEE Transactions on Wireless Communications* (2021).

[7] C. Studer and G. Durisi, *Quantized massive MU-MIMO-OFDM uplink*, *IEEE Transactions on Communications* **64** (2016), no. 6.

[8] S. Jacobsson, U. Gustavsson, G. Durisi, and C. Studer, *Massive MU-MIMO-OFDM uplink with hardware impairments: Modeling and analysis*, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2018.

[9] A. Maltsev, A. Pudeyev, I. Karls, I. Bolotin, G. Morozov, R. Weiler, M. Peter, and W. Keusgen, *Quasi-deterministic approach to mmWave channel modeling in a non-stationary environment*, in *2014 IEEE Globecom Workshops (GC Wkshps)*, IEEE, 2014.

[10] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, *Millimeter wave mobile communications for 5G cellular: It will work!*, *IEEE access* **1** (2013).

[11] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, *Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications*, *IEEE transactions on antennas and propagation* **61** (2012), no. 4.

[12] M. Jacob, S. Priebe, R. Dickhoff, T. Kleine-Ostmann, T. Schrader, and T. Kurner, *Diffraction in mm and sub-mm wave indoor propagation channels*, *IEEE Transactions on Microwave Theory and Techniques* **60** (2012), no. 3.

[13] M. Abdelghany, U. Madhow, and A. Tölli, *Beamspace local LMMSE: An efficient digital backend for mmWave massive MIMO*, in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2019.

[14] M. Abdelghany, M. E. Rasekh, and U. Madhow, *Scalable nonlinear multiuser detection for mmwave massive mimo*, in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, IEEE, 2020.

[15] M. Abdelghany, U. Madhow, and A. Tölli, *Efficient beamspace downlink precoding for mmWave massive MIMO*, in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2019.

[16] M. Abdelghany, U. Madhow, and M. Rodwell, *An efficient digital backend for wideband single-carrier mmWave massive MIMO*, in *to be presented in IEEE Global Communications Conference (Globecom), Waikoloa, Hawaii, Dec. 2019*, 2019.

[17] J. H. Brady and A. M. Sayeed, *Wideband communication with high-dimensional arrays: New results and transceiver architectures*, in *IEEE International Conference on Communication Workshop (ICCW)*, pp. 1042–1047, 2015.

[18] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, *Spatial-and frequency-wideband effects in millimeter-wave massive MIMO systems*, *IEEE Transactions on Signal Processing* **66** (2018) 3393–3406.

[19] A. Puglielli, A. Townley, G. LaCaille, V. Milovanović, P. Lu, K. Trotskovsky, A. Whitcombe, N. Narevsky, G. Wright, T. Courtade, *et. al.*, *Design of energy-and cost-efficient massive mimo arrays*, *Proceedings of the IEEE* **104** (2015), no. 3 586–606.

[20] A. Babakhani, X. Guan, A. Komijani, A. Natarajan, and A. Hajimiri, *A 77-ghz phased-array transceiver with on-chip antennas in silicon: Receiver and antennas*, *IEEE Journal of Solid-State Circuits* **41** (2006), no. 12 2795–2806.

[21] W. Hong, K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, *Study and prototyping of practically large-scale mmwave antenna systems for 5g cellular devices*, *IEEE Communications Magazine* **52** (2014), no. 9 63–69.

[22] E. Cohen, M. Ruberto, M. Cohen, O. Degani, S. Ravid, and D. Ritter, *A cmos bidirectional 32-element phased-array transceiver at 60 ghz with ltcc antenna*, *IEEE Transactions on Microwave Theory and Techniques* **61** (2013), no. 3 1359–1375.

[23] A. Valdes-Garcia, S. T. Nicolson, J.-W. Lai, A. Natarajan, P.-Y. Chen, S. K. Reynolds, J.-H. C. Zhan, D. G. Kam, D. Liu, and B. Floyd, *A fully integrated 16-element phased-array transmitter in sige bicmos for 60-ghz communications*, *IEEE journal of solid-state circuits* **45** (2010), no. 12 2757–2773.

[24] Z. Marzi, D. Ramasamy, and U. Madhow, *Compressive channel estimation and tracking for large arrays in mm-wave picocells*, *IEEE Journal of Selected Topics in Signal Processing* **10** (2016), no. 3 514–527.

[25] H. Yan and D. Cabric, *Compressive initial access and beamforming training for millimeter-wave cellular systems*, *IEEE journal of selected topics in signal processing* **13** (2019), no. 5 1151–1166.

[26] F. Sohrabi and W. Yu, *Hybrid digital and analog beamforming design for large-scale antenna arrays*, *IEEE Journal of Selected Topics in Signal Processing* **10** (2016), no. 3 501–513.

[27] O. Bakr, M. Johnson, J. Park, E. Adabi, K. Jones, and A. Niknejad, *A scalable-low cost architecture for high gain beamforming antennas*, in *2010 IEEE International Symposium on Phased Array Systems and Technology*, pp. 806–813, IEEE, 2010.

[28] G. Zhu, K. Huang, V. K. Lau, B. Xia, X. Li, and S. Zhang, *Hybrid beamforming via the kronecker decomposition for the millimeter-wave massive mimo systems*, *IEEE Journal on Selected Areas in Communications* **35** (2017), no. 9 2097–2114.

[29] J.-C. Chen, *Hybrid beamforming with discrete phase shifters for millimeter-wave massive mimo systems*, *IEEE Transactions on Vehicular Technology* **66** (2017), no. 8 7604–7608.

[30] M. Chung, L. Liu, O. Edfors, and F. Tufvesson, *Millimeter-wave massive mimo testbed with hybrid beamforming*, in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–2, IEEE, 2020.

[31] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, *Hybrid beamforming for massive mimo: A survey*, *IEEE Communications Magazine* **55** (2017), no. 9 134–141.

[32] H. Yan, S. Ramesh, T. Gallagher, C. Ling, and D. Cabric, *Performance, power, and area design trade-offs in millimeter-wave transmitter beamforming architectures*, *IEEE Circuits and Systems Magazine* **19** (2019), no. 2.

[33] O. T. Demir and E. Bjornson, *The bussgang decomposition of nonlinear systems: Basic theory and mimo extensions [lecture notes]*, *IEEE Signal Processing Magazine* **38** (2020), no. 1 131–136.

[34] B. Razavi and R. Behzad, *RF microelectronics*, vol. 2. Prentice Hall New Jersey, 1998.

[35] J. BUSSGANG, *Crosscorrelation functions of amplitude-distorted gaussian signals*, *MIT Res. Lab. Elec. Tech. Rep.* **216** (1952).

[36] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, *Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits*, *IEEE Transactions on Information Theory* **60** (2014), no. 11.

[37] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, *Uplink achievable rate for massive MIMO systems with low-resolution ADC*, *IEEE Communications Letters* **19** (2015), no. 12.

[38] J. Zhang, L. Dai, S. Sun, and Z. Wang, *On the spectral efficiency of massive MIMO systems with low-resolution ADCs*, *IEEE Communications Letters* **20** (2016), no. 5.

[39] L. Xu, X. Lu, S. Jin, F. Gao, and Y. Zhu, *On the uplink achievable rate of massive MIMO system with low-resolution ADC and RF impairments*, *IEEE Communications Letters* **23** (2019), no. 3.

[40] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, *Achievable uplink rates for massive mimo with coarse quantization*, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6488–6492, IEEE, 2017.

[41] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, *Throughput analysis of massive MIMO uplink with low-resolution ADCs*, *IEEE Transactions on Wireless Communications* **16** (2017), no. 6.

[42] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. Wiley, 2014.

[43] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, vol. 159. Springer Science & Business Media, 2012.

[44] U. Madhow and M. L. Honig, *MMSE interference suppression for direct-sequence spread-spectrum CDMA*, *IEEE transactions on communications* **42** (1994), no. 12.

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

[46] B. Hajek, *Random processes for engineers*. Cambridge university press, 2015.

[47] J. Minkoff, *The role of AM-to-PM conversion in memoryless nonlinear systems*, *IEEE Transactions on Communications* **33** (1985), no. 2.

[48] E. Björnson, L. Sanguinetti, and J. Hoydis, *Hardware distortion correlation has negligible impact on UL massive MIMO spectral efficiency*, *IEEE Transactions on Communications* **67** (2018), no. 2.

[49] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, *Quantized precoding for massive MU-MIMO*, *IEEE Transactions on Communications* **65** (2017), no. 11.

[50] S. Ulukus and R. D. Yates, *Adaptive power control and MMSE interference suppression*, *Wireless Networks* **4** (1998), no. 6.

[51] A. Adhikary, J. Nam, J. Ahn, and G. Caire, *Joint spatial division and multiplexing—The large-scale array regime*, *IEEE Trans. Inf. Theory* **59** (Oct, 2013) 6441–6463.

[52] J. Nam, A. Adhikary, J. Ahn, and G. Caire, *Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling*, *IEEE Journal of Selected Topics in Signal Processing* **8** (Oct., 2014) 876–890.

[53] A. Padmanabhan and A. Tölli, *An iterative approach for inter-group interference management in two-stage precoder design*, in *Proc. IEEE Globecom 2018, Abu Dhabi, UAE*, Dec., 2018.

[54] T. Takahashi, A. Tölli, S. Ibi, and S. Sampei, *Layered belief propagation for low-complexity large MIMO detection based on statistical beams*, in *Proc. IEEE International Conference on Communications 2019, Shanghai, China*, May., 2019.

[55] A. Krishnamoorthy and D. Menon, *Matrix inversion using Cholesky decomposition*, in *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 70–72, IEEE, 2013.

[56] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[57] U. Madhow, *MMSE interference suppression for timing acquisition and demodulation in direct-sequence CDMA systems*, *IEEE Transactions on Communications* **46** (1998), no. 8 1065–1075.

[58] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[59] S. Yang and L. Hanzo, *Fifty years of MIMO detection: The road to large-scale MIMOs*, *IEEE Communications Surveys & Tutorials* **17** (2015), no. 4 1941–1988.

[60] C. Tang, C. Liu, L. Yuan, and Z. Xing, *High precision low complexity matrix inversion based on Newton iteration for data detection in the massive MIMO*, *IEEE Communications Letters* **20** (2016), no. 3 490–493.

[61] F. Wang, C. Zhang, J. Yang, X. Liang, X. You, and S. Xu, *Efficient matrix inversion architecture for linear detection in massive MIMO systems*, in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pp. 248–252, IEEE, 2015.

[62] J. Zeng, J. Lin, and Z. Wang, *An improved Gauss-Seidel algorithm and its efficient architecture for massive MIMO systems*, *IEEE Transactions on Circuits and Systems II: Express Briefs* **65** (2018), no. 9 1194–1198.

[63] Z. Wu, Y. Xue, X. You, and C. Zhang, *Hardware efficient detection for massive MIMO uplink with parallel Gauss-Seidel method*, in *2017 22nd International Conference on Digital Signal Processing (DSP)*, pp. 1–5, IEEE, 2017.

[64] C. Studer, S. Fateh, and D. Seethaler, *ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation*, *IEEE Journal of Solid-State Circuits* **46** (2011), no. 7 1754–1765.

[65] C. Zhang, X. Liang, Z. Wu, F. Wang, S. Zhang, Z. Zhang, and X. You, *On the low-complexity, hardware-friendly tridiagonal matrix inversion for correlated massive MIMO systems*, *IEEE Transactions on Vehicular Technology* **68** (2019), no. 7 6272–6285.

[66] A. K. Sah and A. K. Chaturvedi, *An MMP-based approach for detection in large MIMO systems using sphere decoding*, *IEEE Wireless Communications Letters* **6** (2016), no. 2 158–161.

[67] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, *Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing*, *IEEE Journal of Selected Topics in Signal Processing* **8** (2014), no. 5 902–915.

[68] M. A. Albreem, M. Juntti, and S. Shahabuddin, *Massive MIMO detection techniques: a survey*, *IEEE Communications Surveys & Tutorials* **21** (2019), no. 4 3109–3132.

[69] M. E. Rasekh, Z. Marzi, Y. Zhu, U. Madhow, and H. Zheng, *Noncoherent mmWave path tracking*, in *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, pp. 13–18, 2017.

[70] K. A. Alnajjar, P. J. Smith, and G. K. Woodward, *Low complexity V-BLAST for massive MIMO*, in *2014 Australian Communications Theory Workshop (AusCTW)*, pp. 22–26, IEEE, 2014.

[71] L. Fang, L. Xu, and D. D. Huang, *Low complexity iterative MMSE-PIC detection for medium-size massive MIMO*, *IEEE Wireless Communications Letters* **5** (2015), no. 1 108–111.

[72] K. Pham and K. Lee, *Low-complexity SIC detection algorithms for multiple-input multiple-output systems*, *IEEE Transactions on Signal Processing* **63** (2015), no. 17 4625–4633.

[73] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, *Transmit beamforming and power control for cellular wireless systems*, *IEEE Journal on Selected Areas in Communications* **16** (Oct, 1998) 1437–1450.

[74] E. Visotsky and U. Madhow, *Optimum beamforming using transmit antenna arrays*, in *1999 IEEE 49th Vehicular Technology Conference (Cat. No.99CH36363)*, vol. 1, pp. 851–856 vol.1, May, 1999.

[75] M. Schubert and H. Boche, *Solution of the multiuser downlink beamforming problem with individual SINR constraints*, *IEEE Transactions on Vehicular Technology* **53** (Jan, 2004) 18–28.

[76] C. W. Tan, M. Chiang, and R. Srikant, *Maximizing sum rate and minimizing MSE on multiuser downlink: Optimality, fast algorithms and equivalence via max-min SINR*, *IEEE Transactions on Signal Processing* **59** (2011), no. 12 6127–6143.

[77] H. V. Poor and S. Verdú, *Probability of error in MMSE multiuser detection*, *IEEE transactions on Information theory* **43** (1997), no. 3 858–871.

[78] D. Simic and P. Reynaert, *A 14.8 dBm 20.3 dB power amplifier for D-band applications in 40 nm CMOS*, in *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pp. 232–235, June, 2018.

[79] T. B. Reed, M. Rodwell, Z. Griffith, P. Rowell, A. Young, M. Urteaga, and M. Field, *A 220 GHz InP HBT solid-state power amplifier MMIC with 90mW POUT at 8.2dB compressed gain*, in *2012 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, pp. 1–4, Oct, 2012.

[80] D. Zhu, J. Choi, and R. W. Heath, *Two-dimensional AoD and AoA acquisition for wideband millimeter-wave systems with dual-polarized MIMO*, *IEEE Transactions on Wireless Communications* **16** (2017), no. 12 7890–7905.

[81] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, *Uplink performance of wideband massive MIMO with one-bit ADCs*, *IEEE Transactions on Wireless Communications* **16** (2017), no. 1 87–100.

[82] M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, *High-throughput data detection for massive MU-MIMO-OFDM using coordinate descent*, *IEEE Transactions on Circuits and Systems I: Regular Papers* **63** (2016), no. 12 2357–2367.

[83] J. Coon and M. Beach, *An investigation of MIMO single-carrier frequency-domain MMSE equalization*, in *London Communications Symposium*, pp. 237–240, 9, 2002.

[84] X. Zhu and R. D. Murch, *Novel frequency-domain equalization architectures for a single-carrier wireless MIMO system*, in *Proceedings IEEE 56th Vehicular Technology Conference*, vol. 2, pp. 874–878, 2002.

[85] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. J. W. Rodwell, *Phase noise analysis for mmWave massive MIMO: a design framework for scaling via tiled architectures*, in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, Mar, 2019.

[86] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. Rodwell, *Phase noise in modular millimeter wave massive mimo*, *IEEE Transactions on Wireless Communications* (2021).

[87] A. A. Farid, M. Abdelghany, U. Madhow, and M. J. Rodwell, *Dynamic range requirements of digital vs. rf and tiled beamforming in mm-wave massive mimo*, in *2021 IEEE Radio and Wireless Symposium (RWS)*, pp. 46–48, IEEE, 2021.

[88] S. H. Mirfarshbafan, S. Taner, and C. Studer, *Smul-fft: A streaming multiplierless fast fourier transform*, *IEEE Transactions on Circuits and Systems II: Express Briefs* **68** (2021), no. 5 1715–1719.

[89] M. A. El-Motaz, O. A. Nasr, and K. Osama, *A cordic-friendly fft architecture*, in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 1087–1092, IEEE, 2014.

[90] D. Simic and P. Reynaert, *A 14.8 dBm 20.3 dB power amplifier for D-band applications in 40 nm CMOS*, in *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, IEEE, 2018.

[91] S. Lloyd, *Least squares quantization in PCM*, *IEEE transactions on information theory* **28** (1982), no. 2.

[92] J. Max, *Quantizing for minimum distortion*, *IRE Transactions on Information Theory* **6** (1960), no. 1.