Embracing Nonlinearities in Future Wireless Communication Systems : Low-Resolution Analog-to-Digital Converters

Khodor Safa

université PARIS-SACLAY

Embracing Nonlinearities in Future Wireless Communication Systems: Low-Resolution Analog-to-Digital Converters

Etudes des Non-linéarités aux Réseaux sans Fil du Futur: Convertisseurs Analogique-Numérique avec une Faible Résolution

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Sciences des Réseaux, de l'Information et de la Communication
Graduate School : Informatique et sciences du numérique
Référent : CentraleSupélec

Thèse préparée dans l'unité de recherche **Laboratoire des signaux et systèmes (Université Paris-Saclay, CNRS, CentraleSupélec)** , sous la direction de **Sheng YANG**, Professeur, le co-encadrement de **Raul DE LACERDA**, Maître de Conférence

Thèse soutenue à Paris-Saclay, le 25 Novembre 2024, par

Khodor SAFA

Composition du jury

Membres du jury avec voix délibérative

Jean-Philippe OVARLEZ Directeur de Recherche, ONERA	Président
Philippe CIBLAT Professeur, Télécom Paris	Rapporteur & Examinateur
Inbar FIJALKOW Professeure, ENSEA, CNRS	Rapporteure & Examinatrice
Maxime GUILLAUD Directeur de Recherche, INRIA Lyon	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASG082

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Embracing Nonlinearities in Future Wireless Communication Systems: Low-resolution Analog-to-Digital Converters.....

Keywords: nonlinearities, wireless communications, analog-to-digital converters, MIMO, maximum likelihood detection, channel estimation

Abstract: With the advent of 5G networks and the road towards 6G already being established, innovative wireless communication technologies including massive multiple-input multiple-output (mMIMO) and millimeter-wave systems are emerging to address the ever-increasing number of mobile users and to support new industry applications. However, higher frequencies and wider bandwidths lead to increased power consumption in radio-frequency (RF) circuits, necessitating more energy-efficient components. At the same time, systems are increasingly susceptible to nonlinear impairments such as phase noise, saturation, and quantization distortions. Understanding the impact of these nonlinearities on transceiver design and fundamental limits becomes essential. This thesis focuses on the nonlinear effects of low-resolution analog-to-digital converters (ADCs) at the receiver. ADC power consumption increases with both bandwidth and resolution, making low-resolution ADCs a practical solution in systems like mMIMO, where power consumption is a major constraint. The first part of this work examines data detection in quantized flat-fading MIMO channels, with various assumptions about the channel state information (CSI). Maximum likelihood (ML) detection is optimal for minimizing the error probability but is computationally expensive. While sphere-decoding (SD) algorithms are commonly used to reduce complexity in unquantized channels, they are not directly applicable to the quantized case due to the discrete nature of observations. To address this, we propose a two-step low-complexity detection algorithm for systems with one-bit comparators. This method approximates the ML metric using a Taylor series and converts the detection problem into a classical integer least-squares optimization, allowing the use of SD algorithms. Numerical results show this approach achieves near-ML performance. This method is also extended to multi-bit scenarios, converging to conventional SD as resolution increases. In more practical scenarios, where only statistical CSI at the receiver (CSIR) is available, we explore data detection under a pilot transmission scheme. The first approach frames channel estimation and detection as binary classification tasks using probit regression. Achievable mismatched rates using the generalized mutual information are then evaluated in comparison to the Bussgang linear minimum mean-square error estimators where it's shown that the performance depends on the choice of the estimator. The second approach jointly processes data and pilot sequences where we encounter challenges related to evaluating multivariate Gaussian probabilities and the combinatorial complexity of optimization. We address the first challenge with the Laplace method that provides us with a closed-form expression, while for the complexity we adapt the SD technique from the perfect CSI case on a surrogate metric. It is of interest for future wireless systems to obtain design guidelines that can accurately explain the trade-off between spectral efficiency and energy consumption. We then investigate the capacity of quantized MIMO channels, which is difficult to characterize due to their discrete nature. Therefore, assuming an asymptotic regime where the number of receive antennas grows large and employing known information-theoretic results from Bayesian statistics on reference priors, the capacity scaling can be characterized for the coherent multi-bit case providing us with an expression that can be used for the analysis of the spectral efficiency and power consumption balance. For the noncoherent case, we apply the same results at the unquantized channel as an upper bound for the quantized channel and identify upper and lower bounds on the scaling for particular values on the coherence interval.

Titre: Etudes des Non-linéarités aux Réseaux sans Fil du Futur: Convertisseurs Analogique-Numérique avec une Faible Résolution.....

Mots clés: non-linéarités, résaux sans fils, convertisseurs analogique-numérique, MIMO, détection par maximum de vraisemblance, estimation du canal

Résumé: Avec l'avancement des communications sans fil vers la 5G et la 6G, de nouveaux défis émergent en raison de l'augmentation des utilisateurs et des applications industrielles. Des technologies comme le massive multiple-input multiple-output (mMIMO) et les systèmes à ondes millimétriques sont développées pour répondre à ces besoins. Cependant, des fréquences et des largeurs de plus élevées entraînent une consommation d'énergie accrue dans les circuits radiofréquences (RF), nécessitant des composants plus économes. Parallèlement, les systèmes deviennent sensibles aux non-linéarités, telles que le bruit de phase et les distorsions de quantification. Comprendre l'impact de ces non-linéarités sur la conception des émetteurs-récepteurs et les limites fondamentales devient essentiel. Cette thèse se concentre sur les effets non linéaires des convertisseurs analogique-numérique (CAN) à faible résolution au récepteur. La consommation d'énergie des CAN augmente avec la largeur de bande et la résolution, rendant les CAN à faible résolution pratiques dans des systèmes comme le mMIMO, où la consommation est cruciale. La première partie examine la détection de données dans des canaux MIMO à évanouissement plat quantifié, avec différentes hypothèses sur l'information de l'état du canal (IEC). La détection par maximum de vraisemblance (MV) est optimale pour minimiser les erreurs, mais elle est coûteuse en calculs. Les algorithmes de décodage sphérique (DS) réduisent la complexité dans les canaux non quantifiés, mais ne s'appliquent pas directement aux canaux quantifiés. Pour y remédier, nous proposons un algorithme de détection à faible complexité en deux étapes pour les systèmes à un bit. Cette méthode utilise une approximation de la métrique MV via une série de Taylor et transforme le problème de détection en optimisation des moindres carrés entiers, permettant d'utiliser les algorithmes de DS. Les résultats montrent que cette approche atteint des performances proches de celles de ML. La méthode est également étendue aux scénarios multi-bits, convergeant vers le DS classique avec une résolution accrue. Dans des scénarios plus pratiques, où seule l'information statistique sur l'état du canal au récepteur (ISECR) est disponible, nous explorons la détection sous un schéma de transmission pilote. La première approche considère l'estimation du canal et la détection comme des tâches de classification binaire. Les taux réalisables en utilisant l'information mutuelle généralisée sont comparés aux estimateurs de Bussgang où on trouve que la performance dépend de l'estimateur choisi. La seconde approche traite conjointement les séquences de données et de pilotes en rencontrant des défis d'évaluation des probabilités gaussiennes et de complexité combinatoire. Le premier défi est résolu par la méthode de Laplace, et pour la complexité, nous adaptons la technique DS du cas avec IEC parfait à une métrique de substitution. Il est crucial d'obtenir des directives de conception pour les futurs systèmes sans fil afin d'expliquer le compromis entre efficacité spectrale et consommation d'énergie. Nous examinons ensuite la capacité des canaux MIMO quantifiés, difficile à caractériser en raison de leur nature discrète. Par conséquent, en supposant un régime asymptotique où le nombre d'antennes de réception augmente et en utilisant des résultats théoriques bien connus de la statistique bayésienne sur les prioris de référence, l'échelonnement de la capacité peut être caractérisé pour le cas multi-bits cohérent, fournissant une expression utile pour l'analyse de l'efficacité spectrale et de la consommation d'énergie. Pour le cas non cohérent, nous appliquons les mêmes résultats au canal non quantifié comme borne supérieure pour le canal quantifié et identifions des bornes supérieures et inférieures sur l'échelonnement pour certaines valeurs de l'intervalle de cohérence.

Synthèse en Français

À mesure que les systèmes de communication évoluent vers de nouvelles générations, les défis au niveau des circuits radiofréquenceS (RF) deviennent de plus en plus importants. Les considérations pratiques concernant la consommation d'énergie au niveau du front-end du récepteur sont particulièrement critiques, notamment avec l'adoption de largeurs de bande plus élevées et d'architectures numériques plus complexes. Pour les systèmes au-delà de la 5G et la 6G, l'objectif est d'atteindre des débits de transmission élevés tout en minimisant la consommation d'énergie des composants RF individuels. Une observation clé est que la consommation énergétique des convertisseurs analogique-numérique (CAN) augmente de manière exponentielle avec la résolution et linéairement avec la fréquence d'échantillonnage. Cela motive l'utilisation d'alternatives à faible résolution, au prix de techniques de traitement du signal plus sophistiquées.

On examine le modèle système du canal multiple-input multiple-output (MIMO) quantifié sous une architecture numérique et des hypothèses de canal à évanouissement plat. Le modèle mathématique est simplifié en considérant une synchronisation parfaite et un traitement analogique, facilitant ainsi la conception et l'évaluation des algorithmes d'estimation de canal et de détection des données. L'étude passe en revue les techniques de pointe pour ces défis, notamment les récepteurs linéaires exploitant la décomposition de Bussgang et les méthodes probabilistes qui intègrent le modèle statistique global du système de communication.

La première contribution de ce travail aborde la détection des données dans les canaux MIMO quantifiés à un bit. En utilisant une interprétation géométrique de la fonction de vraisemblance, il est démontré que l'algorithme décodage sphérique (DS) conventionnel peut être adapté via une approximation du second ordre. Contrairement aux méthodes heuristiques proposées dans la littérature, cette approche utilise l'information de l'état du canal (IEC) disponible pour construire une liste de candidats via la fonction Hessienne. Les expériences numériques montrent que la méthode proposée atteint des performances vector error rates (VER) quasi-optimales par rapport à une borne inférieure sur le maximum de vraisemblance (MV), pour diverses configurations MIMO et tailles de constellations. L'impact de la corrélation spatiale entre les antennes de réception est également considéré, révélant une dégradation générale des performances VER. Néanmoins, l'algorithme conserve une

quasi-optimalité par rapport aux performances MV. De plus, une étude détaillée de la complexité computationnelle de l'algorithme DS met en évidence qu'un choix judicieux de la taille de pas réduit les itérations nécessaires lors de la descente de gradient dans la première étape de l'algorithme.

Ensuite, nous examinons le problème de détection des données lorsque le IEC est indisponible au niveau du récepteur. Deux schémas sont explorés dans des canaux Rayleigh à évanouissement par bloc avec des trames pilotes et de données indépendantes. Le premier utilise un processus explicite en deux étapes impliquant une estimation du canal et une détection des données basées sur le cadre de régression probit. Des matrices pilotes unitaires sont utilisées pour estimer le canal, qui remplace ensuite le canal réel lors de la détection. Des comparaisons sont effectuées avec l'estimation de canal basée sur la décomposition de Bussgang, montrant que les performances VER s'améliorent avec des séquences d'entraînement plus longues pour les signaux QPSK et 16-QAM, en particulier à un haut rapport signal sur bruit (RSB). L'instabilité observée dans la CDF gaussienne sous IEC imparfait est attribuée au choix de l'estimateur plutôt qu'aux limites inhérentes de la CDF. En outre, les débits atteignables avec des métriques sous-optimales sont analysés à l'aide de la GMI. Les résultats montrent que des séquences d'entraînement plus longues permettent d'approcher les débits de communication adaptés avec IEC parfait.

Le second schéma traite conjointement les trames pilotes et de données. Pour un modèle de canal réel, la métrique de détection optimale introduit deux défis: l'évaluation des probabilités orthantes gaussiennes multivariées sans formes fermées et la complexité prohibitive de détection pour des entrées discrètes en grandes dimensions. Ces problèmes sont résolus par une approximation Laplace pour les probabilités orthantes gaussiennes, produisant une expression en forme fermée. L'algorithme DS est ensuite appliqué à cette approximation via une métrique MV de substitution sous-optimale. Pour un signal 4-PAM, la métrique LA se rapproche de la borne inférieure de la métrique optimale exacte, avec des performances qui convergent vers le cas cohérent à mesure que la longueur d'entraînement augmente.

Enfin, cette thèse explore les travaux en cours et futurs visant à étendre ces résultats aux CAN multi-bits à faible résolution et à analyser le compromis entre la consommation énergétique et la dégradation de l'efficacité spectrale au niveau du front-end du récepteur. L'algorithme DS est adapté pour la quantification multi-bits, atteignant des performances VER quasi-optimales par rapport aux bornes inférieures MV. À mesure que la résolution augmente, l'algorithme converge vers le cas DS

conventionnel. En utilisant des asymptotiques théoriques de l'information issues des statistiques bayésiennes, c'est possible de caractériser l'échelle de capacité quand le nombre de récepteurs augmente. Pour les canaux MIMO quantifiés multi-bits cohérents, l'extension du cas un bit au multi-bit fournit une expression de la capacité asymptotique. Les travaux futurs incluent l'intégration de modèles physiques de la consommation énergétique des composants RF avec les expressions de capacité. Pour les canaux non cohérents non quantifiée, la capacité asymptotique sert de borne supérieure dans certains intervalles de cohérence, l'intégration sur l'espace des paramètres produisant généralement des bornes supérieures et inférieures sur l'échelle.

Les recherches futures exploreront également les techniques de suréchantillonnage pour les CAN un bit, qui ont démontré une amélioration des performances. Le problème de précodage de canal en direction descendante avec des convertisseur numérique-analogique à faible résolution au niveau de l'émetteur constitue une autre voie prometteuse. En outre, l'étude d'autres sources de non-linéarité, telles que la saturation des amplificateurs de puissance et le bruit de phase dans les oscillateurs locaux, pourrait fournir des informations plus approfondies sur les limites de performance.

Acknowledgements

As a prelude, I would like to note that these last three years of my PhD journey have been truly valuable. During my stay at the Laboratoire des Signaux et Systèmes at CentraleSupélec, Université Paris-Saclay, I have learned to grow both as a researcher and a person. I had the opportunity to attend conferences where I could interact with experts in my field, exchange ideas, and expand my perspective. There are no words that can fully describe this unique experience, however, it's evident to say that it fosters a person's resilience, rigor, and commitment.

First of all, I would like to thank my supervisors Prof. Sheng Yang and Prof. Raul De Lacerda for the opportunity to work on this exciting topic. They have provided guidance throughout this journey and set challenges for me in order to learn and grow as a researcher, and for that, I am truly grateful.

I would also like to offer my gratitude to my thesis defense jury members Prof. Jean-Philippe Ovarlez, Prof. Inbar Fijalkow, Prof. Philippe Ciblat, and Prof. Maxime Guillaud. I would like to thank them for reviewing my manuscript and offering interesting perspectives for future potential contributions to this topic. I want to also thank Dr. Richard Combes for being part as an invitee and for his valuable insights and discussions during my PhD.

My acknowledgments also go to my lab colleagues and friends Henrique Miyamoto, Raymond Zhang, Elissa Mehanna, and many others with whom I had a wonderful time which made this experience much more enjoyable. I will dearly miss those afternoons discussing each other's work, learning valuable new ideas, and solving new problems on the board.

I would also like to thank the lab administration and personnel. With special mention to Mme. Stéphanie Douesnard and Mme. Catherine Kilic for their help in the organization of the thesis defense, and Prof. Armelle Wautier for opening the opportunity to contribute to teaching courses for the engineering students at the university.

Last but not least my warmest gratitude goes to my family and closest friends who have always been there to support me during this incredible journey.

Contents

List of Figures

List of Acronyms

ADC	analog-to-digital converter
AQNM	additive quantization noise model
AWGN	additive white Gaussian noise
BLMMSE	Bussgang linear minimum mean square error
CDF	cumulative density function
CSCG	circularly symmetric complex Gaussian
CSI	channel state information
CSIR	channel state information at the receiver
DAC	digital-to-analog converter
DFT	discrete-fourier-transform
ENOB	effective number of bits
FoM	figure of merit
GMI	generalized mutual information
HPA	high-power amplifier
LA	laplace approximation
LMMSE	linear minimum mean square error
LNA	low-noise amplifier
LO	local oscillator
MAP	maximum a posteriori
MC	Monte-Carlo
MI	mutual information
MIMO	multiple-input multiple-output
ML	maximum likelihood
mMIMO	massive multiple-input multiple-output
mmWave	millimeter-wave
MRC	maximum ratio combining
MSE	mean square error
NLZF	nonlinear zero-forcing
OFDM	orthogonal frequency division multiplexing
PA	power amplifier
PAM	pulse amplitude modulation
QAM	quadrature-amplitude modulation
QPSK	quadrature phase-shift keying
RF	radio-frequency
SD	sphere-decoding
SER	symbol error rate

SIMO	single-input multiple-output
SISO	single-input single-output
SNDR	signal-to-noise and distortion
SNR	signal-to-noise ratio
VER	vector error rate
VS	Volterra series
ZF	zero-forcing

Chapter 1

Introduction

The evolution of wireless communication systems has been mainly driven by the emergence of new applications, technologies, and the fundamental need to convey information efficiently and reliably. Since its conception and commercialization in the 1980s, the first generation (1G) cellular network has enabled voice communication through analog technology. Its digital successor, the Global System for Mobile Communications (GSM), or 2G, provided improved data rates and supported text messaging applications such as the Short Message Service (SMS). Beginning in the early 2000s, multimedia communication further encompassed video, photo, and reliable Internet access with the advent of 3G and 4G mobile networks. Throughout the last decade, the 5G standard was established to address the challenges presented by the exponential increase in the number of mobile devices and the explosion in data traffic, through enhanced mobile broadband, massive machine-type communications, and ultra-reliable low latency [1]. The discussion on what 6G represents is already ongoing and it is expected to provide orders of magnitude in enhancements of prior generations' key performance indicators, as well as incorporate revolutionary new use cases and applications such as multisensory extended reality and fully autonomous connected robotics and systems [2, 3]. The advancement towards each generation comes with its own set of challenges that require a re-thinking of standards, and motivate the development of innovative and enabling technologies [4]. For example, 5G networks saw the deployment of massive multiple-input multiple-output (mMIMO) and millimeter-wave (mmWave) to address problems of spectrum scarcity and spectral efficiency [5]. Frequency bands in the sub-terahertz regime are also being explored as viable candidates for 6G systems. It's expected that in the 6G era, there will be extensive deployment of non-terrestrial networks including unmanned aerial vehicles and satellites along with plans to integrate them with cellular systems infrastructure to ensure global coverage and reliability [6]. Indeed, interest in satellite communication is being rekindled, and the development of high throughput satellites will play a prominent role in fulfilling growing and exigent requirements driven by new trends in spectrum usage, energy efficiency, and ubiquitous connectivity [7, 8].

1.1 . Background and Motivation

Adopting these new technologies necessitates, in general, a closer study of the overall system since direct implementation of current architectures and strategies might not be as straightforward. The main common aspects of the technology trends mentioned above are the increase in bandwidth, higher carrier frequencies, and energy-efficient constraints that have primary roles in the design of radio-frequency (RF) circuitry. An RF architecture of a communication system typically includes components such as power amplifiers (PAs), local oscillators (LOs), and analog-to-digital converters (ADCs). The performance of individual components can depend on the intended application and operating conditions. Subsequently, nonlinear characteristics and distortions can manifest due to the design choices and inevitable imperfections present in the circuit. We begin by briefly surveying the different kinds of nonlinearities in RF components to get a better overview of their characteristics and the incurred challenges.

1.1.1 . RF Impairments and Nonlinearities

A critical component in a wireless communication system is the ADC that converts analog signals to the digital domain, whether for storage purposes or efficient implementation of signal processing algorithms. The function of an ADC can be generally decomposed into two fundamental operations: sampling the input signal and then quantizing the result into a finite set of values. The sampling operation is conducted in the sample-and-hold circuitry susceptible to aperture noise. This impairment is attributed to noise present in the sampling circuit itself and in addition to phase noise affecting the clock that generates the sampling signal [9–11]. Given an input $x(t)$, aperture noise can be formally defined as the sample-to-sample uncertainty and can be related to the delay in the switching circuit. In other words, assuming a sampling period T_s and a time instant nT_s, the sampled output is effectively

$$z(nT_s) = x(nT_s + \varepsilon(nT_s)), \tag{1.1}$$

where $\varepsilon(nT_s)$ is a random process that can be assumed to be Gaussian, but more intricately depends on the contribution from both sampling circuit noise errors and the clock phase noise model [12, 13]. If the perturbation is small, a first-order Taylor approximation of the noise effect can be written in a discrete model representation as

$$\begin{aligned} z[n] &= x[n] + \varepsilon[n] \frac{\mathrm{d}x(t)}{\mathrm{d}t}\bigg|_{t=nT_s} \\ &= x[n] + e[n], \end{aligned} \tag{1.2}$$

where the jitter term $e[n]$ is seen to be dependent on the signal's rate of change. The effect of aperture noise can be severe, especially in wideband applications where high-frequency components of the input signal can contribute to large deviations from the sampled value. As we will elaborate upon later, aperture noise is seen as the limiting factor in the design of high-speed and high-resolution ADCs [14, 15]. Quantization distortion is due to the fundamental limitation of representing continuous-valued inputs with high precision according to a fixed number of bits [16, 17]. The resolution is specified by the ADC manufacturer and measured in terms of the signal-to-noise and distortion (SNDR) which can vary depending on the RF architecture and design choices. The quantization can be seen as a partition of the real line $\mathbb{R}$ with a set of intervals $\mathcal{I} = \{\mathcal{I}_k = (I_{k-1}, I_k]\}$ for $k = 1, \ldots, 2^b$ where b is the resolution. The output $y[n]$ of the quantizer given an input $z[n]$ is assigned to y_k if $z[n] \in \mathcal{I}_k$. The distortion error due to quantization can then be defined as $e[n] = y[n] - z[n]$. Note that the quantization stage can also be subject to integral and differential nonlinearities that result in deviation from the specified quantizer thresholds [18, 19].

We highlight other nonlinear impairments that can be present, for example, phase noise can affect other critical components of communication systems including LOs responsible for up- and down-conversion of signals during transmission [20, 21]. The power of an ideal sinusoidal signal oscillator is concentrated at a particular value, but due to circuit imperfections, the power is going to be spread over a larger range of values. Models of the phase noise generating process obtained analytically and from measurements show that, in general, the phase noise variance can grow quadratically with carrier frequency [20]. Power amplifiers are power-hungry devices that exhibit increased signal distortion close to the saturation point where power efficiency is high. To motivate the nonlinearities present in PAs, we turn our attention to the transmitter. The analog signal is passed through a high-power amplifier (HPA) before being transmitted over the channel, this is to ensure coverage and mitigate losses due to the propagation environment. The Friis formula for the free space path loss (FSPL) explicitly shows this relationship for electromagnetic signals

$$\text{FSPL} = \left(\frac{4\pi d f_c}{c}\right)^2, \tag{1.3}$$

where c is the speed of light, f_c is the carrier frequency and d is the distance between transmit and receive antennas assumed to have unit directivity gains. The relation essentially tells us that the received power degrades with the square of the signal frequency, which becomes more

detrimental as we move to mmWave and sub-THz bands [6, 8]. This power degradation can be compensated by employing highly directive transmit antennas [5] or operating the HPA close to its saturation region, at the cost of pushing the amplifier towards the nonlinear regime causing the transmitted signal to be distorted and resulting in in-band and out-of-band intermodulation interference [22]. The latter can usually be eliminated by filtering the signal after the amplification stage if it lies outside the bandwidth of interest. The PA is a nonlinear and bandpass device that operates on real signals

$$\begin{aligned} x(t) &= \Re\{\tilde{x}(t)e^{j2\pi f_c t}\} \\ &= |A(t)|\cos(2\pi f_c t + \phi(t)), \end{aligned} \tag{1.4}$$

where $\tilde{x}(t) = A(t)e^{j\phi(t)}$ is the complex envelope of the signal. The device is generally represented by its amplitude-modulation to amplitude-modulation (AM-AM) and amplitude-modulation to phase-modulation (AM-PM) conversion characteristics

$$y(t) = G(|\tilde{x}(t)|)\cos(2\pi f_c t + \phi(t) + \Psi(|\tilde{x}(t)|)) \tag{1.5}$$

where the functions $G(\cdot)$ and $\Psi(\cdot)$ represent the AM-AM and AM-PM characteristics respectively, assuming no memory in the system. Denoting the nonlinear operation by $\mathsf{V}[\cdot]$, the most general representation of the transfer function, also incorporating memory effects, is the Volterra series (VS) [23, 24]

$$\begin{aligned} y(t) &= \mathsf{V}\left[x(t)\right] \\ &= \sum_{n=0}^{\infty} \int_0^{\infty} \ldots \int_0^{\infty} h_n(\tau_1, \ldots, \tau_n) \prod_{i=1}^{n} x(t-\tau_i)\, \mathrm{d}\tau_i, \end{aligned} \tag{1.6}$$

where we have the nth order Volterra kernels h_n and time delays τ_i. The device is assumed to be causal, i.e., h_n is 0 for all $\tau_i < 0$, and symmetric in its delays without any loss of generality. The VS can be regarded as a generalization of Taylor series for dynamical systems with memory. In fact, it was shown in [24] that, under certain assumptions of fading memory, any nonlinear system can have a Volterra series kernel representation. Moreover, it can be considered as the most general for PAs in the sense that it captures other well-known polynomial models in the literature for different classes including strictly and quasi-memoryless PAs [25, 26]. To reduce the complexity of this model, the memory is assumed to be truncated up to a finite order. There exist other models in the literature that characterize the AM-AM and AM-PM characteristics depending on the architecture, e.g., the Saleh and Rapp models

were developed to emulate the behavior of traveling wave tube and solid-state power amplifiers, respectively, which can prove useful if a specific application is envisioned [27].

1.1.2 . RF Nonlinearities Challenges

The induced nonlinearities in the system can have adverse effects on the information transmission performance, especially when specific requirements on power consumption are needed. For example, to exploit higher energy efficiencies provided by PAs, the design of waveforms (e.g. discrete Fourier transform-frequency division multiple access) and signal processing techniques (e.g. crest factor reduction) that promote a low peak-to-average power ratio is one approach, as it allows operating the HPA at a low output power back-off while keeping the distortion at a minimum [28,29]. Another approach can include linearizing the transmitted signal through digital pre-distortion techniques at the transmitter by designing inverse functions cascaded with the PA nonlinearity [23, 30]. Predistortion can add more complexity to the transmitter since additional operations are needed to learn the nonlinearity coefficients which can also result in higher power consumption. Alternatively one can adopt nonlinear equalization techniques at the receiver [31–33]. Nonlinearities induced by HPAs are particularly interesting for satellite communication applications where power can be scarce and low complexity solutions processing schemes are crucial [7, 34]. Regarding aperture jitter, compensation techniques have also been proposed in the literature, including models based on VS representation [13, 35]. The effect of quantization noise on receiver detection schemes has also received significant attention, especially for low-resolution cases where the effect is most severe [36–40]. On another note, understanding the fundamental limitations in terms of the spectral efficiency is important in the design of communication systems, which can become more challenging when incorporating nonlinearity effects. Studying the mutual information with HPA distortion at the transmitter depends on the assumed model and whether the coefficients are known a priori. For example, achievable rates with amplifier nonlinearities have been studied in [41] using the VS model and random coding arguments, and in [42,43] for multiple-input multiple-output (MIMO) systems assuming the Saleh and polynomial models and Gaussian signaling. Similarly for the quantized channel, the capacity of the quantized additive white Gaussian noise (AWGN) channel has been characterized for the one-bit single-input single-output (SISO) case in [44], and the achievable rates for the MIMO channel with fading have been studied in [45, 46].

Energy efficiency is becoming an added dimension in wireless systems design [47, 48], and more importantly from the receiver point of view

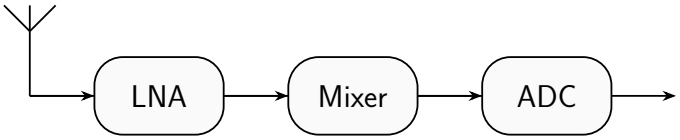

Figure 1.1: Simplified RF frontend component structure composed of a receive antenna, LNA, a mixer for downconversion including oscillators and an ADC.

where an increased overall system complexity raises important practical questions on feasibility in terms of costs and power consumption [49]. Consider the simplified receiver architecture as shown in Fig 1.1 composed mainly of an antenna, a low-noise amplifier (LNA), a mixer, and an ADC [38]. Recent surveys and roadmaps demonstrate that, in general, as frequency increases, power consumption grows for LNAs and ADCs [14, 15, 50]. Moreover, nonlinear distortions arise if higher power efficiency is required, manifesting as intermodulation distortion in LNAs, phase noise in the mixer and ADC sampling circuitry, and quantization errors if low-resolution ADCs are employed. Regarding the imposed challenges presented earlier, it is of interest to study the communication system in this *nonlinear* paradigm. However, a joint study of the overall combined effects can be challenging given the different facets of the problem that can depend on the technology, application, and requirements. We focus hereafter on the ADC component of the RF chain.

1.2 . Analog-to-Digital Converters

As in any RF circuit containing active or passive components, noise sources would arise due to unavoidable imperfections whose effects can vary according to hardware designs. Nevertheless, the performance of each ADC architecture can be summarized and evaluated according to different metrics, or figure of merits (FoMs) that depend on sampling frequency, resolution, and power consumption. The most notable survey that first explored trends of ADC technology was conducted in 1999 by Walden [14]. The results therein show a tradeoff between resolution and the increase in sampling frequency which is attributed to the dominant effects of aperture noise present in the sampling circuitry. Walden's proposed FoM F_W, relates the power dissipation P to the bandwidth B and resolution b[1],

$$P = F_W \cdot B \cdot 2^b, \tag{1.7}$$

[1]In practice, b is the effective number of bits (ENOB), which is calculated from SNDR measurements as ENOB $= (\text{SNDR} - 1.62)/6.02$. This metric can be lower than the stated number of bits.

and is used to analyze the trends in performance. Moreover, the author stipulated that future design trends in ADC technologies will continue to compromise between speed and resolution for the design of power-efficient converters. The work of Le et al. in 2005 [51] extends the analysis for ADC data up to seven years further and refutes his pessimistic point of view; their findings indicate an exponential increase in the trend towards the design of fast, power-efficient and high-resolution ADC technology. Their results show that certain architectures are more relevant for specific applications than others; *sigma-delta* converters are practical for high-resolution applications with lower bandwidth requirements whereas *flash* ADCs are more useful for high-speed and low-resolution scenarios. More recently, Murmann re-explores this topic in his 2015 article [15] which includes data from more modern designs. They show that ADC designs with moderate to high resolutions exhibit a quadrupling in power consumption for every bit increase, whereas for low-resolution converters the Walden FoM remains a valid approximation, i.e., power consumption only doubles with every increasing bit of resolution. This behavior is demonstrated in Fig 1.2 reproduced from [52] which shows data collected for state-of-the-art ADC designs reported in the IEEE International Solid-State Circuit Conferences (ISSCC) and VLSI Circuits Symposium, where we also show the Schreier FoM F_S that captures the quadrupling effect in the power for higher resolutions. Murmann also re-emphasizes the speed-resolution design limitation due to aperture noise as predicted in [14].

In light of the previous discussion, it becomes evident that while, on the one hand, there have been significant technological advancements in ADC designs that can potentially accommodate the increase in bandwidths for future wireless technologies, there remain, on the other hand, important practical considerations. For example, Eq. (1.7) shows that power consumption grows linearly with the bandwidth and exponentially in the resolution which can become prohibitive. In addition, considering a fully digital architecture for a mMIMO system where the number of antenna elements can reach the order of thousands, the use of low-quality RF components becomes imperative to deter high costs [47, 48, 51, 53]. For this purpose, the use of low-resolution ADCs becomes attractive in terms of power and cost, and consequently, there will be newly imposed challenges on various classical aspects regarding channel estimation, data detection, and capacity computation which we survey in the following section. In this thesis, our focus will be on the ADC component which constitutes the bottleneck in terms of energy consumption and information transfer for baseband processing, especially for mmWave, sub-THz, and mMIMO systems. Incorporating the contribution of nonlinearities

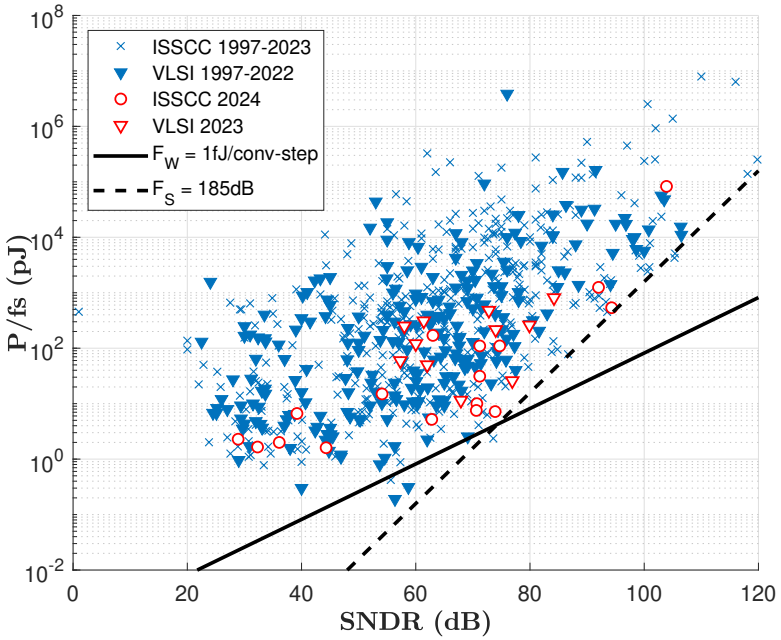

Figure 1.2: ADC energy performance survey data collected from IEEE ISSCC and VLSI Symposium.

due to amplification and phase noise is an interesting direction for future work that merits a detailed study.

1.2.1 . Challenges and Related Work for Low-Resolution ADCs

Due to the severe quantization, the application of classical data detection, channel estimation, and signal processing techniques is not a straightforward feat as we enter the nonlinearity paradigm. Extensive research was dedicated in the early 2000s to exploring the effects of quantized channels on system performance, with ultra-wideband communications being a main drive in that direction [51, 54, 55]. The fundamental capacity limits of quantized single-input single-output (SISO) real additive white Gaussian noise channels were studied in [56] and [44], the authors proved that discrete input distributions are capacity-achieving and binary phase shift keying is optimal for the special case of one-bit quantization. Under Rayleigh fading with statistical channel state information at the receiver (CSIR), the rate expression is known [57] and on-off quadrature phase-shift keying (QPSK) is optimal. Moreover, theoretical work on channel and parameter estimation for the SISO channel also considered *dithering* to improve estimation performance [58–60]. In general, the use of dither consists of adding artificial noise before quantization in order to shape the statistical distribution of the quantization

error [61], this can result in useful signal processing techniques that are amenable to analysis [62]. Under coarse quantization, MIMO systems can be beneficial when exploiting the system's spatial oversampling. One of the earliest works on this topic considered the problems of channel estimation using an expectation-maximization framework for quantized signals and pilot design [63], where it was demonstrated that the performance in terms of mean-square error is dependent on the type of pilot signals used [64]. Finally, the capacity degradation with one-bit quantization is a factor of $2/\pi$ compared to the ideal case, at least in the low signal-to-noise ratio (SNR) regime and with symmetric quantizers [55]. This loss can be further compensated with asymmetric quantizers [65], or in some cases correlated additive noise [66].

With the prospect of mmWave and mMIMO for future wireless systems, major challenges arose with increased system complexity due to nonlinearities, resulting in intractable expressions when it comes to performance evaluation and receiver design. To this end, research on low-resolution ADCs gained further traction in later years focusing on studying capacity bounds, linear approximations, and machine learning methods. Mo and Heath [45] derive lower and upper bounds on the one-bit MIMO deterministic channel at finite SNR assuming perfect CSI at the transmitter and receiver. They demonstrate that channel inversion precoding with QPSK signaling is capacity-achieving when the channel matrix is full row rank and semi-unitary. Furthermore, the authors in [67] expand the mutual information up to second-order in the low SNR regime and show that QPSK is uniquely optimal for deterministic channels, and capacity achieving for the ergodic case with CSIR. Li et al. [68] leverage the Bussgang decomposition to linearize the flat-fading Rayleigh one-bit channel, and derive lower bound approximations in the low SNR regime on the achievable rates. The authors also propose the Bussgang linear minimum mean square error (BLMMSE) estimator for the channel and extend classical linear data detectors, e.g., maximum ratio combining (MRC) and zero-forcing (ZF) for the quantized channel. These results were generalized to the multi-bit case in [46]. Assuming a pilot training scheme and one-bit transceivers, Gao et al. [69] provide an algorithm to compute a lower bound on the capacity using the replica method in the asymptotic regime for large-scale channels. They show that the training length may be set smaller than the number of users in the system. Considering practical channel estimation and data detection methods, Wen et al. [70] present a joint channel and data detection algorithm for low-resolution ADCs based on generalized approximate message passing (GAMP) which yields best estimation results in the Bayesian sense, albeit at a high computational complexity. Choi et al. [71] propose a two-stage

data detection and channel estimation algorithms based on projected gradient descent which constrains the norm of the estimated signal. Similarly, for the orthogonal frequency division multiplexing (OFDM) MIMO channel, Studer and Durisi [72] formulate both tasks into convex optimization problems that are solved using the forward-backward splitting algorithm. Mollen et al. [73] linearize the wide-band quantized channel and consider the regime where the number of taps goes to infinity. Wan et al. [74] generalize the BLMMSE estimator by optimizing over the quantizer threshold to reduce the mean square error (MSE), this results in improved performance compared to fixed zero-thresholding. The authors in [75] study in detail the performance with the Bussgang decomposition for the class of least-squares estimators in single-input multiple-output (SIMO) systems, they highlight on the *stochastic resonance* effect [76] and show that the optimal SNR for channel estimation decreases with increasing pilot length. In addition, a series of recent work considers machine learning tools to address receiver design complexity [40, 77–81]. There is also a plethora of contributions on low-resolution ADCs that explore other directions such as mixed architectures, oversampling and transmit precoding [37, 82–85].

1.3 . Main Contributions and Organization

In this thesis, we consider the communication scenario in a fully digital MIMO channel with low-resolution ADCs at each receive antenna. This channel model also assumes perfect synchronization and analog matched-filtering [46, 73]. In Chapter 2, we review relevant state-of-the-art techniques that address this problem considering different channel models: linear techniques based on the Bussgang decomposition and additive quantization noise model (AQNM) that linearize the channel, and probabilistic channel models that treat the likelihood function. In Chapter 3, we look at the problem of data detection with perfect CSIR in the extreme case where only a one-bit of resolution is available. It is well known that ML detection is optimal in the sense of minimizing the probability of error. However, the optimization is a hard combinatorial problem when the input is discrete since the size of the search space to consider grows exponentially with the number of users in the system and modulation order. The usual approach is first to minimize the ML objective function in the continuous domain and then, either map into the discrete constellation or construct a list of candidate points based on specific heuristics where the original ML function is eventually evaluated [40, 71, 79, 81]. Our contribution in this aspect is as follows:

- We propose a data detection algorithm that first obtains an initial

estimate using first-order methods. The ML metric is then expanded in Taylor series form up to second-order around this point, and the level sets are exploited to define an ellipsoid containing the list of candidate points. This formulation allows us to recover the classical integer least-squares problem where efficient sphere-decoding (SD) algorithms can be readily employed.

- The algorithm greatly reduces the search space of candidate points, and is shown to be near-optimal with respect to the ML oracle lower bound method proposed in [86] in terms of vector error rates (VERs).

A discussion on the computational complexity is also provided. Finally, the algorithm is evaluated in a typical mMIMO channel exhibiting spatial correlation of the receive antennas at the base station. In Chapter 4, we consider a more realistic scenario where no channel state information (CSI) is available a priori at the receiver and adopt a pilot transmission strategy. In the first part and under the framework of probit regression due to the similarity with the channel model, we assume a pilot transmission scheme that first estimates the channel parameters and then performs data detection according to these estimates. We assess the performance numerically by plotting the symbol error rates (SERs) and the generalized mutual information (GMI) for the mismatched achievable rates under the different metrics. Using this formulation, we address the instability highlighted in [40, 79, 87] for the probit model under imperfect CSI and show that it's effectively related to the choice of the estimator. In the second part, we focus on a real channel for simplicity and formulate the optimal metric that processes the pilot and data signals jointly. Computing the ML metric directly requires averaging over the channel statistics which do not have a closed and tractable form. In addition, it is still a hard combinatorial problem with respect to the discrete transmitted signal. Our contribution in this direction is then as follows:

- Given the difficulty of directly averaging over the channel statistics, we reformulate the optimal metric such that averaging is conducted over the channel's posterior distribution conditioned on pilot data. Integrating with respect to this distribution is still as difficult given its non-Gaussianity. Therefore, we resort to approximating it using the Laplace method [88], this allows the evaluation of the likelihood in closed-form.

- To address the combinatorial aspect of the problem, we leverage the SD algorithm presented in the previous part. Application of this algorithm to the approximate metric is not straightforward given

the data dependence of the parameters. Therefore, we propose a workaround where we employ a surrogate metric over which the SD algorithm can be applied and construct a list of candidate points over which we evaluate the proposed approximation.

In Chapter 5, we look at the communication problem when low-resolution ADCs are employed at the receiver side. We show that the SD algorithm can be extended to the multi-bit case. Then, we focus our attention on understanding the trade-off between RF power consumption and spectral efficiency degradation due to quantization distortion. As a first step in that direction, we look at the capacity problem of the quantized MIMO channel. In the coherent case and for a multi-bit quantizer, it's possible to obtain the scaling of the capacity in the asymptotic regime where the number of receive antennas grows large. This is done using information-theoretic asymptotic results in Bayesian statistics. For the noncoherent case, we begin by looking at the asymptotic capacity of the unquantized case as an upper bound for particular cases of the coherence interval length. In general, obtaining exact expressions of the capacity scaling is not straightforward due to the difficulty in integrating over the parameter space. We conclude in Chapter 6 with final remarks on areas of improvement, future perspectives, and potential research directions.

1.4 . List of Publications

The content of this thesis can be found in our previously published work

- K. Safa, R. De Lacerda, and S. Yang, "Channel Estimation and Data Detection in MIMO channels with 1-bit ADC using Probit Regression," *2023 IEEE Information Theory Workshop (ITW)*. Saint-Malo, France: IEEE, Apr. 2023, pp. 457–461.

- K. Safa, R. Combes, R. de Lacerda and S. Yang, "Data Detection in 1-bit Quantized MIMO Systems," *IEEE Transactions on Communications*, vol. 72, no. 9, pp. 5396-5410, Sept. 2024

The following is ongoing work and future perspectives

- K. Safa, R. De Lacerda, and S. Yang, "Capacity of Block Fading Low-Resolution MIMO Channel: Asymptotics and Bounds", *work in progress*

- K. Safa, R. De Lacerda, and S. Yang, "Spectral and Energy Efficient System Design in Low-Resolution ADCs in MIMO Channels", *work in progress*

Chapter 2

System Model and Receiver Design

Due to the nonlinearity introduced by the quantizer, classical techniques for channel estimation and data detection do not apply directly and there are several approaches to address these difficulties. In this chapter, we elaborate on the system model assumed throughout the thesis along with the general assumptions for the quantized MIMO channel. We present the relevant state-of-the-art work that addresses the problem of data detection under quantizer nonlinearities. The mathematical notations adopted are as follows:

Notation

Lowercase letters refer to scalars, e.g., x. Column vectors and matrices are given in boldface lowercase and capital letters respectively, e.g., $\boldsymbol{x}$ and $\boldsymbol{A}$. The tilde accent is used when referring to complex versions of these quantities, e.g., $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{A}}$. We denote the transpose and hermitian transpose operators as $\boldsymbol{A}^{\mathsf{T}}$ and $\tilde{\boldsymbol{A}}^{\mathsf{H}}$, respectively. The operator $\mathsf{diag}(\cdot)$ returns a diagonal matrix with the diagonal elements of its matrix input. The real and imaginary part operators are given by, respectively, $\Re\{\cdot\}$ and $\Im\{\cdot\}$ and are applied component-wise if the input is a vector or matrix. The natural and base 2 logarithms are denoted with ln and log, respectively. Sets are represented in calligraphic font, e.g., $\mathcal{A}$. Identity matrices of dimension n are given by $\boldsymbol{I}_n$. The notation $||.||$ represents the ℓ_2 norm of a vector. The spectral norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ is defined as $||\boldsymbol{A}||_2 = \sup_{\boldsymbol{x}\neq 0} \frac{||\boldsymbol{A}\boldsymbol{x}||}{||\boldsymbol{x}||}$. The standard big O and little o notations are $O(\cdot)$ and $o(\cdot)$, respectively. The standard normal distribution is denoted by $\phi(x)$, and its cumulative density function (CDF) by $\Phi(x)$.

2.1 . System Model

We consider the MIMO transmission model with M transmit antennas and N receivers at the base station where each receive antenna is equipped with a b-bit quantizer as shown in Fig. 2.1. The received signal is

$$\tilde{\boldsymbol{y}} = \tilde{\mathsf{Q}}_b(\tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{z}}), \tag{2.1}$$

where we have:

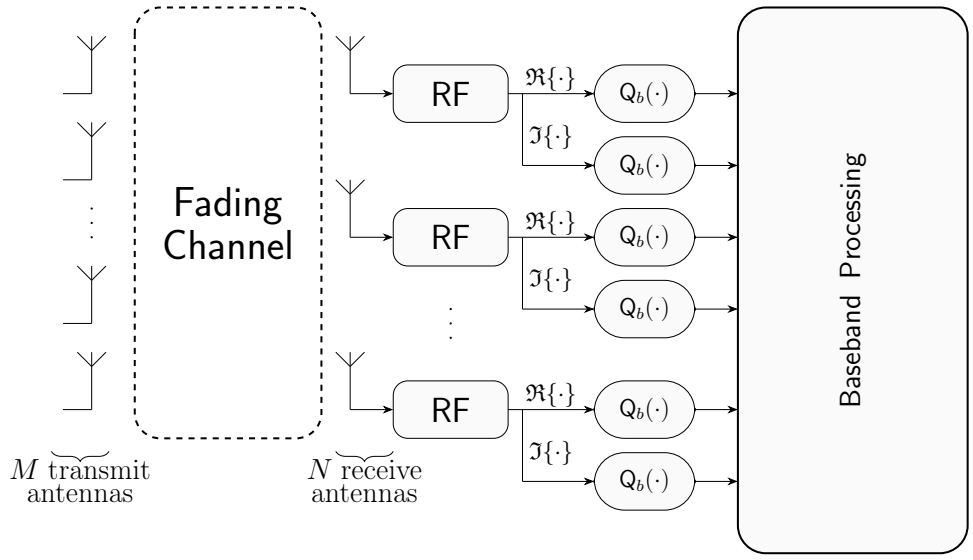

Figure 2.1: MIMO fading channel with M transmit and N receive antennas with quantizers Q_b assigned to each in-phase and quadrature-phase component of the signal.

- The flat-fading channel matrix $\tilde{\boldsymbol{H}} \in \mathbb{C}^{N\times M}$ with zero-mean, unit variance, independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian (CSCG) components $\tilde{h}_{n,m} \sim \mathcal{CN}(0,1)$.

- The transmitted signal $\tilde{\boldsymbol{x}}$, an M-dimensinal complex vector where each element belongs to a complex constellation denoted as $\tilde{\mathcal{X}}$, e.g., QPSK or 16-quadrature-amplitude modulation (QAM).

- Additive noise vector $\tilde{\boldsymbol{z}} \in \mathbb{C}^N$ that is assumed independent of the input and channel, with i.i.d. circularly symmetric complex Gaussian elements having variance $\tilde{\sigma}^2$. In our setting, we define the SNR as $\tilde{\rho} = 1/\tilde{\sigma}^2$.

- A uniform midriser quantizer $\tilde{\mathsf{Q}}_b$ with resolution b and spacing δ that operates on each real and complex dimension of the received signal independently, i.e., given a complex number $\tilde{r}$, $\tilde{\mathsf{Q}}_b = \mathsf{Q}_b(\tilde{r}) + j\mathsf{Q}_b(\tilde{r})$ with $\mathsf{Q}_b(\cdot)$ defined as below.

- The quantized received signal $\tilde{\boldsymbol{y}}$ with elements belonging to a finite set of complex numbers $\tilde{\mathcal{Y}}_b$.

We will also find it convenient for simplicity of notation and derivations to work with the real-equivalent form of this channel that can be decom-

posed as

$$\begin{bmatrix}\Re\{\tilde{\boldsymbol{y}}\}\\ \Im\{\tilde{\boldsymbol{y}}\}\end{bmatrix} = \mathsf{Q}_b\left(\begin{bmatrix}\Re\{\tilde{\boldsymbol{H}}\} & -\Im\{\tilde{\boldsymbol{H}}\}\\ \Im\{\tilde{\boldsymbol{H}}\} & \Re\{\tilde{\boldsymbol{H}}\}\end{bmatrix}\begin{bmatrix}\Re\{\tilde{\boldsymbol{x}}\}\\ \Im\{\tilde{\boldsymbol{x}}\}\end{bmatrix} + \begin{bmatrix}\Re\{\tilde{\boldsymbol{z}}\}\\ \Im\{\tilde{\boldsymbol{z}}\}\end{bmatrix}\right),$$
$$\boldsymbol{y}_{2N} = \mathsf{Q}_b(\boldsymbol{H}_{2N\times 2M}\boldsymbol{x}_{2M} + \boldsymbol{z}_{2N}), \tag{2.2}$$

with the subscripts explicitly shown here to denote the adjusted new dimensions. The effective noise variance is then $\sigma^2 = \tilde{\sigma}^2/2$, the extended M-dimensional constellation is now $\mathcal{X}^{2M}$ such that $\mathcal{X}$ represents the real counterpart of $\tilde{\mathcal{X}}$. The quantizer Q_b operates element-wise and is defined as the map from $\mathbb{R}$ to a finite set of values $\mathcal{Y}_b = \{\nu_1, \ldots, \nu_l, \ldots, \nu_{2^b}\}$ associated with the set of intervals $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_l, \ldots, \mathcal{I}_{2^b}\}$,

$$\nu_l = \delta\left(l - \frac{1}{2} - 2^{b-1}\right), \quad l = 1, 2, \ldots, 2^b. \tag{2.3}$$

Since we assume a uniform quantizer, the spacing for each interval $\mathcal{I}_l = (I_{l-1}, I_l]$ is $I_l - I_{l-1} = \delta$ for $l = 2, \ldots, 2^b - 1$, and we define $I_0 = -\infty$ and $I_{2^b} = \infty$. The points in $\mathcal{Y}_b$ are taken as the midpoints of each corresponding interval[1]. In the extreme case where $b = 1$, we have a simple zero-threshold comparator $\tilde{\mathsf{Q}}_1(\tilde{r}) = \mathrm{sgn}(\Re\{\tilde{r}\}) + j\mathrm{sgn}(\Im\{\tilde{r}\})$ that outputs the sign of the input signal

$$\mathrm{sgn}(a) = \begin{cases} 1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$

Remark: The discrete-time channel model assumes that conventional digital signal processing (DSP) operations such as down-conversion and matched filtering are conducted in the analog domain, in addition to having perfect synchronization between the transmitter and receiver. In particular, to the low-resolution ADCs literature, designing receivers with analog processors is an interesting research topic [89, 90]. Similarly, for the synchronization requirement, there have been some proposed algorithms in the literature [39, 91], and a practical system demonstrator is proposed in [92] for analog synchronization in a one-bit ADC system.

The nonlinear nature of the quantized channel imposes several difficulties in retrieving the transmitted signal $\tilde{\boldsymbol{x}}$. In the classical setting where the receiver has access to the signal assuming infinite precision and that the channel matrix $\tilde{\boldsymbol{H}}$ is perfectly known, the optimal metric under Gaussian additive noise is the ML receiver equivalent to the nearest neighbor detector under square loss. This is no longer the case due

[1]More precisely, on the boundaries where saturation can occur, these points are effectively half the step size δ apart from the finite valued thresholds I_1 and I_{2^b-1}.

to the discrete nature of the observations. This becomes more challenging in a more practical setting when the channel is not known to either the transmitter or receiver. In this case, assuming the channel is block-fading over a coherence interval of length T, the receiver can form an estimate of the channel and use it in place of the true channel realization during the data detection process. More generally, in a situation where only statistical information is available, non-coherent detection can be an option [93]. The transmitter then allocates a portion of the coherence interval to transmit pilot information T_p and the remaining duration T_d is used for data detection. The transmission is then split into two stages with the subscripts p and d referring to pilot and data respectively

$$\left[\tilde{\boldsymbol{Y}}_p \; \tilde{\boldsymbol{Y}}_d\right] = \tilde{\mathrm{Q}}_b\Big(\tilde{\boldsymbol{H}}\left[\tilde{\boldsymbol{X}}_p \; \tilde{\boldsymbol{X}}_d\right] + \left[\tilde{\boldsymbol{Z}}_p \; \tilde{\boldsymbol{Z}}_d\right]\Big). \tag{2.4}$$

We assume the pilot transmission matrix $\tilde{\boldsymbol{X}}_p$ is unitary which can be achieved, for example, with a truncated discrete-fourier-transfor (DFT) matrix or the identity matrix [37, 79, 94].

One of the most common techniques to deal with quantization is to linearize the channel based on the Bussgang decomposition [95], in addition to the AQNM [96, 97] allowing the design of low-complexity linear channel estimation and data detection receivers [37, 94]. Other methods include maximizing the likelihood function induced by this channel which can result in a better performance albeit at a higher computational complexity [70, 71, 81].

2.2 . Channel Linearization

2.2.1 . Bussgang Decomposition

The Bussgang decomposition is a useful technique that allows writing the output of a deterministic function as a sum of its Gaussian input multiplied by a constant term referred to as the Bussgang gain, and a distortion term that is uncorrelated with the input and any other jointly Gaussian random variable [66, 75, 98]. Given an input u that follows a CSCG distribution and a deterministic function $f(\cdot)$, the output $v = f(u)$ can be equivalently expressed as

$$\begin{aligned} v &= w \cdot u + e \\ &= \mathbb{E}\left[ru^*\right]\mathbb{E}\left[|u|^2\right]^{-1} \cdot u + e. \end{aligned} \tag{2.5}$$

In fact, this decomposition can be seen as the LMMSE estimate of the output v given the input u, where e is the distortion error term that is uncorrelated with u [98]. Constraining the problem to the class of linear

estimators, we would like to find the unbiased estimate $\hat{v}$ such that

$$\hat{v} = w \cdot u + b, \tag{2.6}$$

which implies $b = \mathbb{E}[\hat{v}] - w\mathbb{E}[u]$. For simplicity and without loss of generality we can therefore assume the input is zero-mean and set $b = 0$. By the orthogonality principle, the estimator should be the best among all linear estimators in the mean square error sense, i.e., the error term should be uncorrelated with the input $\mathbb{E}[(v - \hat{v})u^*] = 0$. Expanding the latter we obtain the expression of the linear scale α in Eq. (2.5). We can proceed similarly for the vector case assuming $f(\cdot)$ operates element-wise on the input vector to obtain

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e} \\ &= \boldsymbol{C}_{\boldsymbol{vu}}\boldsymbol{C}_{\boldsymbol{u}}^{-1}\boldsymbol{u} + \boldsymbol{e}, \end{aligned} \tag{2.7}$$

where $\boldsymbol{C}_{\boldsymbol{ru}} = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{u}^{\mathsf{H}}\right]$ is the cross-covariance between the input and the output and $\boldsymbol{C}_{\boldsymbol{u}} = \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^{\mathsf{H}}\right]$ is the auto-covariance of the input. We can also obtain an expression of the distortion term auto-covariance as

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{e}} &= \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathsf{H}}\right] \\ &= \mathbb{E}\left[(\boldsymbol{v} - \boldsymbol{W}\boldsymbol{u})(\boldsymbol{v} - \boldsymbol{W}\boldsymbol{u})^{\mathsf{H}}\right] \\ &= \boldsymbol{C}_{\boldsymbol{v}} - \boldsymbol{C}_{\boldsymbol{vu}}\boldsymbol{C}_{\boldsymbol{v}}^{-1}\boldsymbol{C}_{\boldsymbol{vu}}^{\mathsf{H}} \end{aligned} \tag{2.8}$$

We can apply these results directly to our channel model in Eq. (2.1) for the multi-bit quantizer $\mathsf{Q}_b(\cdot)$. Denote the input before quantization by $\tilde{\boldsymbol{r}} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{z}}$ and $\boldsymbol{C}_{\tilde{\boldsymbol{r}}}$ its autocovariance matrix, the linear representation of Eq. (2.1) is

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \boldsymbol{W}_b\tilde{\boldsymbol{r}} + \boldsymbol{e} \\ &= \boldsymbol{W}_b\tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \boldsymbol{W}_b\tilde{\boldsymbol{z}} + \boldsymbol{e}, \end{aligned} \tag{2.9}$$

where the Bussgain gain matrix can be written as

$$\begin{aligned} \boldsymbol{W}_b &= \boldsymbol{C}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{r}}}\boldsymbol{C}_{\tilde{\boldsymbol{r}}}^{-1} \\ &= \mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}} \sum_{i=1}^{2^b} \frac{\nu_i}{\sqrt{\pi}} \left(e^{-I_{i-1}^2 \mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-1}} - e^{-I_i^2 \mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-1}} \right). \end{aligned} \tag{2.10}$$

A detailed derivation of this result can be found in Appendix A where we use Price's theorem [99] [98] to compute $\boldsymbol{C}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{r}}}$ that highlights the assumptions required to obtain this form which includes mainly the Gaussianity of the input. When $b = 1$, we have

$$\boldsymbol{W}_1 = \sqrt{\frac{4}{\pi}} \mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}. \tag{2.11}$$

This decomposition allows the design of simple linear receivers akin to the classical channel without quantization. We can also obtain an expression of $\boldsymbol{C}_{\tilde{\boldsymbol{y}}}$ similarly by direct application of Price's theorem

$$\begin{aligned}\boldsymbol{C}_{\boldsymbol{e}} =& \frac{4}{\pi}\Bigg[\arcsin\Big(\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\Re\{\boldsymbol{C}_{\tilde{\boldsymbol{r}}}\}\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\Big) \\ &+ j\arcsin\Big(\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\Im\{\boldsymbol{C}_{\tilde{\boldsymbol{r}}}\}\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\Big) \\ &- \mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\boldsymbol{C}_{\tilde{\boldsymbol{r}}}\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})^{-\frac{1}{2}}\Bigg]. \end{aligned} \tag{2.12}$$

2.2.2 . Additive Quantization Noise Model

Another common method found in the literature to represent the quantized channel is with the AQNM [39, 66, 97, 100]. It can be seen as a special case of the Bussgang decomposition when the quantizer function is specified. In particular, it mainly relies on the fact that $\mathsf{Q}_b(\cdot)$ is designed in the MMSE sense [17, 101–103] such that the output can be written

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{r}} + \tilde{\boldsymbol{q}}, \tag{2.13}$$

where $\tilde{\boldsymbol{q}}$ is the quantizer distortion noise per receive element in $\tilde{\boldsymbol{r}}$. Denoting β as the distortion factor assigned per receive element in $\tilde{\boldsymbol{y}}$ and defining it as the ratio

$$\beta = \frac{\mathbb{E}\left[|q_i|^2\right]}{\tilde{\sigma}_i^2}, \tag{2.14}$$

then proceeding similarly as in Eq. (2.7) we obtain the following expression [40, 97]

$$\tilde{\boldsymbol{y}} = (1-\beta)\tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + (1-\beta)\tilde{\boldsymbol{n}} + \tilde{\boldsymbol{n}}_q. \tag{2.15}$$

The result is obtained by following the reasoning in [66, 96] which reduces the cross-covariance matrix to

$$\boldsymbol{C}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{r}}} = (1-\rho_q)\boldsymbol{C}_{\tilde{\boldsymbol{r}}}. \tag{2.16}$$

Note that the Bussgang decomposition gives a more general expression as it does not require any assumption on the choice of the quantizer itself and does not make any approximations for the quantizer distortion noise $\tilde{\boldsymbol{n}}_q$. There are other linearization variants of the Bussgang decomposition. For example, in [74] the authors propose a generalized version for the one-bit channel that considers a non-zero threshold. More recently, [104] proposed a novel linearization of the channel that depends on a second-order Hermite polynomial expansion of the quantization function.

2.2.3 . Linear Receivers

The linearized channel allows the design of low-complexity linear receivers similar to those of a classical Gaussian channel which include ZF, MRC, and LMMSE equalizers. Inspecting Eq. (2.9) obtained with the Bussgang decomposition, we have the following equivalent channel and additive noise effects

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{G}}_b \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}}. \tag{2.17}$$

Assuming perfect CSI, let $\boldsymbol{F}$ denote the receiver combining filter such that the transmitted signal estimate is obtained as [40, 46]

$$\hat{\boldsymbol{x}} = \boldsymbol{F}\tilde{\boldsymbol{y}}. \tag{2.18}$$

This estimate is usually re-scaled to match the input signal energy before forcing its components to the discrete values in $\tilde{\mathcal{X}}$. Based on this equivalent channel, we obtain the Bussgang MRC (BMRC)

$$\boldsymbol{F}_{\text{BMRC}} = \mathsf{diag}\left(\tilde{\boldsymbol{G}}_b^{\mathsf{H}} \tilde{\boldsymbol{G}}_b\right)^{-1} \tilde{\boldsymbol{G}}_b^{\mathsf{H}}, \tag{2.19}$$

the Bussgang ZF receiver

$$\boldsymbol{F}_{\text{BZF}} = \left(\tilde{\boldsymbol{G}}_b^{\mathsf{H}} \tilde{\boldsymbol{G}}_b\right)^{-1} \tilde{\boldsymbol{G}}_b^{\mathsf{H}}, \tag{2.20}$$

and the Bussgang MMSE (BMMSE)

$$\boldsymbol{F}_{\text{BMMSE}} = \tilde{\boldsymbol{G}}_b^{\mathsf{H}} \boldsymbol{C}_{\tilde{\boldsymbol{y}}}^{-1}. \tag{2.21}$$

In the special case of one-bit quantizers, we have for the BMMSE

$$\boldsymbol{F}_{\text{BMMSE}} = \tilde{\boldsymbol{G}}_1^{\mathsf{H}} \left(\tilde{\boldsymbol{G}}_1 \tilde{\boldsymbol{G}}_1^{\mathsf{H}} + \boldsymbol{C}_{\boldsymbol{e}}\right)^{-1}, \tag{2.22}$$

with $\boldsymbol{C}_{\boldsymbol{e}}$ obtained according to Eq. (2.12). When the channel matrix needs to be estimated, referring to the pilot transmission strategy in Eq. (2.4) and focusing on the one-bit case, we can also obtain the Bussgang LMMSE of the channel [46, 68]

$$\hat{\boldsymbol{H}} = \frac{\alpha}{2 - \frac{4}{\pi} + \tilde{\sigma}^2\alpha^2 + T_p\alpha^2} \tilde{\boldsymbol{Y}}_p \tilde{\boldsymbol{X}}_p^{\mathsf{H}}. \tag{2.23}$$

where $\alpha = \sqrt{\frac{4}{\pi(M+\tilde{\sigma}^2)}}$. This estimator is computed assuming unitary pilot transmission matrices and that the input covariance matrix is diagonally dominant. It's important to note that the distortion term in Eq. 2.9 is not necessarily Gaussian, and although it's uncorrelated with the input by

the orthogonality principle, they remain dependent. Therefore, analyzing the optimality in the MSE sense of these receivers in general is not as well-studied as for the Gaussian case. The work in [75] considers a more analytical and in-depth look at this problem for the SIMO channel. For the AQNM, we can also derive the Wiener filter [40, 96]

$$\boldsymbol{F}_{\text{WF}} = \tilde{\boldsymbol{H}}^{\mathsf{H}}((1-\beta)\boldsymbol{C}_{\tilde{\boldsymbol{r}}} + \beta\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}}))^{-1}. \tag{2.24}$$

This filter is obtained under the assumption of an MMSE quantizer design, in addition to approximating the output covariance matrix as

$$\boldsymbol{C}_{\tilde{\boldsymbol{y}}} \approx (1-\beta)(\boldsymbol{C}_{\tilde{\boldsymbol{r}}} + \beta\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}})) \tag{2.25}$$

which gives the following expression for the distortion term covariance

$$\boldsymbol{C}_{\tilde{\boldsymbol{n}}_q} \approx \beta(1-\beta)\mathsf{diag}(\boldsymbol{C}_{\tilde{\boldsymbol{r}}}). \tag{2.26}$$

2.3 . Statistical Model Based Methods

2.3.1 . Maximum Likelihood Detection with Perfect CSIR

An alternative approach to studying the design of receivers is through optimizing the likelihood function of the channel. Considering the data detection problem in a perfect CSIR situation, we first map each element in the received vector $\boldsymbol{y}_n$ to its associated index for the corresponding interval, i.e., from $y_n \in \mathcal{Y}_b$ to $\mathsf{y}_n \in \{1, 2, \ldots, l, \ldots, 2^b\}$. The likelihood function for the received vector $\boldsymbol{y}$, given a uniformly transmitted vector $\boldsymbol{x}$ and channel $\boldsymbol{H}$ can then be written as

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) &= \prod_{n=1}^{2N} \boldsymbol{P}(r_n \in \mathcal{I}_{\mathsf{y}_n}|\boldsymbol{h}_n,\boldsymbol{x}) \\ &= \prod_{n=1}^{2N}\left[\Phi\left(\frac{I_{\mathsf{y}_n} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right) - \Phi\left(\frac{I_{\mathsf{y}_n-1} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\right].\end{aligned} \tag{2.27}$$

This decomposition is valid under the assumption that the noise covariance matrix is diagonal. Note that the likelihood implicitly depends on the resolution b and step size δ. This can be written directly for the one-bit case with $y_n \in \{-1, 1\}$

$$\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) = \prod_{n=1}^{2N}\Phi\left(\frac{y_n\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right). \tag{2.28}$$

These expressions are based on the real-equivalent form of the channel in Eq. (2.2). The maximum likelihood detector seeks to maximize the

log of the likelihood function in Eq. (2.27), or equivalently, minimize its negative

$$\begin{aligned}\boldsymbol{x}^* &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad -\ln\left[\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{H},\boldsymbol{x})\right] \\ &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad -\sum_{n=1}^{2N}\ln\left[\Phi\left(\frac{I_{\mathsf{y}_n}-\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)-\Phi\left(\frac{I_{\mathsf{y}_n-1}-\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\right] \\ &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad \ell_b(\boldsymbol{x}). \end{aligned} \tag{2.29}$$

Even when the channel is known, given the discrete nature of the input signal, computing this metric has a complexity that grows exponentially in the number of users and constellation size as it requires to search over a space of size $|\mathcal{X}|^{2M}$. There are several solutions presented in the literature to address the optimization problem in Eq. (2.29) with a particular focus on the one-bit case. One approach is to relax the discrete constraint from $\mathcal{X}^{2M}$ to $\mathbb{R}^{2M}$ as in [71] where the authors propose a two-stage near ML detector. In the first stage, the metric

$$\underset{\substack{\boldsymbol{x}\in\mathbb{R}^{2M}\\ ||\boldsymbol{x}||^2\leq M}}{\arg\min} -\sum_{n=1}^{2N}\ln\left[\Phi\left(\frac{y_n\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\right] \tag{2.30}$$

is computed using projected gradient descent to satisfy the inequality constraint, then normalized after the last iteration to match the input signal energy to obtain $\dot{\boldsymbol{x}}_{\mathrm{nML}}$. This estimate is eventually mapped back into the complex constellation symbol by symbol to match the discrete constraint and retrieve $\ddot{\boldsymbol{x}}_{\mathrm{nML}}$. Note that this procedure is still sub-optimal given the above relaxation. In order to improve the performance, the authors propose a second stage where a constant $\chi > 1$ is fixed, and the following preliminary sets are constructed for each transmit antenna at index m

$$\mathcal{C}_m = \left\{ x \in \tilde{\mathcal{X}} \,\middle|\, \frac{|\dot{x}_m - x|}{|\dot{x}_m - \ddot{x}_m|} < \chi \right\}, \tag{2.31}$$

and then the final candidate set is constructed as the Cartesian product $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_M$, such that

$$\mathcal{C} = \left\{ \check{\boldsymbol{x}} = [\check{x}_1, \ldots, \check{x}_m, \ldots, \check{x}_M]^{\mathsf{T}} \,\middle|\, \check{x}_m \in \mathcal{C}_m, \forall m \right\}. \tag{2.32}$$

The nML solution is finally obtained as

$$\boldsymbol{x}_{\mathrm{nML}} = \underset{\boldsymbol{x}\in\mathcal{C}}{\arg\max}\ \ell_1(\boldsymbol{x}). \tag{2.33}$$

We can see that this technique, although similar to the heuristic of SD, is only based on the Euclidean distance from an initial estimate and does not take into account the effects of the fading channel matrix, as is usually the case. Note that, in addition, the radius χ has to be carefully tuned so that the size of $\mathcal{C}$ is not too large.

The authors in [81] propose a *one-bit sphere-decoding* algorithm to compute Eq. (2.28). The idea is based on converting the metric into a minimum weighted-hamming distance detection problem by first enumerating a codebook that contains all binary output vectors that correspond to noise-free measurements of the possible input vectors in $\mathcal{X}^{2M}$, i.e., construct

$$\mathcal{C}_W = \left\{ \boldsymbol{c} = \operatorname{sgn}(\boldsymbol{H}\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}^{2M} \right\}, \tag{2.34}$$

then the ML metric for the one-bit case is re-written as

$$\begin{aligned} \boldsymbol{x}^* &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad \mathsf{d}_w(\boldsymbol{y}, \boldsymbol{c}; \boldsymbol{w}, \tilde{\boldsymbol{w}}) \\ &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad \sum_{n=1}^{2N} \left(w_i \mathbb{1}_{\{y_n \neq c_n\}} + \tilde{w}_n \mathbb{1}_{\{y_n = c_n\}} \right), \end{aligned} \tag{2.35}$$

where the weights $w_n = -\ln Q\left(\frac{-y_n \boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{x}}{\sigma}\right)$, $\tilde{w}_n = -\ln\left[1 - Q\left(\frac{-y_n \boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{x}}{\sigma}\right)\right]$, and $Q(x)$ represents the Gaussian Q-function. Essentially, as a pre-processing stage, the authors consider the problem of building the discrete channel transition matrix. This can be prohibitive as the size of the output vector grows exponentially in N, which is usually taken to be much larger than M. To circumvent this limitation, the authors consider partitioning the output vector and the codewords for the noise-free codebook $\mathcal{C}_W$ into G sub-vectors of size $N_s = 2N/G$. For each sub-vector, the authors construct a sub-list of size L having the closest sub-codeword to the sub-vector in terms of the metric in Eq. (2.35). Now, given a received signal y, the same partition is applied as before and the overall list of codewords or transmitted signals is constructed as the union of the sub-lists corresponding to the partition of the received vector. The final estimate is computed as the maximum of the proposed metric over this sublist. The authors use in addition an approximation to the Q-function to minimize the computational complexity of the metric. Note that this method does not exactly reduce the size of the input search space as the pre-processing stage requires enumerating all noise-free output binary vectors for each transmitted input in $\mathcal{X}^{2M}$.

2.3.2 . Data Detection with Imperfect CSI

In the case where we need to obtain an estimate of the channel according to the pilot transmission scheme Eq. (2.4), we can look at the

posterior distribution of the channel matrix given the pilot data. To obtain a concise expression, we re-write directly the real-equivalent channel model

$$\begin{aligned} \boldsymbol{Y}_p &= \mathsf{Q}_b(\boldsymbol{X}_p\bar{\boldsymbol{H}} + \boldsymbol{Z_p}), \\ \begin{bmatrix} \Re\{\tilde{\boldsymbol{Y}}_p\}^\mathsf{T} \\ \Im\{\tilde{\boldsymbol{Y}}_p\}^\mathsf{T} \end{bmatrix} &= \mathsf{Q}_b\Bigg(\begin{bmatrix} \Re\{\tilde{\boldsymbol{X}}_p\} & \Im\{\tilde{\boldsymbol{X}}_p\} \\ -\Im\{\tilde{\boldsymbol{X}}_p\} & \Re\{\tilde{\boldsymbol{X}}_p\} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Re\{\tilde{\boldsymbol{H}}\}^\mathsf{T} \\ \Im\{\tilde{\boldsymbol{H}}\}^\mathsf{T} \end{bmatrix} + \begin{bmatrix} \Re\{\tilde{\boldsymbol{Z}}_p\}^\mathsf{T} \\ \Im\{\tilde{\boldsymbol{Z}}_p\}^\mathsf{T} \end{bmatrix} \Bigg) \end{aligned} \tag{2.36}$$

and the posterior distribution of the real-equivalent channel is therefore given by

$$\begin{aligned} \boldsymbol{P}\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p, \boldsymbol{X}_p\big) &\propto \boldsymbol{P}\big(\boldsymbol{Y}_p|\bar{\boldsymbol{H}}, \boldsymbol{X}_p\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{H}}\big) \\ &= \prod_{n=1}^{N} \prod_{t=1}^{2T_p} \Phi\left(\frac{y_{p,n}^t \boldsymbol{x}_{p,t}^\mathsf{T} \bar{\boldsymbol{h}}_n}{\sigma_p} \right) \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big). \end{aligned} \tag{2.37}$$

The authors in [71] propose an ML estimator similar in structure to their data detector that essentially ignores the prior distribution on the channel. Considering frequency-selective channels and OFDM transmission, the authors in [72] formulate the channel estimation and data detection problems for the multi-bit quantizer under a statistical approximation of the quantized channel likelihood function that assumes a linear model with a Gaussian additive distortion term that is independent of the input with variance equal to the conditional MSE for each quantization label. Mainly, the channel is first estimated using the maximum a posteriori (MAP) estimator and used for the detection process through a convex relaxation to the input space which is then solved using a forward-backward splitting algorithm. A similar approach is adopted in [105] where the data detection is conducted after averaging over the channel estimation errors. The authors in [70], propose a joint channel estimation and data detection procedure that is based on GAMP to compute the best estimates of $\bar{\boldsymbol{H}}$ and $\boldsymbol{X}_d$ jointly in the MSE sense. In particular, for the one-bit scenario, recent work in the literature considers nonlinear machine learning methods to solve the data detection problem. For example, the work in [79] reformulates the detection into a support vector machine (SVM) binary classification problem to retrieve an estimate of the channel used subsequently for data detection under the same formulation. Moreover, the same authors propose the OBMNet architecture in [40] which approximates the Gaussian CDF with a sigmoidal function to employ the deep unfolding technique in the computation of the gradient for optimizing the likelihood function based on an estimate of the channel. A similar deep unfolding architecture, labeled as LoRD-Net is proposed in [78] where

the authors formulate a data-driven algorithm for the blind detection of BPSK signals in one-bit mMIMO systems.

2.4 . Summary

In this chapter, we presented the system model of the quantized MIMO channel assumed in this thesis which mainly focuses on a fully digital architecture with a flat-fading channel. The mathematical model is simplified under perfect synchronization and analog processing assumptions, allowing the design and analysis of channel estimation and data detection algorithms. We cover state-of-the-art methods for dealing with the latter problems, which include linear receivers based on linear channel models using the Bussgang decomposition and probabilistic methods that consider the overall statistical model of the communication system.

Chapter 3

One-Bit ADCs Data Detection with Perfect Channel State Information

This chapter addresses the problem of data detection in one-bit MIMO systems under the assumption that perfect CSI is available at the receiver. The main motivation behind this channel is two-fold in terms of simplicity; only a comparator is needed at the receiver which alleviates the need for adaptive-gain-control at the receiver [49, 55] and it facilitates the analysis of the proposed algorithms. The transmitted signal contains symbols usually chosen from quadrature constellations with a finite number of amplitudes such as QPSK, 16 or 64-QAM. The optimal receiver that minimizes the detection error probability is the ML detector which requires an exhaustive search over the entire constellation. In a MIMO system with multiple transmit antennas, the size of the search set grows exponentially with their number, rendering the computational complexity prohibitive. Sphere-decoding is a heuristic method that aims to reduce this complexity by enumerating the possible transmitted vectors that depend on the received signal. The Euclidean distance metric is optimal when the received signal takes values with infinite precision and the noise is i.i.d. Gaussian. It does not extend directly when they are binary-valued as is the case for the one-bit quantized MIMO channel. In this chapter, we first formulate the ML data detection for the one-bit MIMO channel and show how the SD algorithm can be extended to our system model.

3.1 . Maximum Likelihood Data Detection Problem

Consider the negative log-likelihood function of the real-equivalent channel in Eq. (2.27) with $b = 1$. Given the channel $\boldsymbol{H}$ and a received vector $\boldsymbol{y}$, the ML detector seeks to reduce the detection error probability by finding the transmitted signal $\boldsymbol{x}$ in the set $\mathcal{X}^{2M}$ that minimizes this function

$$\boldsymbol{x}_{\mathrm{ML}} = \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad -\sum_{n=1}^{2N} \ln\left[\Phi\left(\frac{y_n \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\right]. \tag{3.1}$$

As stated earlier, this is a difficult combinatorial problem with a search space of size $|\mathcal{X}|^{2M}$. One useful observation is that $\ell(\boldsymbol{x})$ is log-convex

over its argument which we can check by looking at the Hessian and verifying that the eigenvalues are non-negative. This allows us to relax the discrete constraint and solve the problem using gradient-descent techniques to obtain an initial estimate, then round each coordinate to the nearest integer in $\mathcal{X}$. Note that this is only a heuristic as it ignores the combinatorial nature of the problem. We can try to improve the performance through list detection, which consists of enumerating a set of possible transmitted signals based on specified criteria, e.g. the second stage nML step discussed in Chapter 2 [71, 79] or similar variants [40], and then maximizing the log-likelihood over this list. The choice of the criteria can be crucial as it impacts both the performance and how we can control the size of the list. This second-stage approach aims to reduce the computational complexity while remaining as faithful as possible to the original optimization problem. Under infinite precision assumptions and i.i.d. Gaussian noise, the ML metric reduces to a quadratic loss which can be computed using SD algorithms that reduce the input search space [106]. The nonlinearity presented by the quantizer does not allow for the straightforward application of these algorithms. However, the geometric interpretation of SD gives an intuition as to how we can extend it to our channel model. For this purpose, we review in the upcoming section the basic idea of conventional SD in MIMO channels.

3.2 . Sphere-Decoding with One-Bit ADCs

3.2.1 . Review of Sphere-Decoding

In this section, we will review the basic concepts of SD for the classical channel where the receiver has access to the signal with infinite precision

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}. \tag{3.2}$$

The detection problem for this channel has been extensively studied where different linear receivers can also be employed to retrieve $\boldsymbol{x}$ including ZF and MMSE detectors. It's convenient to translate the constellation to a subset $\mathcal{D}$ of the $2M$-dimensional lattice $\mathbb{Z}^{2M}$, by shifting and scaling the coordinates in the constellation assuming minimum symbol distance equal to 2. This set can be written as

$$\mathcal{D} = \left\{ \boldsymbol{s} = \frac{1}{2}(\boldsymbol{x} + \mathbf{1}_{2M}) : \boldsymbol{x} \in \mathcal{X}^{2M} \right\}, \tag{3.3}$$

where $\mathbf{1}_{2M}$ is the all-one vector of size $2M \times 1$. Including this re-scaling, the ML detector, under i.i.d. Gaussian noise, is the optimal criterion and is given by

$$\boldsymbol{s}_{\mathrm{ML}} = \arg\min_{\boldsymbol{s}\in\mathcal{D}} \quad ||\boldsymbol{t} - \boldsymbol{B}\boldsymbol{s}||^2, \tag{3.4}$$

where $\boldsymbol{B} = 2\boldsymbol{H}$ and $\boldsymbol{t} = \boldsymbol{y} + \boldsymbol{H}\mathbf{1}_{2M}$ which takes values in $\mathbb{R}^{2N}$. This is closely related to the integer least-squares problem where the matrix $\boldsymbol{B}$ is usually referred to as the *lattice-generating matrix* [107], i.e., we have the truncated lattice $\Gamma = \{\boldsymbol{B}\boldsymbol{s} : \boldsymbol{s} \in \mathcal{D}\}$ and we would like to find the closest point in Γ to $\boldsymbol{t}$. Due to the discrete nature of the problem, finding an exact solution to this problem is known to be NP-hard for a general lattice-generating matrix $\boldsymbol{B}$, however, there exist heuristics that can retrieve a solution. For example, the Fincke-Pohst algorithm [108] is one of the earliest SD heuristics that was proposed to solve Eq. (3.4) that we will focus on to illustrate the functionality. The basic idea is to first constrain the search space over a sphere of radius d instead of the entire lattice $\mathcal{D}$, then enumerate all possible points within this sphere in a manner analogous to traversing a tree. In other words, we need to find

$$\left\{\boldsymbol{s} \in \mathcal{D} : ||\boldsymbol{t} - \boldsymbol{B}\boldsymbol{s}||^2 \leq d\right\}. \tag{3.5}$$

The procedure begins with a triangularization of $\boldsymbol{B}$ by performing a QR factorization such that

$$\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_1 \\ \mathbf{0}\end{bmatrix} \tag{3.6}$$

where $\boldsymbol{Q}$ is an orthogonal matrix[1] and $\boldsymbol{R}$ is upper-triangular. With this decomposition, we can write the metric as the following [106, 109]

$$\begin{aligned}\boldsymbol{s}_{\text{ML}} &= \arg\min_{\boldsymbol{s}\in\mathcal{D}} \quad ||\boldsymbol{t} - \boldsymbol{B}\boldsymbol{s}||^2 \\ &= \arg\min_{\boldsymbol{s}\in\mathcal{D}} \quad ||\boldsymbol{R}_1\boldsymbol{s}_{\text{ZF}} - \boldsymbol{R}_1\boldsymbol{s}||^2 + c \\ &= \arg\min_{\boldsymbol{s}\in\mathcal{D}} \quad ||\hat{\boldsymbol{t}} - \boldsymbol{R}_1\boldsymbol{s}||^2 \end{aligned} \tag{3.7}$$

where we have the ZF estimate $\boldsymbol{s}_{\text{ZF}} = (\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{t}$ and c is a constant that is independent of $\boldsymbol{s}$. Exploiting the triangular structure of $\boldsymbol{R}_1$, the Fincke-Pohst algorithm searches for the integer-valued coordinates, while moving in descending order from $m = 2M$ to $m = 1$, that verify the radius constraint. Once we find a point in $\mathcal{D}$ that is compatible, we store it and proceed until the algorithm terminates before outputting the point with the smallest Euclidean distance. In a practical communication scenario, the received signal $\boldsymbol{r}$ is not arbitrary and is usually a perturbed version of the lattice point by noise components that follow a known distribution. In addition, the matrix $\boldsymbol{B}$ also exhibits a known structure or statistical behavior. Under these assumptions, the authors in [110] show that SD has an average complexity that is polynomial in M, motivating

[1] An orthogonal matrix $\boldsymbol{A}$ is defined such that $\boldsymbol{A}\boldsymbol{A}^\mathsf{T} = \boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \boldsymbol{I}$.

its implementation in practical systems [109]. One of the key questions with SD is how to choose the radius d in advance. This radius can be set as a scale of the SNR by observing that the noise vector magnitude follows a χ^2 distribution. Therefore, we can choose d such that, with high probability, the magnitude of the noise vector is less than d. Another approach would be to set the radius value to infinity as this can guarantee to find at the first iteration a single point, then update the radius accordingly. The SD algorithm cannot be applied directly to Eq. (3.1) since we do not have a quadratic form, however, its functionality provides us with the intuition on how such an extension is possible. First observe that given a good initial estimate $\hat{\boldsymbol{x}}$, we would like to enumerate a list of candidate points in its vicinity that might contain the ML solution. We can choose to enumerate the points by defining a sphere around $\hat{\boldsymbol{x}}$. It's evident that the parameters of the sphere should also take advantage of the available CSI at the receiver. We will elaborate in detail on this reasoning in the upcoming subsection.

3.2.2 . Likelihood Function and Taylor Series Approximation

Given an initial estimate $\boldsymbol{x}_0$ that we can obtain with gradient-descent iterations, we expand the negative log-likelihood function in Eq. (3.1) in Taylor series form up to second order around this point to obtain

$$\ell(\boldsymbol{x}) = \ell(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{T}}\nabla_{\ell(\boldsymbol{x}_0)} + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{T}}\nabla^2_{\ell(\boldsymbol{x}_0)}(\boldsymbol{x}-\boldsymbol{x}_0) \tag{3.8}$$

where $\nabla_{\ell(\boldsymbol{x}_0)}$ and $\nabla^2_{\ell(\boldsymbol{x}_0)}$ are respectively the gradient and Hessian operators evaluated at $\boldsymbol{x}_0$ and given by the following expressions after a simple application of the chain rule

$$\begin{cases} \nabla_{\ell(\boldsymbol{x})} = \dfrac{\partial \ell(\boldsymbol{x})}{\partial \boldsymbol{x}} = -\dfrac{1}{\sigma}\displaystyle\sum_{n=1}^{2N} \kappa\left(\frac{y_n \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right) y_n \boldsymbol{h}_n, & (3.9) \\ \nabla^2_{\ell(\boldsymbol{x})} = \dfrac{\partial^2 \ell(\boldsymbol{x})}{\partial \boldsymbol{x}^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{n=1}^{2N} \eta\left(\frac{y_n \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right) \boldsymbol{h}_n \boldsymbol{h}_n^{\mathsf{T}}, & (3.10) \end{cases}$$

where we define the following functions for convenience

$$\begin{cases} \kappa(u) = \dfrac{\phi(u)}{\Phi(u)}, & (3.11) \\ \eta(u) = \kappa(u)\big[u + \kappa(u)\big]. & (3.12) \end{cases}$$

This second order approximation allows us to describe a quadratic surface around this initial estimate $\boldsymbol{x}_0$. If the optimal solution $\boldsymbol{x}^*$ is given, then the Taylor expansion will describe the best quadratic approximation of the likelihood function at that point. Although this statement

is obvious, it provides the motivation behind the algorithm. Mainly, if we can find a quadratic surface described by its data-dependent parameter pair $(\nabla_{\ell(\boldsymbol{x})}, \nabla^2_{\ell(\boldsymbol{x})})$ that is close to the optimal and search the points in its vicinity, then we might be able to retrieve the ML solution that maximizes (3.1). The advantage of this formulation is that we can leverage the geometry of the problem, at least up to second-order, to remain as faithful as possible to the original objective function. Moreover, the approximation takes into account the effect of the initial estimate and CSI on which it depends nonlinearly through the functions in Eq. (3.11) and (3.12). A summary of the proposed algorithm is given in Algorithm 1 and outlined in details as the following:

Step 1: We can obtain an initial estimate of the solution $\hat{\boldsymbol{x}}$, by relaxing the discrete condition and iterating using gradient descent with a step size set to ζ,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \zeta \nabla_{\ell(\boldsymbol{x}_t)}. \tag{3.13}$$

Note that since the objective function is convex, the step size ζ can be set optimally according to its smoothness factor (Lipschitz constant) in order to ensure convergence. We elaborate more on this in the discussion on the computational complexity in the following section. This gives us the initial estimate $\hat{\boldsymbol{x}} \in \mathbb{R}^{2M}$. While this results in a solution under the relaxed constraint, we still need to retrieve a discrete solution in $\mathcal{X}^{2M}$. A naive approach would be to round each real coordinate to the closest integer, i.e., projecting to the closest neighbor with respect to the Euclidean distance. We refer to the overall result (iterative gradient-descent and then rounding) as the nonlinear zero-forcing (NLZF) solution

$$\boldsymbol{x}_{\text{NLZF}} = \text{Proj}_{\mathcal{X}^{2M}}(\hat{\boldsymbol{x}}), \tag{3.14}$$

where the projection operation $\text{Proj}_{\mathcal{X}^{2M}}(\cdot)$ simply means in this context, rounding the coordinates of $\hat{\boldsymbol{x}}_i$ for $i = 1, \ldots, 2M$ to the closest points in the constellation $\mathcal{X}^{2M}$.

We expand the log-likelihood function around $\hat{\boldsymbol{x}}$ to obtain a quadratic function. For any fixed level set of $\ell(\boldsymbol{x})$ and assuming the gradient term is negligible, this defines the boundary of an ellipsoid

$$(\boldsymbol{x} - \hat{\boldsymbol{x}})^{\mathsf{T}} \nabla^2_{\ell(\hat{\boldsymbol{x}})} (\boldsymbol{x} - \hat{\boldsymbol{x}}) = d. \tag{3.15}$$

Fig. 3.1 gives illustrates closely the geometric intuition of the algorithm for two dimensions. The eigenvalues of the Hessian represent the variation of our quadratic approximation along each principle axis. For example, a large eigenvalue indicates a sharp decrease of the approximate likelihood, therefore it is reasonable to exclude points along that direction

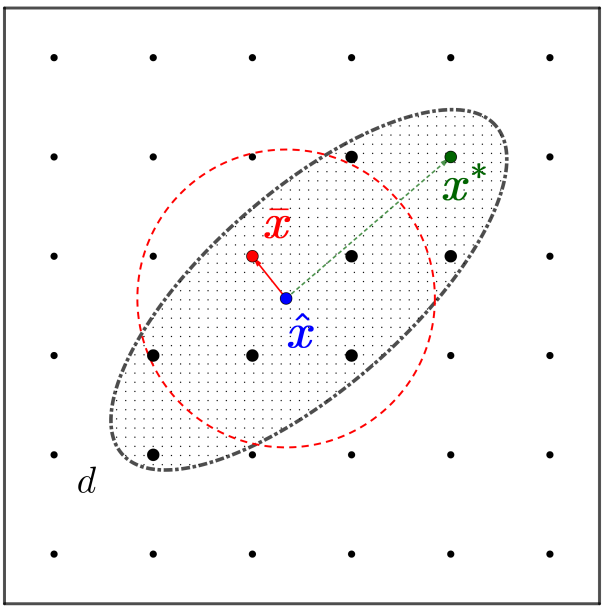

Figure 3.1: Illustration of second stage SD in two dimensions. The initial estimate from stage 1, the nearest neighbor based on the Euclidean distance and the ML estimate are $\hat{\boldsymbol{x}}$, $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}^*$, respectively. The shaded ellipse boundary is defined by d, the dashed circle represents the search region under the Euclidean distance and the enumerated points are represented in a larger scale.

when enumerating our list. This property is lost if we were to consider points based only on a fixed Euclidean distance from $\hat{\boldsymbol{x}}$ as shown by the dashed circle. Rounding the coordinates directly to find the closest point to $\hat{\boldsymbol{x}}$, we would get an incorrect solution $\bar{\boldsymbol{x}}$. We can add more candidate points by taking a larger ellipsoid, which is well-motivated since we expect these new points to remain within this region described by the objective function's second-order dynamics. Obviously, and as noted in the previous remark, this depends on whether the quadratic approximation in (3.15) is sufficiently accurate in describing these variations, which we show through numerical experiments is indeed the case for regimes of interest.

Step 2: We first perform a Cholesky decomposition on the Hessian, $\nabla^2_{\ell(\hat{\boldsymbol{x}})} = \boldsymbol{R}^\mathsf{T}\boldsymbol{R}$, which we can do since the log-likelihood is convex. Then from (3.15) we obtain the condition

$$\begin{aligned} ||\mathbf{R}(\hat{\boldsymbol{x}} - \boldsymbol{x})||^2 &\leq d \\ ||\boldsymbol{t} - \boldsymbol{R}\boldsymbol{x}||^2 &\leq d \\ ||\hat{\boldsymbol{t}} - \tilde{\boldsymbol{R}}\boldsymbol{s}||^2 &\leq d. \end{aligned} \tag{3.16}$$

Translating $\boldsymbol{x}$ to the integer truncated lattice $\mathcal{D}$, we now see that this

coincides exactly with the objective function presented in Eq. (3.7) of SD for unquantized channels with the vector $\hat{\boldsymbol{t}} = \boldsymbol{R}(\hat{\boldsymbol{x}} - \mathbf{1}_{2M})$ and the upper-triangular matrix $\tilde{\boldsymbol{R}} = 2\boldsymbol{R}$ acting as the observations and effective channel, respectively. The final steps consist of populating the list where the original ML metric will be evaluated.

Step 3: We can now use any efficient SD algorithm, for example Fincke-Pohst [108], to construct a list $\mathcal{S}$ of the closest candidates. The only challenge remaining consists of how we choose the initial radius d which we elaborate upon in this step. The algorithm procedure is analogous to constructing a tree with $2M$ levels whereby a path at depth j corresponds to a partial vector $\boldsymbol{s}_j^{2M} = [s_j, s_{j+1}, \ldots, s_{2M}]$ which is assigned the weight for each given j [106]

$$\varphi\big(\boldsymbol{s}_j^{2M}\big) = \sum_{i=j}^{2M} \left| \hat{t}_i - \sum_{m=i}^{2M} \tilde{r}_{i,m} s_m \right|^2 . \tag{3.17}$$

Fixing d, the SD algorithm enumerates all vectors such that the following set of constraints is satisfied

$$\varphi\big(\boldsymbol{s}_j^{2M}\big) \leq d, \qquad \text{for } j = [1, \ldots, 2M] . \tag{3.18}$$

A key difference in our approach is that we need to fix $|\mathcal{S}|$ to a certain number τ during the enumeration. Therefore, our algorithm needs to output the τ closest vectors or leaf nodes

$$[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k, \ldots, \boldsymbol{s}_\tau] \tag{3.19}$$

that are sorted according to their weights

$$\boldsymbol{d} = [d_1, \ldots, d_k, \ldots, d_\tau], \tag{3.20}$$

where $d_k = \varphi(\boldsymbol{s}_k)$ and $d_1 \leq ... \leq d_k \leq ... \leq d_\tau$. This list can then be defined, after translating back to the constellation $\mathcal{X}$ as

$$\mathcal{S} = \big\{ [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_\tau] \in \mathcal{X}^{2M} \;\big|\; ||\boldsymbol{t} - \boldsymbol{R}\boldsymbol{x}_k||^2 \leq d_\tau \big\} . \tag{3.21}$$

In order to guarantee that the minimum number of points satisfies the list size constraint and to alleviate the problem of initializing the radius, we first define the following event

$$\mathcal{E} = \{\textit{The search sphere is not empty}\}, \tag{3.22}$$

and initialize the elements of the radius vector $\boldsymbol{d}$ to infinity, which also corresponds to the weights of the candidate points sorted in increasing

order. While the event $\mathcal{E}$ remains true, the SD algorithm will then retrieve a point or a leaf node according to the largest weight d_τ in $\boldsymbol{d}$ and then appends it to the list before sorting $\boldsymbol{d}$ and $\mathcal{S}$ in ascending order of weights. Until the list is fully populated, the event $\mathcal{E}$ remains true since the largest radius in $\boldsymbol{d}$ is always set to infinity, therefore the algorithm is guaranteed to enumerate τ points. Once the maximum radius d_τ becomes finite, the algorithm continues to enumerate the remaining points inside the sphere with that given radius, which can only decrease as the algorithm continues, otherwise, this will contradict the constraint. At each remaining iteration, the number of leaf nodes is only finite, and therefore the algorithm is guaranteed to converge and terminate once there are no more points that satisfy the constraint.

Step 4: Finally, we evaluate the likelihood function over the set $\mathcal{S}$ instead of $\mathcal{X}^{2M}$

$$\boldsymbol{x}_{\text{SD}} = \arg\min_{\boldsymbol{x}\in\mathcal{S}} \ \ell(\boldsymbol{x}). \tag{3.23}$$

Algorithm 1 Sphere-decoding with one-bit ADCs

Input: Channel matrix $\boldsymbol{H}$, binary observations vector $\boldsymbol{y}$, noise variance σ, candidates list size τ.

Initialization: Fix the step size ζ, maximum number of iterations is $t_{\max}$, tolerance threshold ε, and the elements of the weight vector $\boldsymbol{d}$ in (3.20) are set to ∞, $|\mathcal{S}| = \varnothing$.

Step 1: Obtain an initial estimate $\hat{\boldsymbol{x}}$ using gradient descent:

while $t \leq t_{\max}$ **and** $|\ell(\boldsymbol{x}_t) - \ell(\boldsymbol{x}_{t-1})| > \varepsilon|\ell(\boldsymbol{x}_{t-1})|$

  $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \zeta\nabla_{\ell(\boldsymbol{x}_t)}$

end

Store $\hat{\boldsymbol{x}} = \boldsymbol{x}_{t_{\text{final}}}$.

Step 2: Perform the Cholesky factorization $\nabla^2_{\ell(\hat{\boldsymbol{x}})} = \boldsymbol{R}^{\mathsf{T}}\boldsymbol{R}$.

Step 3: Begin populating the candidates list $\mathcal{S}$ with the SD algorithm and set $d = d_\tau$:

while $\mathcal{E}$ is true

  Find leaf node $\breve{\boldsymbol{x}}$ such that $\varphi(\breve{\boldsymbol{x}}) \leq d_\tau$

  Append $\mathcal{S} \leftarrow \breve{\boldsymbol{x}}$

  Update $\boldsymbol{d}(\tau) = \varphi(\breve{\boldsymbol{x}})$

  Sort $\boldsymbol{d}$ and $\mathcal{S}$ in ascending order of weights

  Update the sphere radius in (3.16) $d = d_\tau$

end

Step 4: Find $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{S}} \ell(\boldsymbol{x})$.

Output: $\boldsymbol{x}_{\text{SD}} = \boldsymbol{x}^*$

3.2.3 . Computational Complexity

In order to assess the computational complexity of the proposed algorithm, we can analyze separately both stages of finding the initial estimate $\hat{\boldsymbol{x}}$ (**Step 1**) and evaluating the likelihood after constructing the list $\mathcal{S}$ of points inside the sphere (**Step 2**–**4**). The convergence and number of iterations required highly depend on the choice of the step size ζ. Given that the objective function is convex (the Hessian is semi-definite as shown in the previous part), and that the gradient is Lipschitz continuous with constant L, setting the step size $\zeta \leq \frac{1}{L}$ will guarantee convergence with a rate $\mathcal{O}(1/K)$, where K is the number of iterations. And by letting

$$\zeta \leq \frac{\sigma^2}{\lambda^2_{\max}(\boldsymbol{H})}, \tag{3.24}$$

with $\lambda_{\max}(\boldsymbol{H})$ denoting the largest singular value of the matrix $\boldsymbol{H}$, this should guarantee the convergence of the gradient descent algorithm (please see Appendix B for details). Note that this choice of the step size should be sufficient for any quantizer resolution as the upper bound in (3.24) applies for any δ and b. We note that the number of elementary operations required to obtain the largest eigenvalue in (6.11) is in the order of $O(MN^2)$ to compute $\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}$ and $O(M^3)$ to find the eigenvalues. The gradient descent operations scale as $O(KMN)$, where K depends on the step size ζ and the acceptable tolerance threshold ε. We show in our numerical results the average number of iterations required under several assumptions on the quantizer's resolution, SNR, and number of transmit and receive antennas. For the second stage of the algorithm, the number of operations for computing the Cholesky decomposition in **Step 2** scales as $O(M^3)$. Within our framework, assessing the average complexity of constructing the candidates list $\mathcal{S}$ in a light similar to that of [110] would be an interesting direction of future work but out of the scope of this paper. For this purpose, we resort to evaluating the average number of floating points operations (n_{flops}) numerically as will be detailed in the section on numerical experiments. Finally, in **Step 4**, we evaluate the likelihood which requires $O(NM)$ operations and scales with $|\mathcal{S}|$ as $O(|\mathcal{S}|MN)$, which in our case can be fixed as desired.

3.3 . Simulation Results

3.3.1 . ML Lowerbound Criterion

In order to assess the gap between the proposed schemes and optimality in terms of error probability, we adopt the oracle lower bound method proposed in [86]. Assume that $q(\boldsymbol{y}|\boldsymbol{x})$ is the likelihood function

that one would like to maximize. Then, it can be readily shown that the probability of ML detection error is lower bounded as [86]

$$\boldsymbol{P}_{\mathrm{ML}}(\text{error}) \geq \boldsymbol{P}\left\{q(\boldsymbol{y}|\boldsymbol{x}) < q(\boldsymbol{y}|\boldsymbol{x}^{\dagger})\right\}, \tag{3.25}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the actual input and output of the channel, respectively; $\boldsymbol{x}^{\dagger}$ is any point inside the input set (which can depend on the output $\boldsymbol{y}$ or even on the input $\boldsymbol{x}$). In particular, when $\boldsymbol{x}^{\dagger}$ is the ML solution, equality in (3.25) is achieved. When $\boldsymbol{x}^{\dagger}$ is the actual input, the right-hand side of (3.25) is 0 and the inequality is trivial. Intuitively, the inequality says that whenever there exists a point with a strictly higher likelihood than the true input, ML detection cannot return the correct decision.

3.3.2 . Vector Error Rates Performance

We now assess the performance of the proposed one-bit quantized SD algorithm. We consider a MIMO system with several assumptions on the number of transmit antennas, receive antennas, the SNR and the maximum number of candidate points in the set $\mathcal{S}$. We begin by investigating the performance with i.i.d. Rayleigh fading and QPSK signaling (i.e., $\tilde{\mathcal{X}} = \{\pm 1 \pm 1j\}$ or $\mathcal{X} = \{-1, +1\}$ for the real equivalent channel). The results are shown in Fig. 3.2-3.4 where the cardinality of the enumerated set $\mathcal{S}$ in Eq. (3.21) is fixed to 3 points for several MIMO antenna setups. For the case of a 2×16 and 4×32-MIMO system in Fig. 3.2 and Fig. 3.3, the exact ML with exhaustive search, Eq. (3.1), can be computed directly. It can be readily seen that the SD approach achieves the ML performance in terms of VER compared to naive NLZF (3.14). For reference, we also present the lower bound (3.25) for the SD metric which is shown to be tight as expected. As we double the number of transmit and receive antennas, we observe that a candidate set size of 3 points remains near-optimal for a 8×64-MIMO setup in Fig. 3.4c. Given the increased ML complexity, we only show the lower bound which is tight with the proposed SD algorithm solution. Compared to the Euclidean distance rounding, we can attribute this performance improvement to the reasoning provided earlir. In other words, by enumerating the points with respect to the basis described by the original objective function second-order dynamics, $\nabla^2_{\ell(\boldsymbol{x})}$, we are more likely to capture the optimal solution $\boldsymbol{x}^*$ since we do not ignore the available CSI while doing this enumeration, in contrast to the approach proposed by the second stage nML method. Moreover, in contrast to the heuristic proposed in [81], our extension is more aligned with the conventional SD since it does not involve pre-processing stage that depends on listing all possible transmit vectors in $\mathcal{X}^M$.

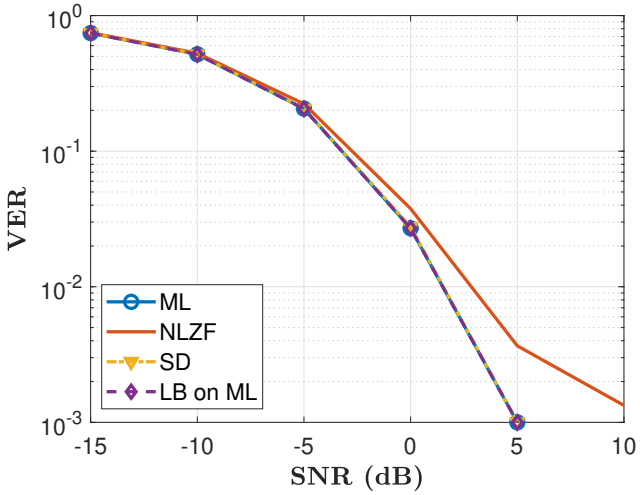

Figure 3.2: VER for 2×16-MIMO with QPSK, perfect CSI, and fixed candidates list cardinality $|\mathcal{S}| = 3$ for varying SNR.

We are also interested in evaluating the performance as we increase the constellation size to 16-QAM, which entails an increase in the search space when computing the ML metric. As a benchmark, we compare our results with the two-stage near ML (nML) method proposed in [71]. Fig. 3.5 shows the achievable VER for a 8×128-MIMO system with Rayleigh fading and 16-QAM signaling with varying sizes of the candidate set $\mathcal{S}$. The radius χ has been set equal to 1.3 and the average size of $\mathcal{C}$ over the SNR range is 4.25 points. We evaluate the performance of the SD algorithm with a starting size of $|\mathcal{S}| = 2$. For a similar size of candidate sets, the SD solution has an improved performance in terms of VER. The performance gradually improves and the lower bound on the ML becomes tighter as we increase the list size to 5. Increasing both the number of transmit antennas and the constellation size, these numerical results indicate how the SD algorithm remains robust to the added search complexity. The performance with NLZF is the weakest since rounding the coordinates based on the Euclidean distance is sub-optimal.

3.3.3 . Performance Assessment in a Spatially Correlated Channel

It is of interest to investigate the performance of our proposed algorithm in a typical single-cell mMIMO environment. We now consider a more realistic setting, where the antennas at the base station exhibit spatial correlation between them. In other words, we assume that the

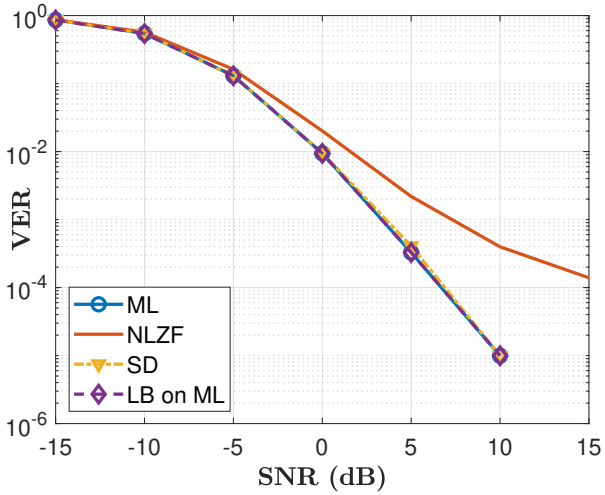

Figure 3.3: VER for 4×32-MIMO with QPSK, perfect CSI, and fixed candidates list cardinality $|\mathcal{S}| = 3$ for varying SNR.

columns in the channel matrix

$$\tilde{\boldsymbol{H}} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m, \ldots, \boldsymbol{c}_M] \tag{3.26}$$

are each drawn according to $\mathcal{CN}(\boldsymbol{0}_N, \boldsymbol{C}_m)$ and are independent amongst each other (i.e., we do not assume that the transmitters are co-located). The channel covariance matrices, $\boldsymbol{C}_m$ are generated according to the model specified in [111] and as detailed in Appendix C. We assume a 10×72-MIMO system with QPSK signaling and that the users average angle of arrivals (AoA) are drawn uniformly from $[-\frac{\pi}{3}, \frac{\pi}{3}]$ where they share the same value of the angular spread AS $= 5°$. The results are shown in Fig. 3.6 where we also compare our proposed approach with the linear Bussgang receivers: the BMMSE Eq. (2.22) and BZF Eq. (2.20). We observe that the effect of receiver spatial correlation has a detrimental effect on the performance. Nevertheless, the proposed approach achieves a performance that is tight with the ML lower bound for that particular channel.

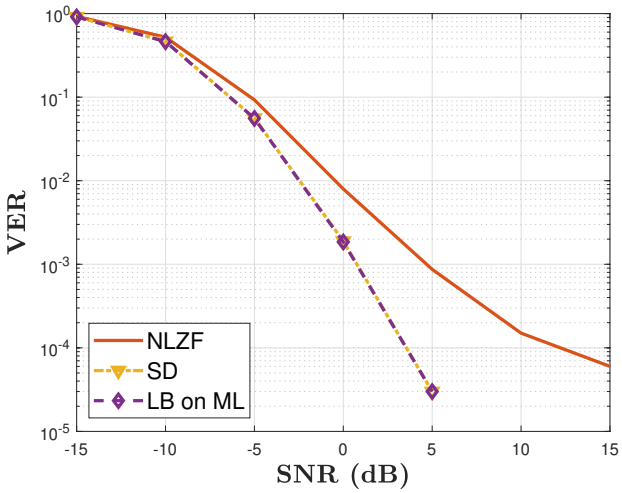

Figure 3.4: VER for 8×64-MIMO with QPSK, perfect CSI, and fixed candidates list cardinality $|\mathcal{S}| = 3$ for varying SNR.

3.3.4 . Numerical Assessment of SD Computational Complexity

We first focus on the first stage of obtaining the initial estimate $\hat{\boldsymbol{x}}$ (**Step 1**). The average number of iterations, K_{SD}, conducted by gradient descent and the first stage of the nML solution [71] are shown in Table 3.1. We note that the number of iterations increases with the SNR. We can justify this behavior by examining the eigenvalues of the Hessian in (3.10) which are governed by the values of the function $\eta(\cdot)$. As the SNR increases, the eigenvalues become very small, indicating that the curvature of the likelihood function is becoming less steep and therefore requiring more iterations in order to reach convergence. Moreover, our approach requires a fewer number of iterations compared to nML. This can be associated with the fact that the chosen step size in our setup is set according to the Lipschitz constant of the gradient, whereas in [71] it is fixed for all SNR values.

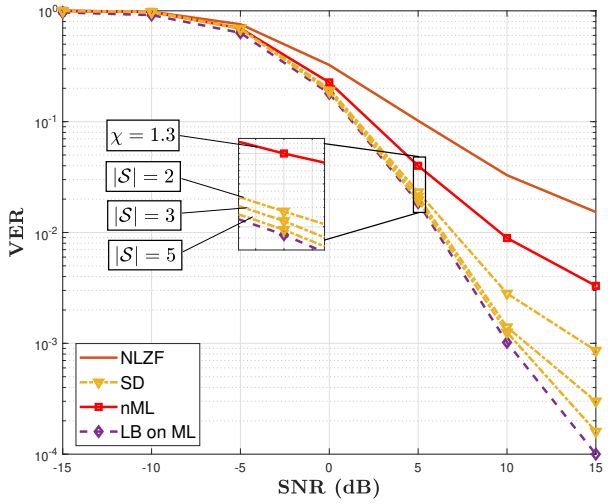

Figure 3.5: VER for 8×128-MIMO with 16-QAM, perfect CSI and varying SNR for several data detection metrics, and different sizes of $\mathcal{S}$.

Table 3.1
Average number of iterations K_{SD} and K_{nML} with $b = 1$, different assumptions on the SNR, constellation size, and antenna configurations.

	$\rho = -10$ dB		$\rho = 10$ dB	
	K_{SD}	K_{nML}	K_{SD}	K_{nML}
QPSK, $M = 8, N = 64$	35.7	44.5	198.3	313.6
16-QAM, $M = 4, N = 64$	32.9	64.1	153.9	267.6
16-QAM, $M = 8, N = 128$	44.6	49.7	193.3	307.6

3.4 . Summary

In this chapter, we have formulated the data detection problem in the one-bit quantized MIMO channel. Based on a geometric interpretation of the likelihood function, it was shown that the conventional SD algorithm can be extended to this channel model through a second-order approximation. In contrast to the proposed heuristics in the literature, the approach exploits the available CSI in constructing the list of candidate points through the Hessian function. In our numerical experiments, we assess the VER performance and show that the proposed approach is near-optimal with respect to an oracle-based lower bound on the ML error probability, under different assumptions on MIMO systems and constellation sizes. Moreover, we look at the effects of spatial correlation between the receive antennas for the given channel of each transmitter, it was

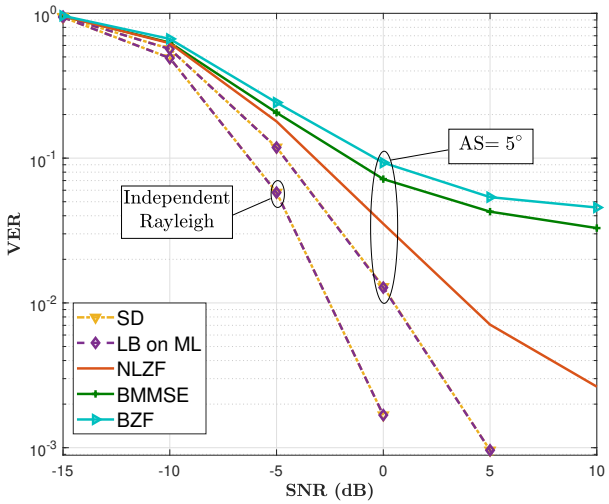

Figure 3.6: VER with QPSK, perfect CSI for a 10×72-MIMO system with one-bit quantization and a list size $|\mathcal{S}| = 5$. The angular spread for the correlated channels case is set to $5°$.

shown that, in general, the correlation degrades the overall performance in terms of VER compared to the i.i.d. case, however, the algorithm remains near-optimal compared to the ML performance. Finally, we study numerically the computational complexity of the first stage in the SD algorithm where it is demonstrated that a judicious choice of the step size reduces the number of iterations required during gradient descent.

Chapter 4

One-Bit ADC Data Detection without Prior Channel State Information

In practice, the realization of the channel might not be readily available at the receiver for the data detection process. Different transmission strategies that address this problem can be used, for example, differential-phase-shift-keying where information is carried in the phase between consecutive data symbols. For the MIMO Rayleigh flat-fading channel, the authors in [112] propose unitary space-time modulation schemes that do not require knowing the channel gain. Another standard approach is to estimate the fading coefficients and conduct the detection process accordingly [93]. The performance can also be improved when the detected data symbols are re-used to refine the estimate of the channel coefficients. This procedure is known as joint channel and data estimation. This is feasible when the channel is assumed to be constant over a sufficiently long duration and the effective data rate is not heavily penalized by the training overhead cost. These strategies can be seen as special cases of addressing the general problem of noncoherent communication [113], where it's of interest to design transmission schemes that can achieve high transmission rates in fading channels without any a priori CSI. We note that the noncoherent capacity for the block-fading MIMO quantized channel is not known and only approximations are available in limiting regimes such as very low SNR [67] or large system limits where both the number of transmitters and receivers grow to infinity [69]. Nevertheless, the same training strategies can be applied, however, with incurred difficulties due to the nonlinearity induced by the quantizer. As discussed in Chapter 2, linear channel estimators based on the Bussgang decomposition have been proposed and were also used to compute lower bounds on the capacity under the assumptions of channel hardening and favorable propagation [46, 68]. These estimators are sub-optimal as they ignore the data dependence of the noise in the residual term due to the linearization. In addition, machine learning techniques were also proposed such as the SVM and deep learning approaches that attempt to learn the channel [40, 78, 79].

In this chapter, we tackle the problem of data detection in the absence of CSI in one-bit quantized MIMO block-fading channels while focusing on two approaches. The first one considers a two-stage channel estima-

tion and data detection procedure that formulates the problem as binary classification based on the statistical "probit model" of the channel. This formulation offers more flexibility in studying achievable rates in the context of mismatched decoding, and in particular, the GMI. We use this approach to investigate the instability of the Gaussian CDF under imperfect CSI as reported in [79, 87]. In the second part, we focus on a real channel model and formulate the ML metric that processes the transmitted input and the pilot data jointly. Evaluating this metric exhibits many challenges that stem from two main difficulties. Mainly, multivariate Gaussian orthant probabilities do not have exact forms in general, therefore closed-form expressions of the ML cannot be obtained. For this purpose, we propose an approximation based on the Laplace method. Secondly, the discrete nature of the optimization is another challenge that we address by adapting the SD problem proposed in Chapter 3.

4.1 . Channel Estimation and Data Detection with Probit Regression

We can formulate the channel estimation and data detection stages problems as binary classification procedures under the framework of *probit regression*, essentially the analog of *logistic regression* with the only difference being the assumption on the latent variable distribution, which is in our case Gaussian. We convert the binary outcomes of the channel output from $\{-1, 1\}$ to $\{0, 1\}$ only to emphasize the similarity with the standard probit model form. Under perfect CSI, for the coherent channel, the ML metric can then be re-written as

$$\begin{aligned} \boldsymbol{x}_{\mathrm{ML}} &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad \boldsymbol{W}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{H}) \\ &= \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad -\sum_{n=1}^{2N} \ln\left[\Phi\left(\frac{\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)^{\frac{y_n+1}{2}} \cdot \left(1-\Phi\left(\frac{\boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\right)^{\frac{1-y_n}{2}}\right]. \end{aligned} \tag{4.1}$$

We refer to this metric as the *matched* ML. We assume a pilot transmission strategy as in Eq. (2.4) which we convert to real-equivalent form similarly to Eq. (2.2).

4.1.1 . Channel Estimation Stage

For the channel estimation of the procedure, we are interested in the pilot transmission stage as per Eq. (2.36)

$$\boldsymbol{Y}_p = \mathsf{Q}_1(\boldsymbol{X}_p\bar{\boldsymbol{H}} + \boldsymbol{Z}_p). \tag{4.2}$$

We note that the pilot matrix $\tilde{\boldsymbol{X}}_p$ is unitary and defined from a truncated DFT matrix of size $T_p \times T_p$ such that $\tilde{\boldsymbol{X}}_p \tilde{\boldsymbol{X}}_p^{\mathsf{H}} = T_p \boldsymbol{I}_M$. The matrices take the following forms

$$\begin{cases} \boldsymbol{Y}_p &= [\boldsymbol{y}_{p,1}, \ldots, \boldsymbol{y}_{p,n}, \ldots \boldsymbol{y}_{p,N}] & \boldsymbol{y}_{p,n} \in \{0,1\}^{2T_p} \\ \boldsymbol{X}_p &= \left[\boldsymbol{x}_{p,1}, \ldots, \boldsymbol{x}_{p,t}, \ldots \boldsymbol{x}_{p,2T_p}\right]^{\mathsf{T}} & \boldsymbol{x}_{p,t} \in \mathbb{R}^{2M} \\ \bar{\boldsymbol{H}} &= \left[\bar{\boldsymbol{h}}_1, \ldots, \bar{\boldsymbol{h}}_n, \ldots, \bar{\boldsymbol{h}}_N\right] & \bar{\boldsymbol{h}}_n \in \mathbb{R}^{2M} \\ \boldsymbol{Z}_p &= [\boldsymbol{z}_{p,1}, \ldots, \boldsymbol{z}_{p,n}, \ldots \boldsymbol{z}_{p,N}], & \boldsymbol{z}_{p,n} \in \mathbb{R}^{2T_p} \end{cases}$$

Taking the negative logarithm of the posterior distribution in Eq. (2.37) based on Bayes' theorem, we can re-write the MAP estimator as N parallel optimization problems for each channel vector

$$\begin{aligned} \hat{\boldsymbol{h}}_n &= \underset{\boldsymbol{h}_n \in \mathbb{R}^{2M}}{\arg\min} \quad \mathcal{L}(\boldsymbol{h}_n) \\ &= \underset{\boldsymbol{h}_n \in \mathbb{R}^{2M}}{\arg\min} \quad -\frac{1}{2T_p} \sum_{t=1}^{2T_p} \left(\tfrac{y_{p,n}^t + 1}{2}\right) \ln\left[\Phi\left(\frac{\boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma}\right)\right] \\ &\qquad + \left(\tfrac{1 - y_{p,n}^t}{2}\right) \ln\left[1 - \Phi\left(\frac{\boldsymbol{h}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma}\right)\right] - \sum_{m=1}^{2M} \ln\left[\boldsymbol{P}\left(\bar{h}_{n,m}\right)\right]. \end{aligned} \tag{4.3}$$

We recognize the first summation on the right-hand side as a "cross-entropy" term $L_{\mathrm{CE}}(\bar{\boldsymbol{h}}_n)$. In fact, we can retrieve almost the same expression by minimizing the Kullback-Leibler divergence between the true distribution of the labels $\boldsymbol{y}_{p,n}$ and the estimates obtained from the probit model

$$\hat{\boldsymbol{y}}_{p,n} = \left[\Phi\left(\frac{\bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,1}}{\sigma}\right), \ldots, \Phi\left(\frac{\bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma}\right), \ldots, \Phi\left(\frac{\bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,2T_p}}{\sigma}\right)\right]^T. \tag{4.4}$$

Assuming a Gaussian prior on the channel vectors with i.i.d. elements, the MAP channel optimization problem can then be summarized as

$$\hat{\boldsymbol{h}}_n = \underset{\boldsymbol{h}_n \in \mathbb{R}^{2M}}{\arg\min} \; L_{\mathrm{CE}}(\bar{\boldsymbol{h}}_n) + \eta_p ||\bar{\boldsymbol{h}}_n||^2. \tag{4.5}$$

We see that the prior plays the role of an L_2 regularization term, where we added a tunable hyperparameter η_p. To this end, it can be seen that the channel estimation procedure is analogous to a binary classification problem where we need to find the hyperplane parameters $\bar{\boldsymbol{h}}_n$ that best minimize the cross-entropy loss function. Furthermore, we re-parameterize the objective function by allowing the SNR to be absorbed

within the channel weights and define $\bar{\boldsymbol{\theta}}_n = \frac{\bar{\boldsymbol{h}}_n}{\sigma}$. This modification assumes that we do not require knowledge of the SNR during the pilot training. This is a convex optimization problem since the Hessian of the objective function in Eq. (4.5) is positive definite. Therefore classical gradient descent techniques can be employed to find the optimal weights $\bar{\boldsymbol{\theta}}_n$ for $n = 1, 2, \ldots, 2N$. At each new iteration $k+1$ we update the weights as

$$\hat{\boldsymbol{\theta}}_{n,k+1} = \hat{\boldsymbol{\theta}}_{n,k} - \zeta_p \nabla \mathcal{L}\Big(\hat{\boldsymbol{\theta}}_n\Big), \tag{4.6}$$

with ζ_p and $\nabla\mathcal{L}(\hat{\boldsymbol{\theta}})$ denoting the step size and the gradient of the loss function respectively. The gradient with the probit model is found as

$$\begin{aligned}\nabla \mathcal{L}\big(\bar{\boldsymbol{\theta}}_n\big) = -\frac{1}{2T_p}\sum_{t=1}^{2T_p}\left[\frac{\frac{y_{p,n}^t+1}{2} - \Phi\Big(\bar{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_{p,t}\Big)}{\Phi\Big(\bar{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_{p,t}\Big)\Big(1-\Phi\Big(\bar{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_{p,t}\Big)\Big)}\right] \\ \times\left[\phi\Big(\bar{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_{p,t}\Big)\right] + 2\eta_p\bar{\boldsymbol{\theta}}.\end{aligned} \tag{4.7}$$

The regularization term, which is only a consequence of the MAP formulation, can add robustness by promoting smaller weights. From a geometric perspective, the L_2 term transforms the optimization into a strongly convex problem which can also lead to faster convergence with an appropriately chosen step size. The final obtained estimate of the weights matrix is then used during the data detection stage to obtain

$$\hat{\boldsymbol{\Theta}} = \Big[\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_n, \ldots, \hat{\boldsymbol{\theta}}_N\Big] \tag{4.8}$$

4.1.2 . Data Detection Stage

We propose two different schemes for the data detection stage. When the complexity is not prohibitive, we can employ an exhaustive mismatched ML (mML) at the receiver as the first method. In other words, instead of $W(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{H})$, we use the estimated coefficients

$$\hat{\boldsymbol{x}}_{\text{mML}} = \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min}\ Q(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\Theta}}). \tag{4.9}$$

We refer to this as the probit mML under exhaustive search. We can also devise the data detection stage in a fashion similar to that of channel estimation. Since data detection is conducted independently at each time instant, we drop the time subscript for brevity. For this purpose, we relax the discrete constraint by assuming a box boundary similar to the one presented by C. Thrampoulidis et al. [114]. The latter condition allows us to constrain the search space by imposing a maximum amplitude on

the coordinates $x_{d,m}$ in $\boldsymbol{x}_{d,m}$. We present the derivations according to the real-channel model, but re-written in the following form

$$\boldsymbol{Y}_d = \mathsf{Q}_1\Big(\breve{\boldsymbol{H}}\boldsymbol{X}_d + \boldsymbol{Z}_d\Big),$$

$$\begin{bmatrix}\Re\{\tilde{\boldsymbol{Y}}_d\}\\ \Im\{\tilde{\boldsymbol{Y}}_d\}\end{bmatrix} = \mathsf{Q}_1\Bigg(\begin{bmatrix}\Re\{\tilde{\boldsymbol{H}}\} & -\Im\{\tilde{\boldsymbol{H}}\}\\ \Im\{\tilde{\boldsymbol{H}}\} & \Re\{\tilde{\boldsymbol{H}}\}\end{bmatrix}\begin{bmatrix}\Re\{\boldsymbol{X}_d\}\\ \Im\{\boldsymbol{X}_d\}\end{bmatrix} + \begin{bmatrix}\Re\{\boldsymbol{Z}_d\}\\ \Im\{\boldsymbol{Z}_d\}\end{bmatrix}\Bigg). \quad (4.10)$$

Under the hard box constraint, we obtain

$$\hat{\boldsymbol{x}}_{d,t} = \underset{|x_{d,m}|\leq \max(\mathcal{X})\,\forall m}{\arg\min} \quad L_{\mathrm{CE}}(\boldsymbol{x}_{d,t}), \quad (4.11)$$

where we define $\breve{\boldsymbol{\theta}} = \frac{\breve{\boldsymbol{h}}_n}{\sigma}$ and let

$$\begin{aligned} L_{\mathrm{CE}}(\boldsymbol{x}_{d,t}) = \frac{-1}{2N}\sum_{n=1}^{2N}&\Big(\tfrac{y_{d,n}+1}{2}\Big)\ln\Big[\Phi\Big(\breve{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{x}_d\Big)\Big]\\ &+\Big(\tfrac{1-y_{d,n}}{2}\Big)\ln\Big[1-\Phi\Big(\breve{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{x}_d\Big)\Big]. \end{aligned} \quad (4.12)$$

The first stage consists of updating the gradient according to

$$\hat{\boldsymbol{x}}_{d,k+1} = \hat{\boldsymbol{x}}_{d,k} - \zeta_d \nabla L_{\mathrm{CE}}(\hat{\boldsymbol{x}}_{d,k}), \quad (4.13)$$

with the gradient taking the form at each coordinate

$$\begin{aligned} \frac{\partial L_{\mathrm{CE}}(\boldsymbol{x}_d)}{\partial \boldsymbol{x}_d} = -\frac{1}{2N}\sum_{n=1}^{2N}&\left[\frac{\frac{y_{d,n}+1}{2} - \Phi\Big(\breve{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_d\Big)}{\Phi\Big(\breve{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_d\Big)\Big(1-\Phi\Big(\breve{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_d\Big)\Big)}\right]\\ &\times\Big[\phi\Big(\breve{\boldsymbol{\theta}}_n^{\mathsf{T}}\boldsymbol{x}_d\Big)\breve{\boldsymbol{\theta}}_n\Big]. \end{aligned} \quad (4.14)$$

The hard constraint is enforced at each iteration by projecting the updated value on $\max(\mathcal{X}) \times B_\infty$, where B_∞ is the l_∞ unit ball. We study the achievable rates in the context of mismatched decoding. In particular, we look at the GMI in its dual form which is given by the following expression [115]

$$I_{\mathrm{GMI}}(\boldsymbol{x},\boldsymbol{y}) = \sup_{s\geq 0} \quad \mathbb{E}_{P_{\boldsymbol{XY}}}\left[\log\frac{q(\boldsymbol{x},\boldsymbol{y})^s}{\sum_{\bar{\boldsymbol{x}}} P_X(\bar{\boldsymbol{x}})q(\bar{\boldsymbol{x}},\boldsymbol{y})^s}\right]. \quad (4.15)$$

It is an achievable rate obtained under the assumption of a random coding argument with an i.i.d codebook and an arbitrary decoding metric $q(\boldsymbol{x},\boldsymbol{y})$. It constitutes a lower bound on the mismatched and matched capacity and can be further optimized under the input distribution, however, it remains only as a lower bound due to the absence of a converse in general. Note that for $s = 1$ and a metric matched to the channel statistics, i.e. $q(\boldsymbol{x},\boldsymbol{y}) = W(\boldsymbol{y}|\boldsymbol{x})$, we recover the known mutual information (MI).

4.2 . Optimal Data Detection for the Real Channel Model with Statistical CSIR

In what follows, for simplicity and ease of illustration, we assume a real channel model. The channel matrix realization is not given a priori and only its statistical distribution is known. The system model is then given as the following

$$\begin{aligned} \boldsymbol{Y} &= \mathsf{Q}\big(\bar{\boldsymbol{H}}\boldsymbol{X} + \boldsymbol{Z}\big), \\ [\boldsymbol{Y}_p\ \boldsymbol{Y}_d] &= \mathsf{Q}\big(\bar{\boldsymbol{H}}[\boldsymbol{X}_p\ \boldsymbol{X}_d] + [\boldsymbol{Z}_p\ \boldsymbol{Z}_d]\big). \end{aligned} \tag{4.16}$$

The transmitted signal is decomposed into two independent frames, where $\boldsymbol{X}_p$ is a pre-determined pilot matrix of size $M \times T_p$ known at both the transmitter and receiver, and $\boldsymbol{X}_d$ is the unknown data matrix whose columns are in $\mathcal{X}^M$ and with total length T_d. The corresponding pilot and data binary observations are given by $\boldsymbol{Y}_p \in \{\pm 1\}^{N\times T_p}$ and $\boldsymbol{Y}_d \in \{\pm 1\}^{N\times T_d}$, respectively. For each transmission phase $i \in \{p, d\}$, we have

$$\boldsymbol{Y}_i = \mathsf{Q}(\bar{\boldsymbol{H}}\boldsymbol{X}_i + \boldsymbol{Z}_i), \tag{4.17}$$

such that

$$\begin{cases} \bar{\boldsymbol{H}} = [\bar{\boldsymbol{h}}_1, \ldots, \bar{\boldsymbol{h}}_n, \ldots, \bar{\boldsymbol{h}}_N]^\mathsf{T}, & \bar{h}_{n,m} \sim \mathcal{N}(0,1) \\ \boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,t}, \ldots, \boldsymbol{x}_{i,T_i}], & \\ \boldsymbol{Y}_i = [\boldsymbol{y}_{i,1}, \ldots, \boldsymbol{y}_{i,n}, \ldots, \boldsymbol{y}_{i,N}]^\mathsf{T}, & \boldsymbol{y}_{i,n} \in \{\pm 1\}^T \\ \boldsymbol{Z}_i = [\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,n}, \ldots, \boldsymbol{z}_{i,N}]^\mathsf{T}, & \boldsymbol{z}_{i,n} \sim \mathcal{N}(0, \sigma_i^2 \boldsymbol{I}_T) \end{cases}$$

Under this formulation, with uniform signaling and assuming the channel statistics are known, the optimal metric is then

$$\begin{aligned} \boldsymbol{X}_d^* &= \arg\max_{\boldsymbol{X}_d}\ \boldsymbol{P}(\boldsymbol{Y}_d, \boldsymbol{Y}_p | \boldsymbol{X}_d, \boldsymbol{X}_p) \\ &= \arg\max_{\boldsymbol{X}_d}\ \mathbb{E}_{\bar{\boldsymbol{H}}}\left[\boldsymbol{P}\big(\boldsymbol{Y}_d, \boldsymbol{Y}_p | \boldsymbol{X}_d, \boldsymbol{X}_p, \bar{\boldsymbol{H}}\big)\right], \end{aligned} \tag{4.18}$$

noting that $\boldsymbol{X}_d \in \mathcal{X}^{M\times T_d}$. The difficulty in evaluating this metric is twofold: we have to average over the channel distribution and solve a hard combinatorial problem as in the perfect CSI case. We highlight that a similar metric is proposed for the unquantized MIMO channel in [116] and a closed-form expression can be obtained when the channel is Gaussian. Due to the severe quantization in this setting, obtaining a tractable expression in general is not possible, to the best of our knowledge. In fact, when the channels are independent at each receive antenna and follow a multivariate Gaussian distribution, it can be shown that (4.18) corresponds to the evaluation of multivariate Gaussian orthant probabilities

which is, in general, a hard problem especially for large dimensions [117]. To simplify the problem, we first assume that we can detect each vector at each time instant t independently, i.e., optimize over $\boldsymbol{x}_{d,t} \in \mathcal{X}^M$ and we can therefore drop the time subscript for the data detection phase. As a second step, conditional on the channel and transmitted data, the two transmission stages are independent and we can re-write (4.18) as

$$\begin{aligned}
\boldsymbol{x}_d^* &= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\max}\ \mathbb{E}\left[\boldsymbol{P}\big(\boldsymbol{y}_d|\boldsymbol{x}_d, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\boldsymbol{Y}_p|\boldsymbol{X}_p, \bar{\boldsymbol{H}}\big)\right] \\
&= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\max}\ \int \boldsymbol{P}\big(\boldsymbol{y}_d|\boldsymbol{x}_d, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\boldsymbol{Y}_p|\boldsymbol{X}_p, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{H}}\big)\, \mathrm{d}\bar{\boldsymbol{H}} \\
&= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\max}\ \boldsymbol{P}(\boldsymbol{Y}_p|\boldsymbol{X}_p) \int \boldsymbol{P}\big(\boldsymbol{y}_d|\boldsymbol{x}_d, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p, \boldsymbol{X}_p\big)\, \mathrm{d}\bar{\boldsymbol{H}} \\
&= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\max}\ \mathbb{E}_{\bar{\boldsymbol{H}}|\boldsymbol{Y}_p, \boldsymbol{X}_p}\left[\boldsymbol{P}\big(\boldsymbol{y}_d|\boldsymbol{x}_d, \bar{\boldsymbol{H}}\big)\right]. \qquad (4.19)
\end{aligned}$$

From a Bayesian perspective, this corresponds to maximizing the "predictive distribution" after averaging over the posterior of the latent channel variables. In other words, the detection is conducted without explicit estimation of the channel. This, in turn, requires a closed-form derivation of the channel's posterior distribution given the pilot information

$$\boldsymbol{P}\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p, \boldsymbol{X}_p\big) = \frac{1}{\boldsymbol{\Omega}} \boldsymbol{P}\big(\boldsymbol{Y}_p|\boldsymbol{X}_p, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{H}}\big), \qquad (4.20)$$

where we have the normalizing factor or the "evidence"

$$\boldsymbol{\Omega} = \int \boldsymbol{P}\big(\boldsymbol{Y}_p|\boldsymbol{X}_p, \bar{\boldsymbol{H}}\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{H}}\big)\, \mathrm{d}\bar{\boldsymbol{H}}. \qquad (4.21)$$

With the independence across receive antennas, we can further simplify (4.20)

$$\begin{aligned}
\boldsymbol{P}\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p\boldsymbol{X}_p\big) &= \prod_{n=1}^{N} \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n|\boldsymbol{Y}_p, \boldsymbol{X}_p\big) \\
&= \prod_{n=1}^{N} \left[\frac{1}{\boldsymbol{\omega}_n} \boldsymbol{P}\big(\boldsymbol{y}_{p,n}|\boldsymbol{X}_p, \bar{\boldsymbol{h}}_n\big) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big)\right] \\
&= \prod_{n=1}^{N} \left[\frac{1}{\boldsymbol{\omega}_n} \prod_{t=1}^{T_p} \Phi\left(\frac{y_{p,n}^t \bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma_p}\right) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big)\right], \qquad (4.22)
\end{aligned}$$

where

$$\boldsymbol{\Omega} = \prod_{n=1}^{N} \boldsymbol{\omega}_n, \text{ and } \boldsymbol{\omega}_n = \int \prod_{t=1}^{T_p} \Phi\left(\frac{y_{p,n}^t \bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma_p}\right) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big)\, \mathrm{d}\boldsymbol{h}_n. \qquad (4.23)$$

Computing the integral in (4.19) with respect to the distribution in (4.20) still does not give any closed tractable form. One can then resort to numerical or Monte-Carlo (MC) sampling methods, but as we stated at the beginning of this section, this can be difficult in general, especially when T_p is large [117]. For this purpose, we propose to retrieve an approximate solution using the Laplace method [88].

4.2.1 . Laplace Approximation of the Channel Posterior Distribution

The Laplace approximation (LA) is a Bayesian method in the sense that it approximates the posterior distribution in (4.20) with a Gaussian. The first step consists of expanding the logarithm of (4.20) in Taylor series form up to second-order around a point $\hat{\boldsymbol{h}}_n$ usually taken to be the MAP estimate, where we expect most of the density to be concentrated. Assuming the channel matrix entries are i.i.d. and follow a Gaussian distribution and unit variance, define

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{H}}) &= \ln\left[\boldsymbol{P}\big(\boldsymbol{Y}_p|\boldsymbol{X}_p,\bar{\boldsymbol{H}}\big)\cdot\boldsymbol{P}\big(\bar{\boldsymbol{H}}\big)\right] \\ &= \ln\left[\prod_{n=1}^{N}\boldsymbol{P}\big(\boldsymbol{y}_{p,n}|\boldsymbol{X}_p,\bar{\boldsymbol{h}}_n\big)\cdot\boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big)\right] \\ &= \sum_{n=1}^{N}\mathcal{L}(\bar{\boldsymbol{h}}_n), \end{aligned} \tag{4.24}$$

where

$$\mathcal{L}(\bar{\boldsymbol{h}}_n) = \sum_{t=1}^{T_p}\ln\left[\Phi\left(\frac{y_{p,n}^t\bar{\boldsymbol{h}}_n^{\mathsf{T}}\boldsymbol{x}_{p,t}}{\sigma_p}\right)\right] - \tfrac{1}{2}||\bar{\boldsymbol{h}}_n||^2. \tag{4.25}$$

Then for each n in parallel we find the MAP estimate, i.e, conduct the following optimization

$$\hat{\boldsymbol{h}}_n = \underset{\bar{\boldsymbol{h}}_n\in\mathbb{R}^M}{\arg\max}\ \mathcal{L}(\bar{\boldsymbol{h}}_n). \tag{4.26}$$

This can be efficiently done using first-order gradient methods, since this objective function is concave in $\bar{\boldsymbol{h}}_n$ with the assumption that the prior distribution on the channel is Gaussian. Exponentiating and normalizing the approximation, the final result is as follows

$$\begin{aligned}\boldsymbol{P}\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p,\boldsymbol{X}_p\big) &\approx Q\big(\bar{\boldsymbol{H}}|\boldsymbol{Y}_p,\boldsymbol{X}_p\big) \\ &\approx \prod_{n=1}^{N}\phi_M\big(\bar{\boldsymbol{h}}_n;\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n\big) \qquad (4.27)\\ &\approx \prod_{n=1}^{N}\frac{1}{\hat{\boldsymbol{\omega}}_n}e^{-\frac{1}{2}(\bar{\boldsymbol{h}}_n-\boldsymbol{\mu}_n)^{\mathsf{T}}\boldsymbol{\Sigma}_n^{-1}(\bar{\boldsymbol{h}}_n-\boldsymbol{\mu}_n)}, \qquad (4.28)\end{aligned}$$

with $\phi_k(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ representing the multivariate Gaussian distribution of a k dimensional vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The Gaussian parameters in (4.28) are given by

$$\begin{cases} \hat{\boldsymbol{\omega}}_n = \sqrt{(2\pi)^M |\boldsymbol{\Sigma}_n|} \\ \boldsymbol{\mu}_n = \hat{\boldsymbol{h}}_n - (\nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)})^{-1} \nabla_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} \\ \boldsymbol{\Sigma}_n = -(\nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)})^{-1}, \end{cases} \tag{4.29}$$

with the gradient and the Hessian of $\mathcal{L}(\bar{\boldsymbol{h}})$ having the following expressions

$$\begin{cases} \nabla_{\mathcal{L}(\bar{\boldsymbol{h}}_n)} = \dfrac{1}{\sigma_p} \displaystyle\sum_{t=1}^{T_p} \kappa\left(\frac{y^t_{p,n} \bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma_p} \right) y^t_{p,n} \boldsymbol{x}_{p,t} - \bar{\boldsymbol{h}}_n \\ \nabla^2_{\mathcal{L}(\bar{\boldsymbol{h}}_n)} = - \left[\dfrac{1}{\sigma_p^2} \displaystyle\sum_{t=1}^{T_p} \eta\left(\frac{y^t_{p,n} \bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma_p} \right) \boldsymbol{x}_{p,t} \boldsymbol{x}_{p,t}^{\mathsf{T}} + I_M \right] \end{cases}$$

and $\kappa(\cdot)$, $\eta(\cdot)$ given as in (3.11) and (3.12) respectively. The detailed derivations can be found in Appendix D.

4.2.2 . Approximation of the ML Metric

After obtaining the Gaussian approximation of the channel's posterior distribution, we can now evaluate (4.19) in order to obtain a closed-form solution, where the expectation is now taken over the multivariate Gaussian in (4.28)

$$\begin{aligned} \boldsymbol{x}_{\mathrm{LA}} &= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\max}\ \mathbb{E}_Q\left[\boldsymbol{P}\left(\boldsymbol{y}_d | \boldsymbol{x}_d, \bar{\boldsymbol{H}} \right) \right] \\ &= \underset{\boldsymbol{x}_d \in \mathcal{X}^M}{\arg\min}\ -\sum_{n=1}^{N} \ln \left[\Phi\left(\frac{y_{d,n} \boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{\mu_n}}{\sqrt{\sigma_d^2 + \boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{\Sigma}_n \boldsymbol{x}_d}} \right) \right]. \end{aligned} \tag{4.30}$$

The derivation of this last equation is given in Appendix E for the sake of completeness. We highlight that this problem is still a hard optimization problem with the discrete nature of $\boldsymbol{x}_d$, in addition to the dependence of the noise on the transmitted signal. However, the evaluation of (4.30) is simpler than that of (4.19) since we no longer have high dimensional orthant probabilities. This approximation can also be seen as a quantized version of the Generalized Gaussian Model (GGM) proposed in [118] where prior to quantization, the noise second-order statistics are data dependent due to the nonlinearity. In particular, we retrieve the *linear self-interference* model as per Definition 1 in [118]

$$\boldsymbol{w} = \varrho(\boldsymbol{x}_d) + \sigma_d \boldsymbol{n}_1 + \boldsymbol{W}(\boldsymbol{x}_d) \boldsymbol{n}_2, \tag{4.31}$$

such that

$$\begin{cases} \varrho(\boldsymbol{x}_d) = \hat{\boldsymbol{H}}\boldsymbol{x}_d \\ \boldsymbol{W}(\boldsymbol{x}_d) = (\boldsymbol{I}_n \otimes \boldsymbol{x}_d^{\mathsf{T}})\boldsymbol{C}, \end{cases}$$

where $\hat{\boldsymbol{H}} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N]$, $\boldsymbol{n}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$, $\boldsymbol{n}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{NM})$ and $\boldsymbol{C} \in \mathbb{R}^{NM \times NM}$ is a block diagonal matrix of $\left[\boldsymbol{C}_1^{\mathsf{T}}, \ldots, \boldsymbol{C}_n^{\mathsf{T}}, \ldots, \boldsymbol{C}_N^{\mathsf{T}}\right]$ such that $\boldsymbol{C}_n$ is retrieved from the Cholesky decomposition of $\boldsymbol{\Sigma}_n$.

4.2.3 . Choice of Pilot Transmission Strategy

The metric in (4.30) is for a general matrix $\boldsymbol{X}_p$, although we expect that the performance to be dependent on the choice of pilot transmission strategies. In this paper we use the identity matrix such that $\boldsymbol{X}_p = \mathbb{1}_\gamma^{\mathsf{T}} \otimes \boldsymbol{I}_M$, wherein for a pilot transmission period T_p, users transmit in a round-robin fashion $\gamma = T_p/M$ symbols. We can further simplify (4.22) for each channel index n as

$$\begin{aligned} \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n | \boldsymbol{Y}_p, \boldsymbol{X}_p\big) &= \frac{1}{\boldsymbol{\omega}_n} \prod_{t=1}^{T_p} \Phi\left(\frac{y_{p,n}^t \bar{\boldsymbol{h}}_n^{\mathsf{T}} \boldsymbol{x}_{p,t}}{\sigma_p}\right) \cdot \boldsymbol{P}\big(\bar{\boldsymbol{h}}_n\big) \\ &= \prod_{m=1}^{M} \left[\prod_{k_m=1}^{\gamma} \frac{1}{\omega_{n,k_m}} \Phi\left(\frac{y_{p,n}^{k_m} \bar{h}_{n,m}}{\sigma_p}\right) \phi(\bar{h}_{n,m}) \right] \qquad (4.32) \\ &= \prod_{m=1}^{M} \left[\frac{1}{\omega_{n,m}} \Phi\left(\frac{\bar{h}_{n,m}}{\sigma_p}\right)^{\alpha_{n,m}} \Phi\left(\frac{-\bar{h}_{n,m}}{\sigma_p}\right)^{\beta_{n,m}} \phi(\bar{h}_{n,m}) \right] \qquad (4.33) \\ &= \prod_{m=1}^{M} p(\bar{h}_{n,m}), \qquad (4.34) \end{aligned}$$

where we defined the sets for each n, m pair

$$\begin{aligned} \mathcal{Y}_{n,m}^{+} &= \{y_{p,n}^{k_m} \in \bar{\boldsymbol{y}}_{p,n} |\, y_{p,n}^{k_m} = 1, k_m = 1, \ldots, \gamma\}, \\ \mathcal{Y}_{n,m}^{-} &= \{y_{p,n}^{k_m} \in \bar{\boldsymbol{y}}_{p,n} |\, y_{p,n}^{k_m} = -1, k_m = 1, \ldots, \gamma\}, \end{aligned}$$

and let $\alpha_{n,m} = |\mathcal{Y}_{n,m}^{+}|$ and $\beta_{n,m} = |\mathcal{Y}_{n,m}^{-}|$, with $\alpha_{n,m} + \beta_{n,m} = \gamma$. Equation (4.32) is obtained by a permutation or re-ordering of the T_p dimensional vector $\boldsymbol{y}_{p,n}$ containing γ one-bit observations for each channel coefficient $\bar{h}_{n,m}$. Letting $\bar{\boldsymbol{y}}_{p,n}$ denote this permuted vector, we then get (4.33) by separating positive and negative observations for each channel coefficient. This decomposition is useful, as it simplifies the numerical integration when evaluating the ML metric in (4.19) during the lower bound computation as we did for the perfect CSI case in Eq. (3.25).

First note that we can re-write

$$\mathbb{E}_{\bar{\boldsymbol{H}}|\boldsymbol{Y}_p,\boldsymbol{X}_p}\left[\boldsymbol{P}\left(\boldsymbol{y}_d|\boldsymbol{x}_d,\bar{\boldsymbol{H}}\right)\right] = \prod_{n=1}^{N}\left[\int_{\mathbb{R}^M}\Phi\left(\frac{y_{d,n}\bar{\boldsymbol{h}}_n^{\mathsf{T}}\boldsymbol{x_d}}{\sigma_d}\right)\boldsymbol{P}\left(\bar{\boldsymbol{h}}_n|\boldsymbol{Y}_p,\boldsymbol{X}_p\right)\mathrm{d}\bar{\boldsymbol{h}}_n\right]. \tag{4.35}$$

Since the channel distributions are independent for each m, this is equivalent to finding the integral, for a given $\boldsymbol{x}_d$ as

$$\prod_{n=1}^{N}\left[\int_{\mathbb{R}}\Phi(s_n)p(s_n)\,\mathrm{d}s_n\right], \tag{4.36}$$

where we define the vector

$$\boldsymbol{g}_n = \frac{y_{d,n}}{\sigma_d}[\bar{h}_{n,1}x_{d,1},\ldots,\bar{h}_{n,m}x_{d,m},\ldots,\bar{h}_{n,M}x_{d,M}], \tag{4.37}$$

and let $s_n = \sum_{m=1}^{M} g_{n,m}$ with $p(s_n)$ obtained from the M-fold linear convolution $(p_{g_{n,1}} * \cdots * p_{g_{n,M}})$. Note that in more general cases, we can also apply the MC sampling method proposed in [119] which we detail in Appendix F.

4.2.4 . Sphere-Decoding under Channel Uncertainty

The optimization problem in (4.30) is still exponential in the input dimension. To address this complexity issue we can adapt the SD algorithm from Chapter 3 to this scenario. The objective is to construct a list of candidate points over which we can evaluate the metric in (4.30). Noting that applying the technique directly is not straightforward, this is due to the additional dependence of the variance term on the transmitted data. We therefore propose the following approximation:

- **Step 1:** We consider a surrogate metric that is convex in the data vector. For example, we can take the MAP channel estimate already obtained during the LA procedure $\hat{\boldsymbol{H}}$ and formulate the mismatched ML metric similarly to Eq. (4.9)

$$\boldsymbol{x}_{\mathrm{MM}} = \underset{\boldsymbol{x}_d\in\mathcal{X}^M}{\arg\min}\; -\sum_{n=1}^{N}\ln\left[\Phi\left(\frac{y_{d,n}\boldsymbol{x}_d^{\mathsf{T}}\boldsymbol{\mu_n}}{\sigma_d}\right)\right]. \tag{4.38}$$

  This metric ignores the data-dependent noise. However, it is indeed convex in $\boldsymbol{x}_d$ when relaxed over $\mathbb{R}^M$.

- **Step 2:** We then perform SD operations and construct a list of candidates

$$\mathcal{M} = \left\{[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{|\mathcal{M}|}]\in\mathcal{X}^M \;\middle|\; ||\boldsymbol{t}-\boldsymbol{U}\boldsymbol{x}_k||^2 \le d\right\}, \tag{4.39}$$

  such that $\boldsymbol{U}$ is given by the Cholesky decomposition of the mismatched log-likelihood's Hessian of (4.38).

- **Step 3:** Finally, we perform the optimization

$$\boldsymbol{x}_{\text{LA-SD}} = \underset{\boldsymbol{x}_d \in \mathcal{M}}{\arg\min} \; -\sum_{n=1}^{N} \ln \left[\Phi \left(\frac{y_{d,n} \boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{\mu}_{\boldsymbol{n}}}{\sqrt{\sigma_d + \boldsymbol{x}_d^{\mathsf{T}} \boldsymbol{\Sigma}_n \boldsymbol{x}_d}} \right) \right], \tag{4.40}$$

  where we retrieve the Laplace approximation sphere-decoding (LA-SD) solution. Although this heuristic is dependent on the choice of the surrogate metric, it will still provide us with a list of candidate points where we can evaluate (4.30).

4.2.5 . Computational Complexity of the LA Approach

Similarly as in the perfect CSI case, we analyze the computational complexity by counting the theoretical number of operations in different stages. The first step is to obtain the initial estimate in (4.26) which is done using gradient descent and grows in the order of $O(LMNT_p)$ where L is te number of required iterations. The second step requires constructing the LA approximate metric after obtaining the parameters in (4.29) which is dominated by computing the inverse of the Hessian for each channel vector and scales as $O(M^3N)$. Finally, we will need to compute the final LA metric in (4.30) which scales as $O(|\mathcal{X}|^M M^2 N)$ when spanning over the entire constellation and can be reduced to $O(|\mathcal{M}|M^2N)$ when applying the SD algorithm.

4.3 . Simulation Results

4.3.1 . Data Detection with Probit Regression Framework

For the simulation setup, we first investigate the performance of QPSK signaling and compute the matched and mismatched achievable rates for a SIMO channel with one transmitter $M = 1$, and $N = 8$ receive antennas. The step sizes are chosen similarly as described in Chapter 3 for the perfect CSI case based on the Lipschitz constant. The optimization over the parameter s in the GMI is done numerically by discretizing it over a fixed range and then taking the maximum rate among this range. It is possible to compute the exact GMI rates which are plotted in Fig. 4.1 along with the MC simulations for comparison. We see that the MC simulations are close to the exact computations. We also look at the SER for data detection using the procedure described earlier for a 4×32-MIMO system with training lengths $T_p = \{5, 10, 20\}$ and compare that with the exact ML and BLMMSE metric from Eq. (2.23) (see Fig. 4.2). We show the results for the exhaustive mML for $T_p = 20$. The BLMMSE estimator is accurate for low SNRs, in this regime the additive Gaussian noise dominates and the Bussgang approximation is

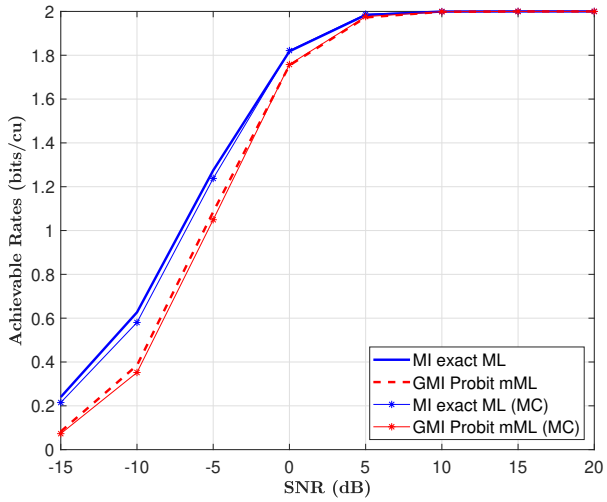

Figure 4.1: Achievable rates with matched ML and mismatched metric with QPSK signaling for 1×8-SIMO and $T_p = 20$ (exact computations and MC simulations)

accurate. As the SNR increases, it can be observed that its performance depreciates compared to that of the probit model. We note that the initial results reported in [120] show a higher instability in that regime which was due to a missing scale factor when computing the estimator. This instability vanishes after including the proper scaling and using the logistic approximation of the Gaussian CDF proposed in [40]. Nevertheless, the conclusion remains the same: the choice of the estimator is what influences the overall performance since we are still using the Gaussian CDF in our model. We also re-examine the mismatched rates for the probit and the BLMMSE metrics as reported in [120] (Figure 4 therein) which we plot in Fig. 4.4 for the same 4×32-MIMO system. We see that after including the correct scaling, the instability in the mid to high SNR regime indeed vanishes. In general, we observe that the performance improves with increasing T_p. A similar behavior is echoed in Fig. 4.3 as the maximum mismatched achievable rates for the given uniform input distribution improve with increasing T_p. We see that the GMI based on the identified model parameters approaches that of the MI with perfect CSI. This confirms that the probit model is indeed stable even with imperfect CSI, in contrast to what has been reported recently [40]. In Fig 4.5, we look at the SER performance in a 16-QAM 2×32-MIMO system, we observe similar behavior in the mid to high SNR regime. The mismatched rates are also plotted in Fig. 4.6. We remark on the following observations. First, the SERs for all metrics, including that with perfect CSI

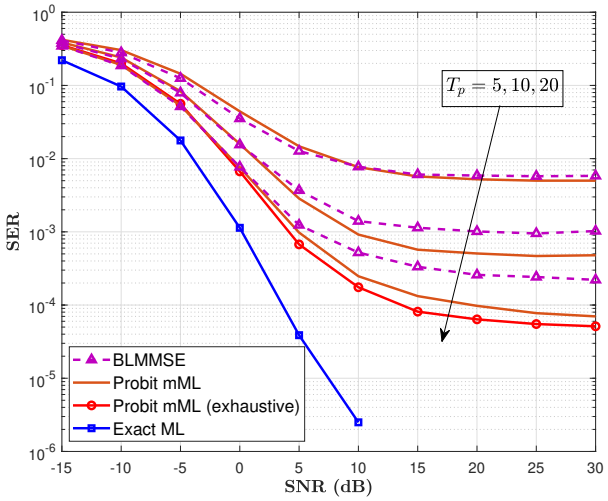

Figure 4.2: SER for exact ML, mismatched probit and BLMMSE with QPSK 4×32-MIMO and increasing training lengths $T_p = \{5, 10, 20\}$.

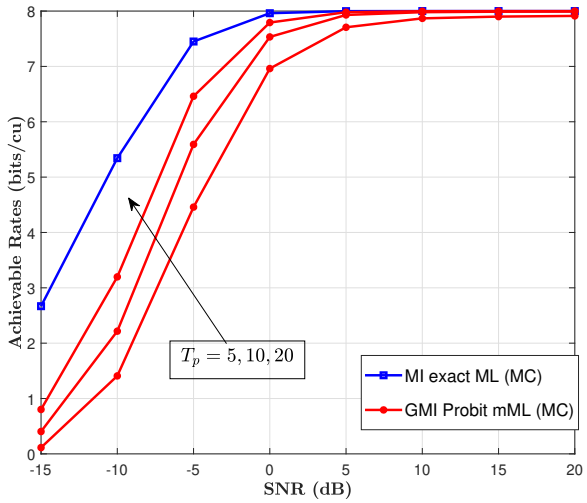

Figure 4.3: GMI for exact ML and mismatched probit model for a 4×32-MIMO with QPSK and increasing training lengths (MC simulations).

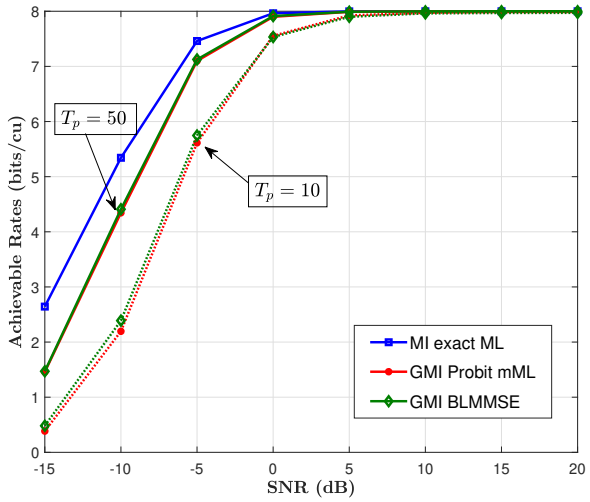

Figure 4.4: Achievable rates under the mismatched Probit and BLMMSE metrics for a 4×32-MIMO with QPSK and $T_p = \{10, 50\}$.

case, exhibit a minimum with respect to the SNR, this can be justified by the following fact: considering the simple scenario where we transmit a point from a 4-PAM constellation, as the SNR increases it becomes evident that some constellation points will "collapse" onto each other resulting in the ML declaring an error. A similar explanation is provided in the recent work where this particular problem is treated [121]. Moreover, this behavior is frequently encountered in the literature on one-bit ADCs, e.g. [46, 73, 75]. Second, the BLMMSE estimator marginally outperforms the probit model in the low SNR regime, this can be attributed to the fact that we do not take into account the SNR during the optimization, i.e., the step size is fixed for all SNR values.

4.3.2 . Data Detection with Statistical CSIR

We now consider the situation where only the channel statistics are known at the receiver and adopt the identity pilot transmission strategy. Figure 4.7 shows the VER for a 2×64 real MIMO system with 4-PAM signaling and training lengths $T_p = \{52, 100\}$. We show in the same figure the performance of the proposed LA metric (4.30) along with the LA-SD method to reduce the search space (4.40). As a benchmark, we also present the VER results with perfect CSI, the mismatched ML from (4.38) with the mismatched channel estimate and the Bussgang linear minimum mean square error (BLMMSE) channel estimator proposed in [68]. Several remarks are in order. We first note that VERs for all metrics also exhibit a minimum with respect to the SNR for the reason explained

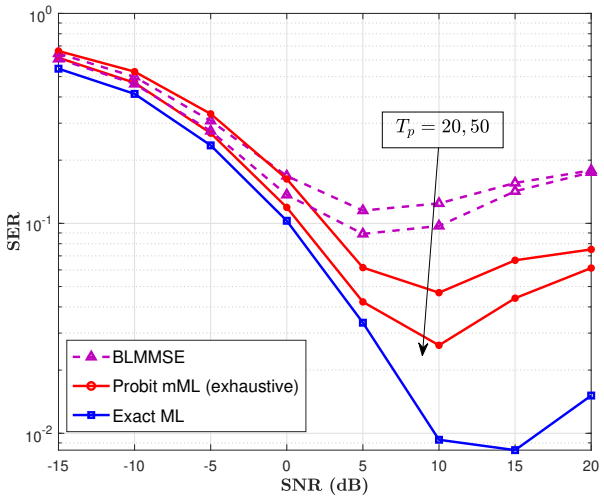

Figure 4.5: SER for exact ML, mismatched probit and BLMMSE with 16-QAM 2×32-MIMO and $T_p = \{20, 50\}$.

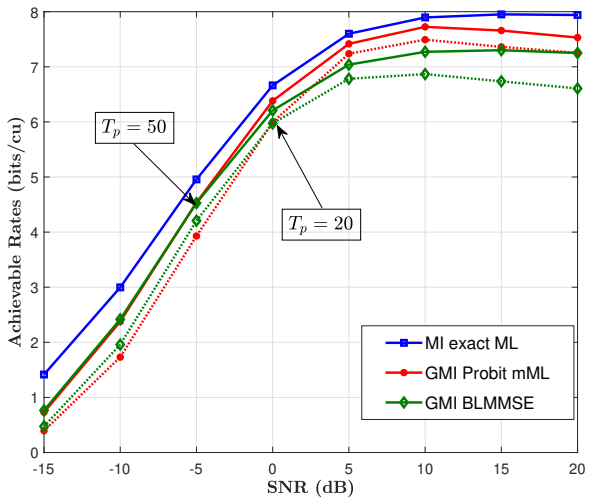

Figure 4.6: Achievable rates under the mismatched Probit and BLMMSE metrics for a 2×32-MIMO with 16-QAM and $T_p = \{20, 50\}$.

earlier. Secondly, for training lengths of 52 and 100, the proposed LA solution is tight with the lower bound of the optimal metric in (4.19). This is conserved as we reduce the search space and employ the LA-SD solution which uses the mismatched ML metric as a surrogate for constructing the list, with $|\mathcal{M}| = 2$ in this setting. Moreover, it can be observed that, compared to other metrics, the LA solution remains near-

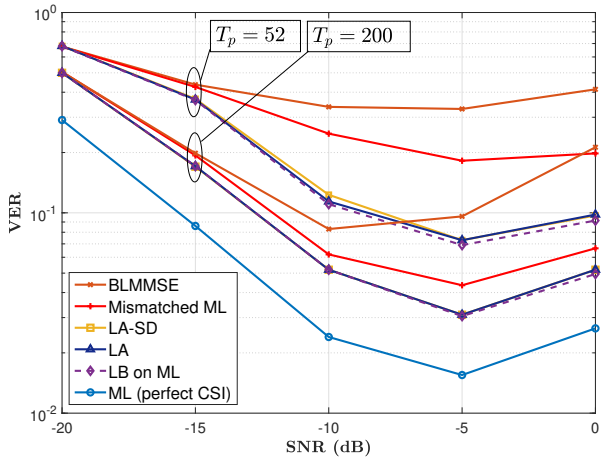

Figure 4.7: VER for a 2×64 real MIMO system with varying SNR and increasing T_p.

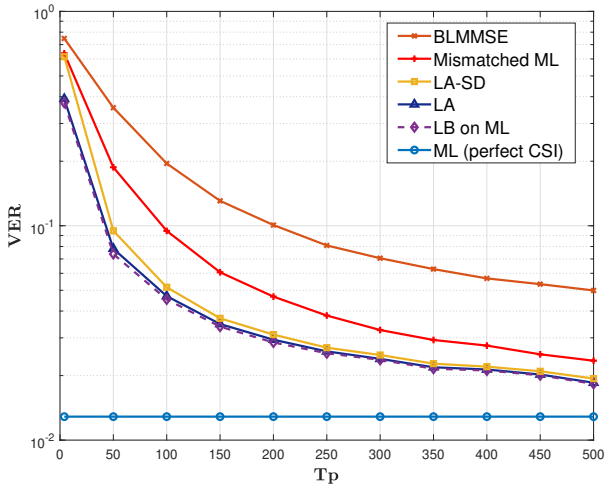

Figure 4.8: VER performance with respect to increasing training lengths for a 2×64 real MIMO channel, 4-PAM signaling, SNR is fixed to -5dB, and $|\mathcal{M}| = 2$ points.

optimal even as we cut the training length in half. We investigate this behavior by fixing the SNR to -5dB and varying the training period as shown in Fig. 4.8. Indeed, the LA solution is tight with the ML lower bound, and as T_p increases, it approaches the performance of the coherent ML detector, as expected. Increasing the number of transmit antennas and keeping N fixed, similar behavior is echoed for a 4×64 real MIMO system in Fig. 4.9. This indicates that the proposed metric

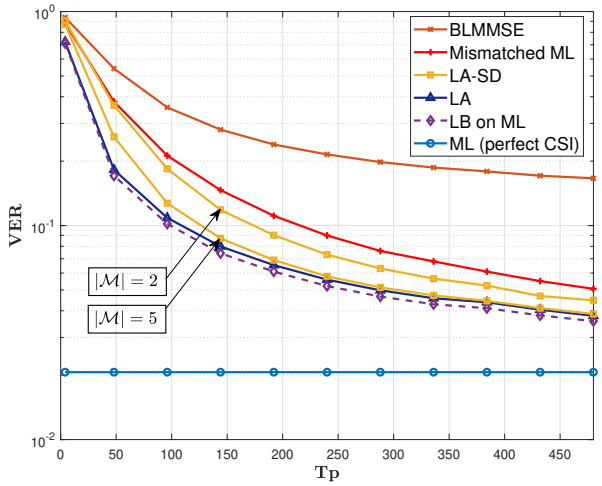

Figure 4.9: VER performance with respect to increasing training lengths for a 4×64 real MIMO channel, 4-PAM signaling, SNR is fixed to -5dB, and $|\mathcal{M}| = 2, 5$ points.

remains robust for small training periods and an increased number of users, as it takes advantage of the learned channel statistics without explicitly estimating it. In both scenarios, the LA-SD metric remains near-optimal, with a deviation in the regime when T_p is small, this can be attributed to the fact that it uses the estimated channel from the mismatched ML for the list construction which can be inaccurate.

4.4 . Summary

In this chapter, we looked at the data detection problem where CSI is unavailable at the receiver. We considered two schemes in block-fading Rayleigh channels where transmission is decomposed into two independent pilot and data frames. The first is a two-stage channel estimation and data detection approach based on the probit regression binary classification framework. Using unitary pilot matrices, the channel is first estimated and used directly as a replacement for the true channel during the detection process. The results are compared with an estimator of the channel obtained based on the Bussang decomposition. In general, the performance improves with increasing training length for QPSK and 16-QAM uniform signaling, with larger gaps in the high SNR regime. Moreover, we address the observation noted in [40] where instability is attributed to the Gaussian CDF under imperfect CSI. It's shown that, instead, the instability is due to the choice of the estimator itself. In ad-

dition, we looked at the mismatched achievable rates with the GMI using the formulated mismatched metrics. It is seen again that increasing the training lengths allows us to approach the achievable rates of matched communication with perfect CSI. In the second part, we move to a more general situation where the pilot and data frames are processed jointly. The optimal metric is then formulated for a real channel model where we establish two main difficulties. The first involves evaluating multivariate Gaussian orthant probabilities which do not have closed forms. The second is detection complexity under the discrete nature of the input which is prohibitive for large dimensions similar to the perfect CSI case. We address the first problem with an approximation using the Laplace method which allows us to obtain a closed-form expression. We then apply the SD algorithm proposed in Chapter 3 to the Laplace approximation through the mismatched ML as a surrogate metric. Assuming a 4-PAM signaling, it's observed that the LA metric is close to optimal with respect to the lower bound on the exact optimal metric with an overall performance approaching the coherent case as the training length increases.

Chapter 5

Communication with Multi-Bit Low-Resolution ADCs

In the previous chapters, we have considered the data detection problem in the extreme case where one-bit ADCs are employed at the receiver, under different assumptions on the channel state information. Understanding the fundamental limitations of communication systems guides their design and provides insights into how to optimize their performance. Since the pivotal work of Claude Shannon in 1948 [122], information theory has become a rich field that paved the path for the conception of now standardized and indispensable data compression and channel coding schemes [123]. The channel capacity is defined as the ultimate limit under which information can be transferred with an arbitrarily vanishing error probability. Characterizing the capacity of a system is challenging from several aspects, as it first requires obtaining a simple yet faithful statistical model of the physical channel model, and second, solving the optimization problem is generally difficult. For example, the effect of channel fading has been extensively studied in the past under several assumptions on the CSI. In fact, tractable expressions even for basic channel models such as SISO Rayleigh fading channel under an average power constraint are not known, and only a characterization of the capacity-achieving distribution is available [124]. This optimization problem can become more intractable as we move to MIMO channels. Nevertheless, insightful results can still be obtained by looking at particular cases or asymptotic regimes in the SNR, number of transmit and receive antennas, or the coherence interval for block-fading channels [93, 113, 125].

The capacity problem of the communication channel while incorporating effects of low-resolution ADCs at the receiver has also received attention in recent years. One of the earliest works considers the AWGN quantized output SISO channel with an average input power constraint [44]. The authors show that for a symmetric quantizer with a resolution b and 2^b output levels, the capacity-achieving distribution is discrete with $2^b + 1$ mass points. For the one-bit case (i.e. $b = 1$), binary antipodal signaling is shown to be optimal, and an exact expression of the capacity is obtained as a function of the SNR. For $b > 1$, no analytical expression is available and the capacity can only be computed numerically, given a fixed quantizer design. For MIMO channels with deterministic

channel state realization, upper and lower bounds on the capacity have been proposed in [45]. In [67], the authors study the capacity of the quantized MIMO case by finding the asymptotic expression of the mutual information in the low SNR regime, where they consider particular cases of the ergodic and noncoherent channels. Achievable rates have also been obtained based on the Bussgang decomposition and AQNM models, along with specific choices of linear receivers [46, 68, 97]. It is evident then that asymptotic expressions and bounding techniques on the capacity can prove useful in providing insightful results on capacity scaling in certain regimes.

Another key aspect of future wireless communication systems design that merits attention is power consumption, particularly that of the receiver [47–49]. Therefore, there has also been recent interest in studying the communication rate and quantization resolution trade-off to obtain insight into practical limitations of the communication system design [85, 97, 126, 127]. The work therein sets assumptions on the choice of the quantizer or the signaling scheme usually taken to be Gaussian. It would be interesting to explore a similar direction where we obtain, as a first step, a tractable expression of the quantized MIMO channel capacity with less strict assumptions. The results by Clarke and Barron [128, 129] could provide us with the necessary tools for that goal. They have been used in [130, 131] to obtain capacity scalings for the coherent and noncoherent one-bit quantized channel in the asymptotic regime where the number of receive antennas is large which could be of particular interest for mMIMO. The results therein could potentially be extended for the multi-bit case. In this chapter, we investigate a direction towards that goal. First, for the coherent case, we re-visit the data detection problem and extend the SD algorithm in the scenario where multi-bit ADCs are employed at the receiver by redefining the likelihood function for the multi-bit quantizer. Then, the asymptotic capacity expression can be found similarly as in [131]. For the noncoherent channel, we start by looking at the unquantized output for special cases of the coherence interval as upper bounds on the one-bit asymptotic channel capacity.

5.1 . Data Detection Extension to Multi-Bit ADCs

Assume now that the receiver is equipped with a uniform midriser quantizer where $b \geq 2$. Consider now the negative log-likelihood function in Eq. (2.29), as in the binary case, we also compute the gradient and

Hessian functions that can be shown to take the following form

$$\begin{cases} \nabla_{\ell_b(\boldsymbol{x})} = -\dfrac{1}{\sigma}\displaystyle\sum_{n=1}^{2N} \kappa_{b,\delta}\left(\dfrac{I_{\mathsf{y}_n} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}, \dfrac{I_{\mathsf{y}_n-1} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\boldsymbol{h}_n, & (5.1) \\ \nabla^2_{\ell_b(\boldsymbol{x})} = \dfrac{1}{\sigma^2}\displaystyle\sum_{n=1}^{2N} \eta_{b,\delta}\left(\dfrac{I_{\mathsf{y}_n} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}, \dfrac{I_{\mathsf{y}_n-1} - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}}{\sigma}\right)\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}}, & (5.2) \end{cases}$$

with similar functions defined as

$$\begin{cases} \kappa_{b,\delta}(u,v) = -\dfrac{\phi(u) - \phi(v)}{\Phi(u) - \Phi(v)}, & (5.3) \\ \eta_{b,\delta}(u,v) = \dfrac{u\phi(u) - v\phi(v)}{\Phi(u) - \Phi(v)} + (\kappa_{b,\delta}(u,v))^2. & (5.4) \end{cases}$$

We highlight that these functions implicitly depend on the resolution b and the step size δ. We can perform the Taylor expansion of the log-likelihood function in Eq. (2.29) around an initial estimate and follow similar steps as in Algorithm 1, i.e., obtain an $\hat{\boldsymbol{x}}$ using gradient iterations and employ the Hessian evaluated at that point to construct the list of candidate points. It can be shown that this extension generalizes the proposed approach as it captures the extreme case where we only have one bit of resolution, and recovers the classical SD problem for the unquantized channel as the quantization becomes finer. To see the latter case, first, observe that Eq. (5.3) and (5.4) are functions of finite difference terms centered around v and depend on the step size δ. As the resolution $b \to \infty$ and $\delta \to 0$ we get for the infinite resolution case $\kappa_\infty \approx v$ and $\eta_{b,\delta} \approx 1$ which result in the following expressions for the gradient and Hessian functions

$$\begin{cases} \nabla_{\ell_\infty(\boldsymbol{x})} = -\dfrac{1}{\sigma}\displaystyle\sum_{n=1}^{2N} \boldsymbol{h}_n(y_n - \boldsymbol{h}_n^{\mathsf{T}}\boldsymbol{x}) = -\dfrac{1}{\sigma}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}), & (5.5) \\ \nabla^2_{\ell_\infty(\boldsymbol{x})} = \dfrac{1}{\sigma^2}\displaystyle\sum_{n=1}^{2N} \boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}} = \dfrac{1}{\sigma^2}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}. & (5.6) \end{cases}$$

It's now evident that by setting Eq. (5.5) to 0 we obtain the well-known least-squares or ZF solution

$$\boldsymbol{x}_{\mathrm{ZF}} = \left(\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{y}, \tag{5.7}$$

along with the Euclidean distance metric for the classical SD problem

$$\boldsymbol{x}_{\mathrm{ML}} = \underset{\boldsymbol{x}\in\mathcal{X}^{2M}}{\arg\min} \quad (\boldsymbol{x} - \boldsymbol{x}_{\mathrm{ZF}})^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{ZF}}) \tag{5.8}$$

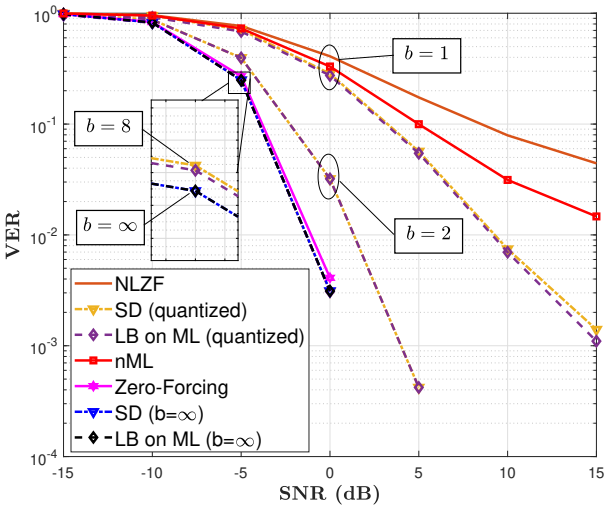

Figure 5.1: VER for 4×64-MIMO with 16-QAM, perfect CSI and varying SNR for several data detection metrics, and assumptions on the resolution b. The list size is fixed to $|\mathcal{S}| = 5$.

We show in Fig. 5.1 for a 16-QAM constellation and a 4×64-MIMO channel simulation results with several assumptions on the quantizer resolution b and a fixed list size of $|\mathcal{S}| = 5$. The step size δ is set to minimize the mean-square error distortion at the output assuming a Gaussian input signal [101]. The SD algorithm is applied in the same manner as before with the adapted gradient and Hessian functions from Eq. (5.1) and (5.2), respectively. As the resolution increases, the VER performance for our proposed SD algorithm improves while approaching that of the unquantized case when the resolution is equal to 8 bits. In the latter case, we show both the exact ZF detector from Eq. (5.7). Note that adding one additional bit of resolution from $b = 1$ to $b = 2$ significantly improves the detection performance, which implies that the effect of employing low-resolution ADCs might not be very severe assuming the base station is equipped with a sufficient number of receive antennas. We next look at the computational complexity of the SD algorithm as a function of the resolution and number of antennas at the receiver. Given that other measures of performance, e.g. running time, can highly depend on the machine and number of processors in use, we focus on counting the average number of floating point operations n_{flops} taken by the algorithm to output the list of candidates for a given fixed size. Fig. 5.2 shows how the number of operations scales while increasing $|\mathcal{S}|$ with different assumptions on the quantizer resolution and number of receiver antennas for $M = 8$ transmitters and a 16-QAM constellation. We observe two

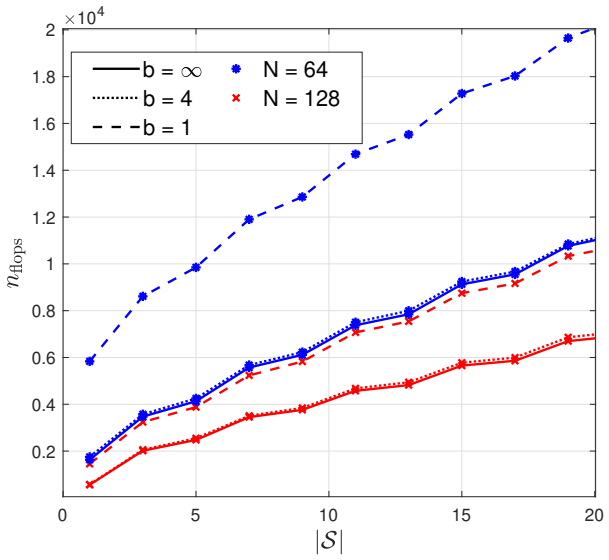

Figure 5.2: Scaling of average floating point operations with list size $|\mathcal{S}|$ in a $8 \times N$-MIMO system, 16-QAM with $M = 8$, SNR$=-5$ dB, and different assumptions on b and N.

interesting behaviors: the number of operations decreases as the number of receive antennas increases, for a fixed transmission scheme and quantizer resolution, indicating the benefit of mMIMO systems. For example, the number of operations with one-bit quantization for a 8×128 is close to that of an unquantized system with only 64 receive antennas. Moreover, a resolution of $b = 4$ seems to be sufficient in approaching the performance of the unquantized channel for both MIMO setups. cc

5.2 . Asymptotic Capacity of Quantized MIMO Channel

We begin by formulating the capacity problem for a real quantized MIMO channel without a priori CSI before presenting the theorem provided in [129]. The channel is assumed to be block fading, stationary, ergodic, and without loss of optimality in terms of capacity, the source can be considered memoryless

$$\boldsymbol{Y}_{N\times T} = \mathsf{Q}_b(\boldsymbol{H}_{N\times M}\boldsymbol{X}_{M\times T} + \boldsymbol{Z}_{N\times T}). \tag{5.9}$$

where we have $h_{nm} \sim \mathcal{N}(0,1)$, $z_{nm} \sim \mathcal{N}(0,\sigma^2)$. From a practical aspect, we usually impose a power constraint on the input set $\mathcal{X}$. In our setting, we assume a peak-power constraint on the columns of $\boldsymbol{X}$, i.e., for any $\boldsymbol{X} \in \mathcal{X}$ we need the columns to satisfy $\{||\boldsymbol{x}_i||^2 \leq \rho, i = 1,\ldots,T\}$. We employ the following Shannon capacity definition of this channel written

as a functional of the input distribution $P_{\boldsymbol{X}}$ and the channel law $P_{\boldsymbol{Y}|\boldsymbol{X}}$

$$C = \max_{P_{\boldsymbol{X}} \in \mathcal{P}} \boldsymbol{I}(P_{\boldsymbol{X}}; P_{\boldsymbol{Y}|\boldsymbol{X}}), \tag{5.10}$$

where $\boldsymbol{I}$ represents the mutual information between the matrix valued random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\mathcal{P}$ is the set of input distributions on $\boldsymbol{X} \in \mathcal{X}$. Obtaining a closed-form expression of Eq. (5.10) is difficult, in general, and as discussed earlier only asymptotic expressions or approximations are known. Let us introduce the following Markov chain

$$\boldsymbol{X} \to \boldsymbol{\theta} \to \boldsymbol{Y}, \tag{5.11}$$

where $\boldsymbol{\theta}$ is a function of $\boldsymbol{X}$ and defined over the space

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}(\boldsymbol{X}) : \boldsymbol{X} \in \mathcal{X}\} \subseteq \mathbb{R}^d. \tag{5.12}$$

We can parameterize the set of the family of output distributions conditional on $\boldsymbol{X}$ by choosing $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and denoting each member of this family as $f_{\boldsymbol{\theta}}(\boldsymbol{Y})$. We emphasize that $f_{\boldsymbol{\theta}}$ is a function of $\boldsymbol{Y}$ but is omitted here for notation simplicity. From the *data-processing inequality* we can upper bound the mutual information as

$$I(P_{\boldsymbol{X}}; P_{\boldsymbol{Y}|\boldsymbol{X}}) \leq I(P_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}}), \tag{5.13}$$

where $P_{\boldsymbol{\theta}}$ is the distribution over $\boldsymbol{\theta}$, then the capacity of the channel in (5.10) can be upper bounded by

$$\begin{aligned} C \leq \underbrace{\sup_{P_{\boldsymbol{\theta}}} \inf_{\boldsymbol{Q}} D(f_{\boldsymbol{\theta}} || \boldsymbol{Q} | P_{\boldsymbol{\theta}})}_{\underline{C}_N} &\overset{(a)}{\leq} \inf_{\boldsymbol{Q}} \sup_{P_{\boldsymbol{\theta}}} D(f_{\boldsymbol{\theta}} || \boldsymbol{Q} | P_{\boldsymbol{\theta}}) \\ &= \underbrace{\inf_{\boldsymbol{Q}} \sup_{\boldsymbol{\theta}} D(f_{\boldsymbol{\theta}} || \boldsymbol{Q})}_{\bar{C}_N} \end{aligned} \tag{5.14}$$

where $\boldsymbol{Q}$ is any other distribution on $\boldsymbol{Y}$. The following result by Clarke and Barron, originally obtained in the context of defining least informative priors in Bayesian statistics, can be used here to show that inequality (a) becomes an equality, asymptotically in N.

Theorem 1 (Clarke and Barron [129])

$$\lim_{N \to \infty} \left[\bar{C}_N - \frac{d}{2} \log \frac{N}{2\pi e} \right] = \log \int_{\boldsymbol{\theta}} |J(\boldsymbol{\theta})|^{\frac{1}{2}} \, \mathrm{d}\boldsymbol{\theta}, \tag{5.15}$$

and similarly the minimax divergence has the same asymptotic value

$$\lim_{N \to \infty} \left[\underline{C}_N - \frac{d}{2} \log \frac{N}{2\pi e} \right] = \log \int_{\boldsymbol{\theta}} |J(\boldsymbol{\theta})|^{\frac{1}{2}} \, \mathrm{d}\boldsymbol{\theta}. \tag{5.16}$$

which implies that, asymptotically, we can find the saddlepoint capacity solution that is uniquely achieved when the optimal *auxiliary* input distribution $P^*_{\boldsymbol{\theta}}$ is the Jeffreys prior $\frac{1}{c}|J(\boldsymbol{\theta})|^{\frac{1}{2}}$. This result is satisfied under certain *regularity* conditions including twice differentiability of $f_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$, the Fisher information matrix $J(\boldsymbol{\theta})$ is positive definite and the family of distributions of $f_{\boldsymbol{\theta}}$ has a one-to-one mapping, i.e., for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ we have $f_{\boldsymbol{\theta}} \neq f_{\boldsymbol{\theta}'}$. These results can be applied for the coherent and non-coherent one-bit quantized channel as shown in [130] and we take a first step in the direction of extending them for the multi-bit case in order to analyze the spectral efficiency and power consumption tradeoff of a communication system.

5.3 . Multi-bit case Receiver Design: Spectral and Energy Efficiency Constraints

Beyond 5G and 6G system design goals are leaning more towards ensuring higher spectral efficiencies while maintaining minimum incurrences in RF power consumption particularly that of the receiver. We have seen that energy efficiency usually comes at the compromise of higher incurred signal distortion which could be due to power amplification, phase noise, or coarse quantization. Consequently, the spectral efficiency can be penalized and an analysis of the trade-off is needed [42, 44]. On the one hand, information-theoretic tools can be leveraged to obtain approximations or bounds on the spectral efficiency of a communication system. On the other hand, this necessitates a faithful yet simple statistical representation of the physical model at hand to incorporate the effects of power consumption. Theorem 1 in the previous section can be useful in obtaining similar results in the regime where the number of receive antennas is large. Such an approximation is not too severe as it's natural for mMIMO or distributed processing systems where the number of receivers can be very high. Moreover, the signaling scheme is not assumed to follow any specific distribution such as Gaussian. In this section, we would like to study the trade-off between the spectral efficiency and the energy consumption of the receiver in a MIMO system assuming for simplicity quantizer distortion only. The work in [126] considers this problem for a SISO link with a line-of-sight transmission. The target transmission rate is set as $R = \mathsf{SE} \times f_s$ where SE is the required spectral efficiency that depends on the SNR and quantizer resolution of the link, and f_s is the Nyquist sampling rate. Assuming an average power constraint and a uniform quantizer, the capacity can be obtained numerically using the cutting-plane algorithm [44]. In this work, the authors assume only the power dissipation due to the ADC. For sampling frequencies between 1

and 100 GHz, power dissipation scale quadratically with the bandwidth and effective number of quantization levels, therefore

$$\begin{aligned}\mathcal{P}_{\mathsf{ADC}} &= \mathsf{const} \cdot 2^{2b} \cdot f_s^{\nu} \\ &= \mathsf{const} \times \left(\frac{R}{\mathsf{SE}}\right)^{\nu} \times 2^{2b}. \end{aligned} \tag{5.17}$$

where $\nu = \{1, 2\}$ is a parameter that reflects the different state-of-the-art power consumption scaling in sampling frequency of the quantizer. For a fixed throughput and power constraint Ω_s, the authors retrieve the ADC parameterization pair $(b_{\mathsf{opt}}, f_{s,\mathsf{opt}})$ that minimizes its power dissipation. For any fixed resolution, $\mathsf{SE}_{\mathsf{max}}$ minimizes the power dissipation of the ADC which is a function of b and SNR. With the latter depending on the average power constraint and sampling rate, this results in an optimization problem that first minimizes the power dissipation over b with SE numerically and then obtains the optimal sampling frequency from the optimal SE. The work in [97] approaches the same problem for a MIMO system assuming the AQNM decomposition presented in Chapter 2 along with transmit and receive beamforming architectures including analog and digital. The output is decomposed as

$$\boldsymbol{y}_q = \tilde{\boldsymbol{H}}\boldsymbol{x} + \tilde{\boldsymbol{n}} \tag{5.18}$$

where $\tilde{\boldsymbol{H}} = (1-\beta)\boldsymbol{H}$ and $\tilde{\boldsymbol{n}} = (1-\beta)\boldsymbol{n} + \boldsymbol{e}$. Assuming the input is Gaussian distributed with covariance $\boldsymbol{C_x}$, the noise $\tilde{\boldsymbol{n}}$ is also Gaussian with the covariance $\boldsymbol{C}_{\tilde{\boldsymbol{n}}} = (1-\rho)(\rho\mathsf{diag}(\boldsymbol{C_r}) + (1-\rho)\boldsymbol{C_n})$, and the system operates over a bandwidth B, the achievable rate is given as

$$\boldsymbol{E_H}\left[B\log\det\left\{\boldsymbol{I} + (\rho\mathsf{diag}\left(\boldsymbol{H}\boldsymbol{C_x}\boldsymbol{H}^{\mathsf{H}}\right) + B\boldsymbol{C_n})^{-1}(1-\rho)\boldsymbol{H}\boldsymbol{C_x}\boldsymbol{H}^{\mathsf{H}}\right\}\right] \tag{5.19}$$

and is maximized in terms of the input covariance matrix while assuming analog and digital combining architectures along with the incurred penalty in terms of energy efficiency. The authors include power consumption induced by the LNA, mixer circuitry and ADC that scale with N. In [49], they follow a similar approach while including saturation distortion effects. A lower bound on the spectral efficiency is obtained based on the Bussgang decomposition and mainly depends on the SNDR. The receiver energy efficiency is defined as the ratio of the rate and the power consumption of LNA, saturation, and ADC modeled according to state-of-the-art FoM available in the literature. The results therein show that, at low resolutions, the noise figure from the LNA is the dominant noise factor and the resolution can be increased to improve the spectral

efficiency. For high resolutions, the thermal noise is the dominant term which limits the spectral efficiency.

We consider here coherent communication where CSI is only available at the receiver and a peak power constraint on the transmitted signal. The capacity for the block-fading channel with side information normalized by the coherence interval T written directly as a function of the random variables is

$$\begin{aligned} C &= \max_{P_{\boldsymbol{X}}} \frac{1}{T} I(\boldsymbol{X}; \boldsymbol{Y}, \boldsymbol{H}) \\ &= \max_{P_{\boldsymbol{X}}} \frac{1}{T} I(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T | \boldsymbol{H}) \\ &= \max_{P_{\boldsymbol{x}}} I(\boldsymbol{x}; \boldsymbol{y} | \boldsymbol{H}). \end{aligned} \tag{5.20}$$

The channel law is given by the joint distribution $P(\boldsymbol{y}, \boldsymbol{H}|\boldsymbol{x})$ which we can decompose as $\prod_{n=1}^{N} P(\boldsymbol{y}_n, \boldsymbol{h}_n|\boldsymbol{x}) P(\boldsymbol{h}_n)$ since the noise is also assumed i.i.d. across antennas. We drop the n subscript in the following as it becomes immaterial. Assuming a symmetric quantizer as described in Chapter 2, we can obtain the Fisher matrix by taking the expectation of negative the Hessian of the log-likelihood function which we obtained in Eq. (5.2) up to a sign. Associating the quantized received vector from $y \in \mathcal{Y}_b$ to $\mathsf{y} = \{1, 2, \ldots, 2^b\}$, defining $u_{\mathsf{y}} = I_{\mathsf{y}} - \boldsymbol{h}^{\mathsf{T}} \boldsymbol{x}$ and $v_{\mathsf{y}} = I_{\mathsf{y}-1} - \boldsymbol{h}^{\mathsf{T}} \boldsymbol{x}$, we can write

$$\begin{aligned} \boldsymbol{J}_{b,\delta}(\boldsymbol{x}) &= \mathbb{E}_{P(\boldsymbol{y},\boldsymbol{h}|\boldsymbol{x})} \left[\nabla^2_{\ell_b(\boldsymbol{x})} \right] \\ &= \mathbb{E}_{\boldsymbol{h}} \Bigg\{ \Bigg[\sum_{\mathsf{y}=1}^{2^b} \left[\frac{u_{\mathsf{y}} \phi(u_{\mathsf{y}}) - v_{\mathsf{y}} \phi(v_{\mathsf{y}})}{\Phi(u_{\mathsf{y}}) - \Phi(v_{\mathsf{y}})} + \left(\frac{\phi(u_{\mathsf{y}}) - \phi(v_{\mathsf{y}})}{\Phi(u_{\mathsf{y}}) - \Phi(v_{\mathsf{y}})} \right)^2 \right] \\ &\qquad \times P(\mathsf{y}|\boldsymbol{h}, \boldsymbol{x}) \Bigg] \boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} \Bigg\} \\ &= \mathbb{E}_{\boldsymbol{h}} \left[\sum_{\mathsf{y}=1}^{2^b} \frac{\left(\phi(u_{\mathsf{y}}) - \phi(v_{\mathsf{y}})\right)^2}{\Phi(u_{\mathsf{y}}) - \Phi(v_{\mathsf{y}})} \boldsymbol{h}\boldsymbol{h}^{\mathsf{T}} \right] \end{aligned} \tag{5.21}$$

where the last equation is obtained given the assumption of a symmetric midriser quantizer. We highlight that the Fisher matrix depends on the choice of the quantizer design specified by b and δ. Similarly as in [131], let $\boldsymbol{g} = \boldsymbol{V}^{\mathsf{T}} \boldsymbol{h}$ where $\boldsymbol{V} = \left[\frac{\boldsymbol{x}}{||\boldsymbol{x}||} \; \tilde{\boldsymbol{V}} \right]$ is unitary and define

$$\xi_{b,\delta}(s) = \sum_{\mathsf{y}=1}^{2^b} \frac{\left(\phi(I_{\mathsf{y}} - s) - \phi(I_{\mathsf{y}-1} - s)\right)^2}{\Phi(I_{\mathsf{y}} - s) - \Phi(I_{\mathsf{y}-1} - s)}, \tag{5.22}$$

we obtain denoting g_1 as the first element in $\boldsymbol{g}$

$$\boldsymbol{J}_{b,\delta}(\boldsymbol{x}) = \boldsymbol{V}\mathbb{E}_{\boldsymbol{g}}\left[\xi_{b,\delta}(||\boldsymbol{x}||g_1)\boldsymbol{g}\boldsymbol{g}^{\mathsf{T}}\right]\boldsymbol{V}^{\mathsf{T}}, \tag{5.23}$$

Following the same notation as in [131], the determinant can be written as

$$|\boldsymbol{J}_{b,\delta}(\boldsymbol{x})| = \zeta_0^{b,\delta}(||\boldsymbol{x}||)^{M-1}\zeta_2^{b,\delta}(||\boldsymbol{x}||) \tag{5.24}$$

where $\zeta_k^{b,\delta}(s) = \mathbb{E}\left[g^k\xi_{b,\delta}(sg)\right]$ where $g \sim \mathcal{N}(0,1)$. We note the additional dependence of the capacity on the choice of the quantizer in terms of b and δ. Applying Theorem 1 in [131] for the multi-bit quantized coherent channel, the asymptotic capacity expression can be written as [131]

$$C = \frac{M}{2}\log\frac{N}{2\pi e} + \log\alpha_{\rho,M}^{b,\delta} + \log V_M + o(1) \tag{5.25}$$

where

$$\alpha_{\rho,M}^{b,\delta} = \int_0^{\sqrt{\rho}} \zeta_0^{b,\delta}(r)^{\frac{M-1}{2}}\zeta_2^{b,\delta}(r)^{\frac{1}{2}}r^{M-1}\,\mathrm{d}r \tag{5.26}$$

and V_M is the volume of a unit ball with dimension M. In contrast to previous work, this expression does not assume any particular input distribution, or linearization of the channel model and includes parameters of interest that can also capture the effect of the receiver power consumption due to the choice of the quantizer parameters in terms of the resolution and effect of the noise figure in the SNR. The previous result provides us in a first step with an expression of the capacity as N becomes large. The second step requires studying the relationship between spectral efficiency degradation and the quantizer resolution, which relates directly to its power consumption. Assuming a digital architecture where each receive antenna is equipped with a dedicated RF chain, future work involves formulating a meaningful optimization problem that can capture the balance with the spectral efficiency by defining a power consumption model of the receiver front-end in a manner similar to [49, 97, 126] and given the data and proposed FoMs in the literature [15]. It would be interesting to derive similar results for the noncoherent channel, therefore we begin by investigating this problem in the following section.

5.4 . Asymptotic Capacity of Noncoherent Channel

Without any a priori CSI, the likelihood function depends on the covariance matrix of the channel Gaussian density function. When $b = 1$,

the channel law of Eq. (5.9) is given as

$$\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{k=1}^{n} \boldsymbol{P}(\boldsymbol{y}_i|\boldsymbol{X}) \tag{5.27}$$

$$= \prod_{k=1}^{n} \int_{\mathcal{K}_{\boldsymbol{y}_k}} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{T/2}} \exp\left\{-\tfrac{1}{2}\boldsymbol{v}_k^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}_k\right\} \mathrm{d}\boldsymbol{v}_k, \tag{5.28}$$

where $\boldsymbol{\Sigma} = \boldsymbol{I}_T + \rho\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ and $\mathcal{K}_{\boldsymbol{y}_k}$ is the orthant probability region such that $\boldsymbol{y}_k \odot \boldsymbol{v_k} \geq 0$, where $\odot$ represents an element-wise product. Theorem 1 has been used in [130] to derive the capacity scaling by finding a carefully chosen one-to-one parameterization between $\boldsymbol{X}$ and $\boldsymbol{\theta}$ such the left-hand side of Eq. (5.14) becomes equality. We will consider upper bounds on the results therein by looking at the unquantized channel and following the same approach. Assuming infinite precision, the channel likelihood function is given by

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{X}), \tag{5.29}$$

where we have N i.i.d. realizations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ according to

$$p(\boldsymbol{y}|\boldsymbol{X}) = \frac{|\boldsymbol{\Sigma}_{\boldsymbol{X}}|^{-\frac{1}{2}}}{(2\pi)^{T/2}} \exp\left\{-\tfrac{1}{2}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{y}\right\}, \tag{5.30}$$

such that $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{I}_T + \rho\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. It can be seen that the channel law in Eq. (5.30) depends on the input through $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, i.e., it is parameterized by the following set

$$\mathcal{F} := \left\{\varphi(\boldsymbol{X}) := \boldsymbol{I}_T + \rho\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} : \boldsymbol{X} \in \mathcal{X}\right\}. \tag{5.31}$$

Consider the set of parameters $\boldsymbol{\theta} := \{\theta_1, \theta_2, \ldots, \theta_d\}$ where $d = \binom{T+1}{2}$ and define

$$\boldsymbol{\Theta} := \left\{\boldsymbol{\theta} \in \boldsymbol{\Gamma} \subset \mathbb{R}^d : \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) \succeq 0\right\} \tag{5.32}$$

as the set of positive semi-definite matrices parameterized by $\boldsymbol{\theta}$ where $\boldsymbol{\Gamma}$ is such that $\theta_{jj} \geq 0$ for $j = k$ and $-1 \leq \theta_{j,k} \leq 1$ for $j \neq k$ for $j, k \in [T]$. We also have the following functions

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{I}_T, \tag{5.33}$$

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) := [\theta_{\{j,k\}}], \quad j, k \in [T]. \tag{5.34}$$

Since $\varphi(\boldsymbol{X})$ is not injective into $\boldsymbol{\Theta}$, the mutual information can be written as

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= I(\varphi(\boldsymbol{X});\boldsymbol{Y}) \\ &\leq \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} I(\boldsymbol{\theta};\boldsymbol{Y}). \end{aligned} \tag{5.35}$$

The following is an equivalent of Lemma 1 in [130] to show the existence of a signaling scheme that achieves an equality above. Let

$$\mathcal{F}' := \{\boldsymbol{\Sigma}(\rho\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \tag{5.36}$$

It is evident that $\mathcal{F} \subset \mathcal{F}'$, this is because we can write

$$\boldsymbol{\Sigma}(\rho\boldsymbol{\theta}) = \boldsymbol{I}_T + \rho\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) \tag{5.37}$$

and

$$\varphi(\boldsymbol{X}) = \boldsymbol{I}_T + \rho\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} \tag{5.38}$$

where $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} \succeq 0$. Furthermore, let $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$ where $\boldsymbol{L}^{\mathsf{T}}$ has columns with magnitude less than or equal to 1, then see that

$$\boldsymbol{\Sigma}(\rho\boldsymbol{\theta}) = \boldsymbol{I}_T + \rho\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} = \varphi(\boldsymbol{L}^{\mathsf{T}}) \in \mathcal{F} \tag{5.39}$$

Similarly, define the function

$$\varrho : \boldsymbol{X} \longrightarrow \boldsymbol{\Theta}, \tag{5.40}$$

such that $\varrho_{\{j,k\}}(\boldsymbol{X}) := \boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_k$ for $\boldsymbol{X} \in \mathcal{X}$, and the canonical set $\mathcal{X}_0$ through the Cholesky decomposition function

$$\boldsymbol{X}_0(\boldsymbol{\theta}) = \mathrm{Chol}\Big(\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\Big)^{\mathsf{T}}, \tag{5.41}$$

then the two functions are the inverse of one another and we obtain a bijection between the two sets. We can now work directly with the following parameterized pdfs for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which is that of the multivariate Gaussian

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{5.42}$$

Note that the regularity conditions of Theorem 1 hold for this family and more generally for the exponential family of distributions as shown in [128]. Applying Barron and Clarke's result for this channel, we have

$$\max_{\boldsymbol{\theta}} I(\boldsymbol{\theta},\boldsymbol{Y}) = \frac{d}{2}\log\left(\frac{N}{2\pi e}\right) + \log\int_{\boldsymbol{\Theta}} |\boldsymbol{J}(\boldsymbol{\theta})|^{\frac{1}{2}}\,\mathrm{d}\boldsymbol{\theta} + o(1). \tag{5.43}$$

The Fisher information is given by

$$\boldsymbol{J}_T(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \ln(f_{\boldsymbol{\theta}}(\boldsymbol{y}))\, \nabla_{\boldsymbol{\theta}}^{\mathsf{T}} \ln(f_{\boldsymbol{\theta}}(\boldsymbol{y}))\right]. \tag{5.44}$$

In our case, for a zero-mean multivariate Gaussian distribution, the Fisher information can be obtained in closed-form as

$$[\boldsymbol{J}_T(\boldsymbol{\theta})]_{i,j} = \frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_j}\right\} \tag{5.45}$$

Proof: First, we define the log-likelihood function

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \ln(f_{\boldsymbol{\theta}}(\boldsymbol{y})) \\ &= -\frac{1}{2}\ln|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} + \text{cte}\end{aligned} \tag{5.46}$$

then we take the partial derivative with respect to θ_i

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i} = \frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\right\} + \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}, \tag{5.47}$$

then the partial derivative with respect to θ_j and applying the chain rule, we obtain

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} =& -\frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i^2}\theta_j + \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\partial \theta_j}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\right\} \\ &+ \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\partial \theta_j} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i^2}\theta_j\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}}{\partial \theta_j}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\boldsymbol{y},\end{aligned} \tag{5.48}$$

which simplifies to

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} =& \frac{1}{2}\mathsf{tr}\left\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_j}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\right\} \\ &- \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_j}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_j}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\boldsymbol{y},\end{aligned} \tag{5.49}$$

The Fisher matrix is obtained by then taking negative of the expectation

$$[\boldsymbol{J}_T(\boldsymbol{\theta})]_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]. \tag{5.50}$$

We will now look at particular cases when the coherence interval is fixed.

5.4.1 . Special Case: T = 1

When the coherence interval is equal to $T = 1$ and $d = 1$, we obtain

$$\sigma_\theta = 1 + \rho\theta. \tag{5.51}$$

Asymptotically in N, the optimal distribution on $\boldsymbol{\theta}$ is Jeffreys prior given by

$$p^*(\boldsymbol{\theta}) = \frac{1}{c}\sqrt{|\boldsymbol{J}(\boldsymbol{\theta})|} \tag{5.52}$$

where we have the Fisher information

$$J(\theta) = \frac{\rho^2}{2(1+\rho\theta)^2} \tag{5.53}$$

with

$$c = \int_0^1 \sqrt{\frac{\rho^2}{2(1+\rho\theta)^2}}\,\mathrm{d}\theta = \frac{1}{\sqrt{2}}\ln(1+\rho). \tag{5.54}$$

The asymptotic capacity grows as $N \to \infty$

$$C = \frac{1}{2}\log\left(\frac{N}{4\pi e}\right) + \log\ln(1+\rho), \tag{5.55}$$

and the optimal distribution on θ is

$$p^*(\theta) = \frac{\rho}{\ln(1+\rho)} \cdot \frac{1}{1+\rho\theta}. \tag{5.56}$$

The optimal signaling is such that

$$x(\theta^*) = \sqrt{\theta^*}, \tag{5.57}$$

this implies that as N goes to infinity, one transmit antenna is sufficient to achieve capacity under a peak power constraint.

5.4.2 . Special Case: T = 2

As T increases, the dimensionality of the problem also grows rendering the computation of certain quantities such as the Fisher matrix along with the integral of its determinant more challenging. We consider now the special case when $T = 2$ and $d = \binom{3}{2} = 3$. We have

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{bmatrix} 1+\rho\theta_1 & \rho\theta_2 \\ \rho\theta_2 & 1+\rho\theta_3 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} = \frac{1}{|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|}\begin{bmatrix} 1+\rho\theta_3 & -\rho\theta_2 \\ -\rho\theta_2 & 1+\rho\theta_1 \end{bmatrix} \tag{5.58}$$

such that $|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| = (1+\rho\theta_1)(1+\rho\theta_3) - \rho^2\theta_2^2 > 0$. From Eq. (5.45), we obtain

$$\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_1} = \rho \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_2} = \rho \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_3} = \rho \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.59}$$

The Fisher information is given by the following form

$$\begin{aligned} &\boldsymbol{J}_2(\boldsymbol{\theta}) \\ &= \frac{\rho^2}{2|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^2} \begin{bmatrix} (1+\rho\theta_3)^2 & -2\rho\theta_2(1+\rho\theta_3) & \rho^2\theta_2^2 \\ -2\rho\theta_2(1+\rho\theta_3) & 2[\rho^2\theta_2^2 + (1+\rho\theta_1)(1+\rho\theta_3)] & -2\rho\theta_2(1+\rho\theta_1) \\ \rho^2\theta_2^2 & -2\rho\theta_2(1+\rho\theta_1) & (1+\rho\theta_1)^2 \end{bmatrix} \end{aligned} \tag{5.60}$$

Where the determinant of the Fisher matrix in Jeffreys' prior is given by

$$|\boldsymbol{J}_2(\boldsymbol{\theta})| = \frac{\rho^6}{4|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^3} = \frac{\rho^6}{4\left[(1+\rho\theta_1)(1+\rho\theta_3) - \rho^2\theta_2^2\right]^3} \tag{5.61}$$

and the normalizing constant is such that $c = \int_{\boldsymbol{\Theta}} \sqrt{|\boldsymbol{J}_2(\boldsymbol{\theta})|}\, \mathrm{d}\boldsymbol{\theta}$. Knowing that

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} : \theta_1\theta_3 - \theta_2^2 \geq 0 \text{ and } 0 \leq \theta_1, \theta_3 \leq 1 \right\}, \tag{5.62}$$

the normalizing constant is equal to

$$\begin{aligned} c &= \int_0^\rho \int_0^\rho \sqrt{\frac{\theta_1\theta_3}{1+\theta_1+\theta_3}} \cdot \frac{1}{(1+\theta_1)(1+\theta_3)}\, \mathrm{d}\theta_1\, \mathrm{d}\theta_3 \\ &= \int_0^\rho \int_0^\rho f(\theta_1,\theta_3) g(\theta_1,\theta_3)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_3 \end{aligned} \tag{5.63}$$

To the best of our knowledge, this integrand does not have any closed-form expression or an anti-derivative, therefore we will seek upper and lower bounds. There are different ways to bound this integral for a fixed value of ρ, the first is a direct application of the extreme value theorem by first noting that for a fixed value of θ_3, the right-hand side function denoted now as $g_{\theta_3}(\theta_1)$ is monotonically decreasing with θ_1 over $[0, \rho]$ and can be upper-bounded as

$$\begin{aligned} g_{\theta_3}(\rho) \leq g_{\theta_3}(\theta_1) &\leq g_{\theta_3}(0) \\ \frac{1}{(1+\rho)(1+\theta_3)} \leq g_{\theta_3}(\theta_1) &\leq \frac{1}{(1+\theta_3)} \end{aligned} \tag{5.64}$$

therefore integrating further with respect to θ_3 we obtain the bounds on c as

$$\kappa_1^{\mathrm{LB}}(\rho) := \frac{1}{1+\rho}\kappa(\rho) \leq c \leq \kappa(\rho) := \kappa_1^{\mathrm{UB}}(\rho), \tag{5.65}$$

where we define the function

$$\begin{aligned}\kappa(\rho) = 2\Bigg[&(2+\rho)\sqrt{\frac{\rho}{1+\rho}}\ln\Big(\sqrt{\rho}+\sqrt{\rho+1}\Big) \\ &- \ln^2(\sqrt{\rho}+\sqrt{\rho+1}) - \rho\Bigg].\end{aligned} \tag{5.66}$$

The second approach relies on an alternative parameterization of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ using the Cholesky decomposition and provides us with tighter bounds. Let the covariance matrix be parameterized such that $\boldsymbol{\Sigma}(\boldsymbol{\ell}) = \boldsymbol{I}_T + \rho \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$. Following the same approach as before, the space becomes

$$\boldsymbol{\Omega} := \big\{\boldsymbol{\ell} \in \mathbb{R}^d : \boldsymbol{\Sigma}(\boldsymbol{\ell}) \succeq 0\big\} \tag{5.67}$$

where $\boldsymbol{l} = \{\ell_1, \ldots, \ell_d\}$ represents the elements of $\boldsymbol{L}$ stacked column-wise. Note that since the Fisher information matrix is covariant under re-parameterization [132], we can write

$$\boldsymbol{J}(\boldsymbol{\ell}) = \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right]^{\mathsf{T}} \boldsymbol{J}(\boldsymbol{\theta}(\boldsymbol{\ell}))\left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right] \tag{5.68}$$

The square root of the determinant of the Fisher matrix is given by

$$\begin{aligned}|\boldsymbol{J}(\boldsymbol{\ell})|^{\frac{1}{2}} &= |\boldsymbol{J}(\boldsymbol{\theta}(\boldsymbol{\ell}))|^{\frac{1}{2}} \cdot \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right| \\ &= \frac{\rho^3}{2|\boldsymbol{\Sigma}(\boldsymbol{\theta}(\boldsymbol{\ell}))|^{\frac{3}{2}}} \cdot \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right| \\ &= \frac{2\rho^3}{(1+\rho^2\ell_1^2\ell_3^2 + \rho\ell_2^2 + \rho\ell_3^2 + \rho\ell_1^2)^{\frac{3}{2}}} \cdot \ell_1^2\ell_3.\end{aligned} \tag{5.69}$$

We can use the following identity to upper and lower bound the determinant

$$\begin{aligned}|\boldsymbol{\Sigma}(\boldsymbol{\theta}(\boldsymbol{\ell}))| &= |\boldsymbol{I}_T + \rho^2 \boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}| \\ &= 1 + \rho^2|\boldsymbol{L}\boldsymbol{L}^T| + \rho\mathsf{tr}\left\{\boldsymbol{I}_T\right\}\mathsf{tr}\left\{\boldsymbol{L}\boldsymbol{L}^T\right\} - \rho\mathsf{tr}\left\{\boldsymbol{I}_T\boldsymbol{L}\boldsymbol{L}^T\right\} \\ &= 1 + \rho^2|\boldsymbol{L}\boldsymbol{L}^T| + \rho\mathsf{tr}\left\{\boldsymbol{L}\boldsymbol{L}^T\right\} \\ &= 1 + \rho^2\ell_1^2\ell_3^2 + \rho(\ell_1^2 + \ell_2^2 + \ell_3^2) \\ &\geq 1 + \rho^2\ell_1^2\ell_3^2 + \rho(\ell_1^2 + \ell_3^2)\end{aligned} \tag{5.70}$$

which we can also upper bound as

$$|\boldsymbol{\Sigma}(\boldsymbol{\theta}(\boldsymbol{\ell}))| \leq 1 + \rho^2\ell_1^2\ell_3^2 + \rho^2\ell_1^2\ell_2^2 + \rho(\ell_1^2 + \ell_2^2 + \ell_3^2) \tag{5.71}$$

These two choices of the bounds along with converting to spherical coordinates allow us to integrate over the region where we can retrieve tighter upper and lower bounds on c as explicit functions of ρ

$$\kappa_2^{\mathrm{LB}}(\rho) \leq c \leq \kappa_2^{\mathrm{UB}}(\rho) \tag{5.72}$$

where

$$\begin{aligned}
\kappa_2^{\mathrm{LB}}(\rho) &= \frac{4}{1+\rho} \cdot (\rho - \sqrt{\rho(1+\rho)}\sinh^{-1}(\sqrt{\rho})) \\
&\quad + 8\left(\sinh^{-1}(\sqrt{\rho}) - \sqrt{\frac{\rho}{1+\rho}}\right) \cdot \tanh^{-1}\left(\frac{\sqrt{1+\rho}-1}{\sqrt{\rho}}\right) \\
\kappa_2^{\mathrm{UB}}(\rho) &= \frac{4}{\sqrt{1+\rho}} \cdot \left(\sqrt{1+\rho}\sinh^{-1}(\sqrt{\rho}) - \sqrt{\rho}\right) \cdot (\sqrt{\rho} - \arctan(\sqrt{\rho}))
\end{aligned} \tag{5.73}$$

We compare these bounds in Fig. 5.3. We see that the second bounding technique gives us tighter scalings of the capacity compared with the numerical integration for increasing values of ρ which are shown to be tightest for small peak power constraints. For a fixed $\rho = 5$ dB and increasing values of N, we plot in Fig. 5.4 the asymptotic capacity of the unquantized channel for this particular case compared to that of the one-bit quantized channel result obtained in [130]. As expected from the data processing inequality, for a fixed value of N and ρ, the capacity of the unquantized channel is an upper bound to that of the one-bit case. For a fixed ρ, the comparison can provide us with a rough estimate of the number of receive antennas needed to approach the spectral efficiency of the unquantized channel.

5.4.3 . Special Case: T = 3

When $T = 3$ the dimension of the input space now involves 6 parameters. We can proceed similarly to the case when $T = 2$ and obtain the determinant of the Fisher information matrix as

$$|\boldsymbol{J}_3(\boldsymbol{\theta})|^{\frac{1}{2}} = \frac{\rho^6}{2\sqrt{2}|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^2} \tag{5.74}$$

Using the Cholesky parameterization we have $\boldsymbol{\Sigma}(\boldsymbol{\ell}) = \boldsymbol{I}_T + \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\ell})$ and the square root of the Fisher matrix determinant is then given by

$$\begin{aligned}
|\boldsymbol{J}(\boldsymbol{\ell})|^{\frac{1}{2}} &= |\boldsymbol{J}(\boldsymbol{\theta}(\boldsymbol{\ell}))|^{\frac{1}{2}} . \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right| \\
&= \frac{\rho^6}{2\sqrt{2}|\boldsymbol{\Sigma}(\boldsymbol{\ell})|^2} \cdot 8\ell_1^3\ell_3^2\ell_6,
\end{aligned} \tag{5.75}$$

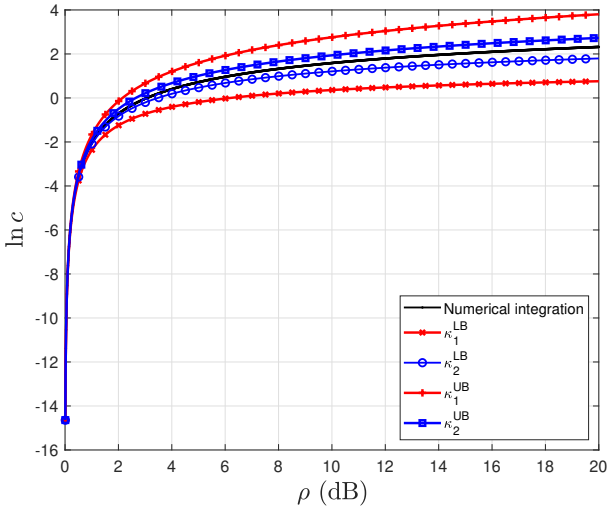

Figure 5.3: Normalizing constant as a function of ρ in linear scale along with the updated upper and lower bounds for $T = 2$.

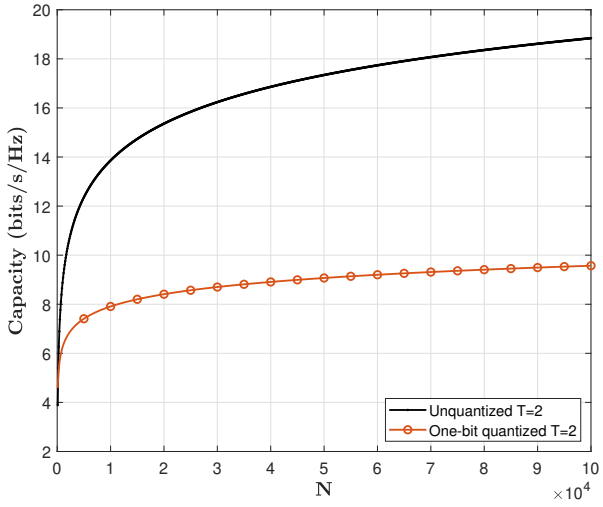

Figure 5.4: Comparison between the asymptotic capacity of the unquantized and one-bit quantized channel for $T = 2$ and $\rho = 5$ dB.

Let $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \lambda_3\}$ denote the eigenvalues of $\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$, then using the following identity

$$\exp\left\{\mathsf{tr}\left\{\boldsymbol{I} - \boldsymbol{A}^{-1}\right\}\right\} \leq |\boldsymbol{A}| \leq \exp\{\mathsf{tr}\left\{\boldsymbol{A} - \boldsymbol{I}\right\}\}, \tag{5.76}$$

we obtain an upper bound on the Fisher matrix determinant as

$$|\boldsymbol{\Sigma}(\boldsymbol{\ell})| \leq \exp\left\{\rho \sum_{i=1}^{6} \ell_i^2\right\} \tag{5.77}$$

The integral lower bound is thus

$$c \geq 2\sqrt{2}\rho^6 \int_{\boldsymbol{\Omega}} \ell_1^3 \ell_3^2 \ell_6 \exp\left\{-2\rho \sum_{i=1}^{6} \ell_i^2\right\} \mathrm{d}\boldsymbol{\ell} \tag{5.78}$$

With the spherical parameterization, we manage to obtain the following lower bound on this integral

$$c \geq \frac{\pi^2}{256\sqrt{2}} \cdot \frac{(e^{2\rho} - 2\rho - 1)^3}{e^{6\rho}}. \tag{5.79}$$

For general $T > 3$, Jeffreys' prior obtained as the limit of the conjugate prior of the Inverse-Wishart distribution with $T - 1$ degrees of freedom can be written as [133]

$$|\boldsymbol{J}_T(\boldsymbol{\theta})|^{\frac{1}{2}} \propto \frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{T+1}{2}}}. \tag{5.80}$$

Following the same Cholesky parameterization as before, the integral of the Fisher matrix determinant may be written as

$$\begin{aligned} \int_{\boldsymbol{\Theta}} |\boldsymbol{J}(\boldsymbol{\theta})|^{\frac{1}{2}} \, \mathrm{d}\boldsymbol{\theta} &= \int_{\boldsymbol{\Omega}} |\boldsymbol{J}_T(\boldsymbol{\theta}(\boldsymbol{\ell}))|^{\frac{1}{2}} \cdot \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\ell}}\right| \mathrm{d}\boldsymbol{\ell} \\ &= 2^T \int_{\boldsymbol{\Omega}} \prod_{i=1}^{T} (1 + \rho \ell_{ii}^2)^{-(T+1)} (\ell_{ii})^{T-i+1} \, \mathrm{d}\boldsymbol{\ell} \end{aligned} \tag{5.81}$$

For this particular case and $T > 3$ ongoing work is underway to obtain an upper bound on the integral to get a better understanding of the scaling of the capacity when N is large. The main challenge towards this goal is in obtaining meaningful bounds on the determinant of the Fisher matrix for any T that can result in tight scalings of the capacity. The difficulty lies in integrating over the space $\boldsymbol{\Theta}$ or $\boldsymbol{\Omega}$ in general, and although reparameterizing the space can sometimes provide us with simpler expressions to obtain closed-form expressions of these bounds, it is not evident how to do that for general T. An alternative approach would be to consider the capacity scaling thereafter in asymptotic regimes of the SNR. Future objectives involve studying the scaling once a multi-bit quantizer is introduced.

5.5 . Summary

In this chapter, we presented preliminary studies for ongoing and future work that aims to extend the work to the multi-bit low-resolution ADCs case and to analyze the trade-off between power consumption and spectral efficiency degradation at the receiver front end. We have seen that the SD algorithm in Chapter 3 can be extended to the multi-bit case, and exhibits near-optimal performance in terms of VERs with respect to the ML lower bounds. The approach captures as the resolution grows to infinity the conventional SD case. Employing known information-theoretic asymptotics used in Bayesian statistics, we can characterize the capacity scaling in the asymptotic regime where the number of receive antennas is large. For the coherent multi-bit quantized MIMO channel, a straightforward extension of the one-bit case as presented in [131] is possible and we can obtain an expression of the asymptotic capacity. A future step is identifying meaningful yet simple physical models that capture the power consumption of RF components that can be analyzed jointly with the capacity expression. For the noncoherent channel, we look at the unquantized asymptotic capacity as an upper bound when the coherence interval takes particular values. Integrating over the parameter space does not result in closed-form expressions in general, therefore we identify upper and lower bounds on the scaling.

Chapter 6

Conclusion and Future Perspectives

As we move towards newer generations of communication systems, the difficulties imposed on the RF circuitry level become more prevalent. Practical considerations have to be taken on the overall energy consumed at the receiver front-end, especially as we strive towards employing higher bandwidths and more complex digital architectures. Beyond 5G and 6G system design goals aim to ensure high transmission rates while keeping power consumption due to individual RF components at a minimum. We have seen that, for ADCs, this consumption scales exponentially in resolution and linearly in their sampling frequency. This incentivizes the use of low-resolution alternatives, at the cost of more complex signal processing techniques. We have considered in this thesis the problems of optimal data detection under two extreme cases of CSI availability in a quantized narrowband block-fading Rayleigh MIMO channel. In Chapter 3, we addressed the case of optimal data detection under a perfect CSI case with one-bit ADCs. The computational complexity under a discrete constellation constraint grows exponentially in the number of transmitters and constellation size. For the classical case with infinite resolution, SD is a technique that aims to reduce the search space under the square-loss metric. We then showed that SD can be applied under coarse quantization, and in contrast to SD-like heuristics [71, 81] in the literature, the approach is more in line with the conventional method that does indeed reduce the search space while taking advantage of the available CSI. Moreover, near-optimality is assessed with respect to a numerical lower bound on the ML metric in terms of VERs for different constellation sizes and assumptions on channel correlations. In addition, the approach is extended to the multi-bit quantized case and captures naturally the infinite resolution case. In Chapter 4, we looked at the scenario where only statistical CSI is available at the receiver. Assuming transmission is conducted independently between pilot and data frames, we first present a two-stage channel estimation and data detection approach based on probit regression. This approach is adopted given its direct similarity with the binary classification problem. In addition, it allows us to investigate the instability of the probit model under imperfect CSI. First, it is demonstrated that the instability is essentially related to the BLMMSE estimator and not the model. Second, the GMI is used to compute achievable rates with the mismatched metrics obtained from

the probit model and BLMMSE estimators. Generally, it's shown that the performance approaches that of the matched case as the training length increases, with degradation when 16-QAM signaling is used. In the second part, we look at a more general approach where the pilot and data frames are processed jointly at the receiver. For simplicity, we assumed a real channel model. After formulating the optimal metric in this case, we identify two main difficulties that involve the evaluation of multivariate Gaussian orthant probabilities which do not have closed-form expressions, and the combinatorial optimization as in Chapter 3. We then present an approximation using the Laplace method and extend the SD algorithm based on the mismatched ML as a surrogate metric. It's shown through numerical simulations that the proposed approximation achieves an improved near-optimal performance for 4-PAM signaling schemes over one-shot estimators such as the mismatched ML and the Bussgang techniques in terms of VERs. In Chapter 5, we turned our attention to power consumption and spectral efficiency trade-off assessment by first looking at the capacity problem of the quantized channel. We present preliminary work motivated by Clarke and Barron's results to obtain an asymptotic expression of the capacity. The objective is to use information-theoretic results coupled with physical constraints on the system to derive design guidelines for future wireless communication systems. For the coherent channel, we present the straightforward extension of the results in [131] for the multi-bit quantizer and highlight the dependence of the capacity expression on different parameters that can be useful for the energy consumption design guidelines. A future step is to identify a meaningful model of the receiver front-end power consumption. Aspiring for a similar goal with the noncoherent channel, we begin by investigating the asymptotic capacity of the unquantized channel as an upper bound for the quantized case for particular values of the coherence interval. When $T = 1$, an exact expression of the capacity is obtainable. When $T = 2$, the Fisher matrix determinant does not have a closed-form and we manage to obtain tight upper and lower bounds with respect to the numerical integration. For $T = 3$, we present a lower bound on the capacity scaling. The main challenge for general T is the difficulty in integrating over the induced parameter space. This motivates future work in investigating these bounds for asymptotic regimes of other parameters of the capacity such as the SNR. Moreover, studying the scaling of the capacity for the multi-bit quantized case is also an interesting direction.

6.1 . Future Perspectives

6.1.1 . Improvements and Extensions

The research work on communication systems with low-resolution ADCs is rich and each area can offer a plethora of opportunities to investigate. In our work, we have assumed Nyquist sampling at the receiver. It's known that oversampling in the extreme case of one-bit ADCs can improve the system performance [83,134–136]. Moreover, we have not made any design assumptions about the transmitter, which opens a line of different research perspectives. For example, coupled with the oversampling approach, we can investigate optimal transmission strategies and signaling schemes [84,137]. Under a more general assumption of low-resolution digital-to-analog converters (DAC) and considering the downlink case, the design of precoding schemes is another direction [37,85,138]. Assuming frequency-selective channels, employing OFDM with low-resolution ADCs is not straightforward, since after quantization applying the DFT does not result in the channel singular values that we can directly equalize [72, 73, 139]. In addition to the current contributions, several extensions can be explored. For example, analyzing analytically the SD complexity in the average sense in a manner similar to [110] where the channel statistics and the noise distribution are considered. It would be also useful to assess the performance of the proposed metrics in more practical scenarios where we have coded transmission. For the data detection problem under statistical CSI, future work involves extending the LA method to the complex channel and exploring the effect of channel spatial correlations. We also note that the Laplace method is one technique that allows us to approximate the optimal metric in Eq. (4.19), it would be interesting to consider more general models that can describe this metric more accurately.

6.1.2 . Other Sources of Nonlinearities

The current work can be extended to encompass other sources of nonlinearities. For example, we have not assumed any impairments in the quantization operation such as integral and differential nonlinearities [19]. Furthermore, aperture uncertainty effects in the sample-and-hold circuitry could also impose additional difficulties, which are not assumed in this work [7]. Another source of signal distortion can be due to power amplification (PA) which can be present both at the transmitting and receiving ends. Inexpensive and power-efficient amplifiers are inherently nonlinear which can cause distortions that affect the transmission rates. A pioneering work that deals with PA distortion for satellite communications is due to Benedetto et al. [31] for their design of a nonlinear equalizer of the Volterra channel that estimates the coefficients based on

a minimum mean square error (MMSE) criterion. We note that these techniques have been presented in the context of satellite communication channels where the power is very limited and high efficiency is necessary. Regarding the performance evaluation of MIMO systems with nonlinear amplifier distortion, Fozooni et al. [43] analyze the ergodic achievable rate where the Bussgang decomposition is exploited to decompose the output of the power amplifier such that the distortion noise is additive. Moreover, the authors in [42] present single-carrier optimal transmit beamforming techniques in the presence of memoryless nonlinear PAs in flat Rayleigh fading channels, and compute bounds on the average symbol error probability (SEP). As for mMIMO systems where dedicating RF chains for each transmit antenna can be impractical, the authors in [140] shed light on evaluating the performance degradation of hybrid OFDM digital-analog mMIMO systems. Much less work has focused on the joint impairments due to both PA and ADC distortions which can be an interesting direction for receiver design and spectral-power efficiency trade-off analysis.

Appendices

Appendix A

In this section, we provide a detailed derivation of the Bussgang gain in Eq. (2.10) obtained from linearizing the multi-bit quantized MIMO channel using Price's theorem which states that, as a particular case, given two jointly Gaussian x and y as inputs to two deterministic functions $f_1(\cdot)$ and $f_2(\cdot)$, the covariance component at the outputs is given by

$$\frac{\partial \mathbb{E}\left[f_1(x) f_2(y)\right]}{\partial \sigma_{xy}} = \mathbb{E}\left[\frac{\partial f_1(x)}{\partial x} \cdot \frac{\partial f_2(y)}{\partial y}\right]. \tag{6.1}$$

We can now proceed to show the result in Eq. (2.10). The derivation assumes that $\tilde{\boldsymbol{r}}$ follows a CSCG distribution. First, we express the quantizer function as

$$\mathsf{Q}_b(x) = \sum_{l=1}^{2^b} \nu_l \mathbb{1}_{\{x \in \mathcal{I}_l\}} \tag{6.2}$$

To obtain the Bussgang gain $\boldsymbol{W}_b$ we will need to compute the cross-correlation matrix

$$\begin{aligned} \boldsymbol{C}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{r}}} = \mathbb{E}\left[\tilde{\boldsymbol{y}}\tilde{\boldsymbol{r}}\right] &= \begin{bmatrix} \mathbb{E}[\tilde{r}_1 \tilde{y}_1^*] & \dots & \mathbb{E}[\tilde{r}_1 \tilde{y}_N^*] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\tilde{r}_N \tilde{y}_1^*] & \dots & \mathbb{E}[\tilde{r}_N \tilde{y}_N^*] \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}[\tilde{\mathsf{Q}}_b(\tilde{y}_1) \tilde{y}_1^*] & \dots & \mathbb{E}[\tilde{\mathsf{Q}}_b(\tilde{y}_1) \tilde{y}_N^*] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\tilde{\mathsf{Q}}_b(\tilde{y}_N) \tilde{y}_1^*] & \dots & \mathbb{E}[\tilde{\mathsf{Q}}_b(\tilde{y}_N) \tilde{y}_N^*] \end{bmatrix} \end{aligned} \tag{6.3}$$

Noting that

$$\begin{aligned} \mathbb{E}\left[\tilde{\mathsf{Q}}_b(\tilde{y}_i) \tilde{y}_j^*\right] =& \mathbb{E}\left[\left(\mathsf{Q}_b(\tilde{y}_i^R) \tilde{y}_j^R + \mathsf{Q}_b(\tilde{y}_i^I) \tilde{y}_j^I\right)\right] \\ &+ j\mathbb{E}\left[\left(\mathsf{Q}_b(\tilde{y}_i^I) \tilde{y}_j^R - \mathsf{Q}_b(\tilde{y}_i^R) \tilde{y}_j^I\right)\right] \end{aligned} \tag{6.4}$$

We apply Eq. (6.1) where now $f_1(\cdot)$ is the quantization operation Q_b whose derivative is

$$\frac{\partial \mathsf{Q}_b(x)}{\partial x} = \sum_{l=1}^{2^b} \nu_l \left[\delta(x - I_{l-1}) - \delta(x - I_l)\right], \tag{6.5}$$

such that $\delta(x)$ represents the Dirac delta function. In this setting, $f_2(\cdot)$ is simply the identity function. Using the circular symmetry property of the input, the real and imaginary components of the input vector $\tilde{\boldsymbol{r}}$ have the same auto-covariance, while their cross-covariance matrix is skew-symmetric. Therefore let $\tilde{\sigma}_{i,j} = [\boldsymbol{C}_{\tilde{\boldsymbol{r}}}]_{i,j}$ we can write Eq. (6.4) after applying the Dirac's function sifting property,

$$\mathbb{E}\left[\tilde{\mathsf{Q}}_b(\tilde{y}_i)\tilde{y}_j^*\right] = \tilde{\sigma}_i^{-\frac{1}{2}} \sum_{l=1}^{2^b} \frac{\nu_l}{\sqrt{\pi}} \left(e^{-\frac{I_{l-1}^2}{\tilde{\sigma}_i^2}} - e^{-\frac{I_l^2}{\tilde{\sigma}_i^2}} \right) \cdot (\tilde{\sigma}_{i,j}), \tag{6.6}$$

expanding in matrix form and multiplying by the right with $\boldsymbol{C}_{\tilde{\boldsymbol{r}}}$, we obtain Eq. (2.10).

Appendix B

The gradient is $L-$Lipschitz by the following sufficient condition on the Hessian's spectral norm

$$||\nabla^2_{\ell_b(\boldsymbol{x})}||_2 = \lambda_{\max}(\nabla^2_{\ell_b(\boldsymbol{x})}) \leq L, \tag{6.7}$$

for some $L > 0$ for all $\boldsymbol{x} \in \mathbb{R}^{2M}$ and where $\lambda_{\max}(\cdot)$ denotes the largest singular value of the corresponding matrix. This last condition can be readily established by first re-writing (5.2) into a matrix form

$$\nabla^2_{\ell_b(\boldsymbol{x})} = \frac{1}{\sigma^2} \boldsymbol{H}^{\mathsf{T}} \boldsymbol{\Upsilon}_{b,\delta} \boldsymbol{H}, \tag{6.8}$$

where

$$\boldsymbol{\Upsilon}_{b,\delta} = \mathsf{diag}(\eta_{b,\delta}(y_1), \ldots, \eta_{b,\delta}(y_n)). \tag{6.9}$$

We obtain by the sub-multiplicative property of matrix norms

$$\begin{aligned} ||\nabla^2_{\ell_b(\boldsymbol{x})}||_2 &\leq \frac{1}{\sigma^2} ||\boldsymbol{H}^{\mathsf{T}}||_2 \cdot ||\boldsymbol{\Upsilon}_{b,\delta}||_2 \cdot ||\boldsymbol{H}||_2 \\ &\leq \frac{1}{\sigma^2} \lambda^2_{\max}(\boldsymbol{H}) \cdot \lambda_{\max}(\boldsymbol{\Upsilon}_{b,\delta}) \\ &\leq \frac{\lambda^2_{\max}(\boldsymbol{H})}{\sigma^2}. \end{aligned} \tag{6.10}$$

The last inequality follows knowing that (5.4) is upper bounded by 1. Therefore we can choose L in (6.7) as

$$\begin{aligned}\lambda_{\max}(\nabla^2_{\ell_b(\boldsymbol{x})}) \leq L &= \frac{1}{\sigma^2}\max_{i=1,\dots,2N}\lambda_i^2, \\ &= \frac{1}{\sigma^2}\lambda^2_{\max}(\boldsymbol{H}),\end{aligned} \tag{6.11}$$

where the λ_i are the singular values of $\boldsymbol{H}$.

Appendix C

We present in this section how the spatial correlation model is generated for users channel vectors at the base station and as described in [111]. This model assumes a massive MIMO single-cell scenario in a frequency flat fading channel. The channel vectors for each user are given by

$$\boldsymbol{c}_m = \int_\Theta \boldsymbol{v}(\theta)c_m(\theta)\,\mathrm{d}\theta, \tag{6.12}$$

where $c_m(\theta)$ and $\boldsymbol{v}(\theta)$ represent, respectively, the channel gain function and base station response vector as a function of the incidence angle θ, which we assume to lie within a fixed range in $\Theta = [-\frac{\pi}{2}, \frac{\pi}{2}]$. The correlation matrix for the channel $\boldsymbol{c}_m$ of transmit antenna m is given as

$$\begin{aligned}\boldsymbol{C}_m &= \mathbb{E}\left[\boldsymbol{c}_m\boldsymbol{c}_m^{\mathsf{H}}\right] \\ &= \int_\Theta \boldsymbol{v}(\theta)\boldsymbol{v}(\theta)^{\mathsf{H}}S_m(\theta)\,\mathrm{d}\theta \\ &\approx \boldsymbol{V}\boldsymbol{R}_m\boldsymbol{V}^{\mathsf{H}},\end{aligned} \tag{6.13}$$

with $S_m(\theta)$ representing the power angular spread density. The last expression represents the approximation for a mMIMO system where $\boldsymbol{V}$ is a scaled DFT matrix with components

$$V_{i,j} = \frac{1}{\sqrt{N}}\exp\left\{-j\frac{2\pi}{M}(i-1)(j-1-N/2)\right\}, \quad 1 \leq i,j \leq N \tag{6.14}$$

The matrix $\boldsymbol{R}_m$ is diagonal with elements equal to

$$R_n^m = NS_m(f(\alpha_{n-1}))(f(\alpha_n) - f(\alpha_{n-1})), \quad n = 1,2,\dots,N. \tag{6.15}$$

Where $\alpha_n = n/N$ and $\theta = f(\alpha) = \arcsin(2\alpha - 1)$. Given each transmitter average AoA θ_m and the angular spread ς_m, the PAS density assumed is the Laplacian [141]

$$S_m(\theta) = \frac{1}{\sqrt{2}\varsigma_m(1-\exp\{-\sqrt{2}\pi\varsigma_m\})}\cdot\exp\left\{-\sqrt{2}\frac{|\theta-\theta_m|}{\varsigma_m}\right\}. \tag{6.16}$$

The mean AoAs are drawn uniformly from the interval $[-\frac{\pi}{3}, \frac{\pi}{3}]$ and are used in the computation of each channel vector $\boldsymbol{c}_m$ correlation matrix.

Appendix D

We first conduct a Taylor expansion for each channel vector in (4.26)

$$\begin{aligned}\mathcal{L}(\bar{\boldsymbol{h}}_n) \approx \mathcal{L}(\hat{\boldsymbol{h}}_n) + (\bar{\boldsymbol{h}}_n - \hat{\boldsymbol{h}}_n)^T \nabla_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} \\ + \frac{1}{2}(\bar{\boldsymbol{h}}_n - \hat{\boldsymbol{h}}_n)^T \nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} (\bar{\boldsymbol{h}}_n - \hat{\boldsymbol{h}}_n).\end{aligned} \tag{6.17}$$

Exponentiating this result and expanding the terms in the parenthesis we obtain

$$\begin{aligned}f(\bar{\boldsymbol{h}}_n) &= \exp\left\{\mathcal{L}(\bar{\boldsymbol{h}}_n)\right\} \\ &= \exp\left\{k(\hat{\boldsymbol{h}}_n) - \boldsymbol{b}_n^T \bar{\boldsymbol{h}}_n + \tfrac{1}{2}\bar{\boldsymbol{h}}_n^T \nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} \bar{\boldsymbol{h}}_n\right\},\end{aligned} \tag{6.18}$$

where we defined

$$\boldsymbol{b}_n = \nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} \hat{\boldsymbol{h}}_n - \nabla_{\mathcal{L}(\hat{\boldsymbol{h}}_n)}, \tag{6.19}$$

and $\exp\left\{\tilde{k}(\hat{\boldsymbol{h}}_n)\right\}$ is only a constant dependent on $\hat{\boldsymbol{h}}_n$ and $\boldsymbol{b}_n$ that eventually cancels out. Completing the square in (6.18) we get

$$f(\bar{\boldsymbol{h}}_n) \propto \exp\left\{\tfrac{1}{2}(\bar{\boldsymbol{h}}_n - \boldsymbol{\mu}_n)^T \nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} (\bar{\boldsymbol{h}}_n - \boldsymbol{\mu}_n)\right\}. \tag{6.20}$$

Identifying with the Gaussian (4.28), we obtain the corresponding parameters in (4.29)

$$\begin{cases}\boldsymbol{\mu}_n = \hat{\boldsymbol{h}}_n - (\nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)})^{-1} \nabla_{\mathcal{L}(\hat{\boldsymbol{h}}_n)} \\ \boldsymbol{\Sigma}_n = -(\nabla^2_{\mathcal{L}(\hat{\boldsymbol{h}}_n)})^{-1}\end{cases} \tag{6.21}$$

The final step is to normalize by integrating, which can be easily done for a multivariate Gaussian distribution

$$\begin{aligned}a_n &= \int f(\bar{\boldsymbol{h}}_n)\, \mathrm{d}\bar{\boldsymbol{h}}_n \\ &= e^{\tilde{k}(\hat{\boldsymbol{h}}_n)} \sqrt{(2\pi)^M |\boldsymbol{\Sigma}_n|} \int \phi(\boldsymbol{h}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\, \mathrm{d}\boldsymbol{h}_n \\ &= e^{\tilde{k}(\hat{\boldsymbol{h}}_n)} \sqrt{(2\pi)^M |\boldsymbol{\Sigma}_n|}.\end{aligned} \tag{6.22}$$

Finally, we obtain the result in (4.28)

$$Q(\bar{\boldsymbol{H}} | \boldsymbol{Y}_p, \boldsymbol{X}_p) = \prod_{n=1}^{N} \frac{1}{a_n} f(\bar{\boldsymbol{h}}_n). \tag{6.23}$$

Appendix E

The derivation of (4.30) relies on straightforward applications of Gaussian properties which we outline in this part. Since the channels are independent, we can re-write the expectation for each channel

$$\begin{aligned}\mathbb{E}_Q\left[\boldsymbol{P}\big(\boldsymbol{y}_d|\boldsymbol{x}_d,\bar{\boldsymbol{H}}\big)\right] &= \prod_{n=1}^{N}\mathbb{E}_Q\left[\boldsymbol{P}\big(\boldsymbol{y}_{d,n}|\boldsymbol{x}_d,\bar{\boldsymbol{h}}_n\big)\right] \\ &= \prod_{n=1}^{N}\int\left[\Phi\left(\frac{y_{d,n}\bar{\boldsymbol{h}}_n^T\bar{\boldsymbol{x}}_d}{\sigma_d}\right)\right]\times\phi_M(\bar{\boldsymbol{h}}_n;\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)\,\mathrm{d}\bar{\boldsymbol{h}}_n \\ &= \prod_{n=1}^{N} I_n(\boldsymbol{x}_d), \end{aligned} \tag{6.24}$$

where the expectation over Q in this context should be understood as taken over each channel $\bar{\boldsymbol{h}}_n$ due to independence. We introduce the latent variable $v_n = \bar{\boldsymbol{h}}_n^T\boldsymbol{x}_d + z_n$ and drop the n subscript for brevity in what follows. We can re-write

$$\begin{aligned} I(\boldsymbol{x}_d) &= \int_{\mathbb{R}}\mathbb{E}_Q\left[\boldsymbol{P}\big(y_d,v|\boldsymbol{x}_d,\bar{\boldsymbol{h}}_n\big)\right]\mathrm{d}v \\ &= \int_{\mathbb{R}}\boldsymbol{P}(y_d|v)\mathbb{E}_Q\left[\boldsymbol{P}\big(v|\boldsymbol{x}_d,\bar{\boldsymbol{h}}_n\big)\right]\mathrm{d}v \end{aligned} \tag{6.25}$$

such that conditional on v, y_d is deterministic where

$$\boldsymbol{P}(y_d|v) = \begin{cases}\mathbb{1}_{\{v\geq 0\}}, & y_d = 1\\ \mathbb{1}_{\{v<0\}}, & y_d = -1\end{cases} \tag{6.26}$$

and $\mathbb{1}_{\{.\}}$ represents the indicator function. We can therefore express (6.24) for each n as

$$I(\boldsymbol{x}_d) = \int_0^{\infty}\int_{\mathbb{R}^M}\phi(v, y_d\bar{\boldsymbol{h}}^T\boldsymbol{x}_d,\sigma_d^2).\phi_M(\bar{\boldsymbol{h}};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\bar{\boldsymbol{h}}\,\mathrm{d}v. \tag{6.27}$$

Marginalizing over the distribution of $\bar{\boldsymbol{h}}$ in this last equation for each n, and taking the negative logarithm we get (4.30).

Appendix F

To compute the metric in Eq. (4.18), we will need to generate samples of the channel that follow the posterior distribution from Eq. (4.20). We can make use of the following theorem proposed in [119]

Theorem (Durante [119]): *If* $\boldsymbol{y} = [y_1, y_2, \ldots, y_i, \ldots, y_T]^T$ *is a vector of conditionally independent binary data from a probit model* $(y_i|\boldsymbol{x}_i, \boldsymbol{h}) \sim \text{Bern}(\Phi(\boldsymbol{x}_i^T\boldsymbol{h}))$, *for* $i = 1, \ldots, k$ *and* $\boldsymbol{h} \sim \mathcal{N}_M(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *then*

$$(\boldsymbol{h}|\boldsymbol{y}, \boldsymbol{X}) \sim \text{SUN}_{M,T}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, \boldsymbol{\eta}, \boldsymbol{\Gamma}) \tag{6.28}$$

Where $\text{SUN}_{M,T}$ represents the skewed-unified normal distribution family with density function

$$\phi_M(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\boldsymbol{\Phi}_T\Big(\boldsymbol{\eta} + \boldsymbol{\Delta}^T\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{w}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}), \boldsymbol{0}_T, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^T\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Delta}\Big)}{\boldsymbol{\Phi}_T(\boldsymbol{\eta}; \boldsymbol{0}_T, \boldsymbol{\Gamma})} \tag{6.29}$$

such that $\boldsymbol{\Phi}_T$ denotes the T-dimensional cumulative multivariate Gaussian integral evaluated at $\boldsymbol{\eta}+\boldsymbol{\Delta}^T\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{w}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})$ representing the orthants with mean $\boldsymbol{0}_T$ and covariance matrix $\boldsymbol{\Gamma} - \boldsymbol{\Delta}^T\bar{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Delta}$. $\boldsymbol{w}$ is a diagonal matrix containing the square roots of the diagonal elements in $\boldsymbol{\Sigma}$, and we have the following relations

$$\begin{cases} \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^T \\ \boldsymbol{\Sigma} = \boldsymbol{w}\bar{\boldsymbol{\Sigma}}\boldsymbol{w} \\ \boldsymbol{\Delta} = \bar{\boldsymbol{\Sigma}}\boldsymbol{w}\boldsymbol{D}^T\boldsymbol{s}^{-1} \\ \boldsymbol{\eta} = \boldsymbol{s}^{-1}\boldsymbol{D}\boldsymbol{\mu} \\ \boldsymbol{\Gamma} = \boldsymbol{s}^{-1}(\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T + \boldsymbol{I}_T)\boldsymbol{s}^{-1} \end{cases} \tag{6.30}$$

where $\boldsymbol{D} = \text{diag}(y_1, \ldots, y_T)\boldsymbol{X}$ and $\boldsymbol{s} = \text{diag}\big\{(\boldsymbol{d}_1^T\boldsymbol{\Sigma}\boldsymbol{d}_1+1)^{1/2}, \ldots, (\boldsymbol{d}_T^T\boldsymbol{\Sigma}\boldsymbol{d}_T+1)^{1/2}\big\}$.

Corollary (Durante [119]): *If* $(\boldsymbol{h}|\boldsymbol{y}, \boldsymbol{X})$ *has the unified skew-normal distribution from Theorem 1, then*

$$(\boldsymbol{h}|\boldsymbol{y}, \boldsymbol{X}) \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{w}\{\boldsymbol{V}_0 + \bar{\boldsymbol{\Sigma}}\boldsymbol{w}\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T + \boldsymbol{I}_T)^{-1}\boldsymbol{s}\boldsymbol{V}_1\} \tag{6.31}$$

with the following distributions

$$\begin{cases} \boldsymbol{V}_0 \sim \mathcal{N}\big(\boldsymbol{0}_M, \bar{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}}\boldsymbol{w}\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T + \boldsymbol{I}_T)^{-1}\boldsymbol{w}\bar{\boldsymbol{\Sigma}}\big) \\ \boldsymbol{V}_1 \sim \mathcal{TN}\big(-\boldsymbol{s}^{-1}\boldsymbol{D}\boldsymbol{\mu}, \boldsymbol{0}_T, \boldsymbol{s}^{-1}(\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T + \boldsymbol{I}_T)\boldsymbol{s}^{-1}\big) \end{cases} \tag{6.32}$$

where the notation $\stackrel{d}{=}$ means equality in distribution and $\mathcal{TN}(\boldsymbol{l}, \boldsymbol{a}, \boldsymbol{C})$ is a truncated normal distribution from below $\boldsymbol{l}$ with mean and covariance matrix $\boldsymbol{a}$ and $\boldsymbol{C}$, respectively. Looking at the simplified posterior from Eq. (4.33), it can be directly verified that $h_{n,m}$, implicitly conditioned on the data, follows the skew-normal distribution after a simple re-parameterization according to the previous theorem

$$p(\bar{h}_{n,m}) = \phi(\bar{h}_{n,m}, 0, 1)\frac{\boldsymbol{\Phi}_\gamma\{\frac{\bar{h}_{n,m}}{\sqrt{2}}\bar{\boldsymbol{X}}_\gamma, 0, \frac{1}{2}\boldsymbol{I}_\gamma\}}{\boldsymbol{\Phi}_\gamma\big\{\boldsymbol{0}, \frac{1}{2}(\boldsymbol{I}_\gamma + \bar{\boldsymbol{X}}_\gamma\bar{\boldsymbol{X}}_\gamma^T)\big\}}, \tag{6.33}$$

where we have $\bar{\boldsymbol{X}}_\gamma = \frac{1}{\sqrt{\sigma_p}}\text{diag}(y_{p,n}^1, \ldots, y_{p,n}^\gamma)\mathbb{1}_\gamma$. Since everything is independent, to perform MC integration we can simply generate the random variables $h_{n,m}$ in parallel according to the Corollary:

$$\begin{cases} v_0 \sim \mathcal{N}(0, 1 - \bar{\boldsymbol{X}}_\gamma^T(\bar{\boldsymbol{X}}_\gamma\bar{\boldsymbol{X}}_\gamma^T + \boldsymbol{I}_\gamma)^{-1}\bar{\boldsymbol{X}}_\gamma) \\ \boldsymbol{V}_1 \sim \mathcal{TN}_\gamma(\boldsymbol{0}; \boldsymbol{0}_\gamma, \frac{1}{2}(\bar{\boldsymbol{X}}_\gamma\bar{\boldsymbol{X}}_\gamma^T + \boldsymbol{I}_\gamma)) \end{cases}$$

Therefore,

$$\bar{h}_{n,m} \overset{d}{=} v_0 + \sqrt{2}\bar{\boldsymbol{X}}_\gamma^T(\bar{\boldsymbol{X}}_\gamma\bar{\boldsymbol{X}}_\gamma^T + \boldsymbol{I}_\gamma)^{-1}\boldsymbol{V}_1. \tag{6.34}$$

The truncated normal distribution can be generated according to the method proposed in [117].

Bibliography

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] H. Viswanathan and P. E. Mogensen, "Communications in the 6G Era," *IEEE Access*, vol. 8, pp. 57 063–57 074, 2020.

[3] N. Rajatheva, I. Atzeni, E. Bjornson, A. Bourdoux, S. Buzzi, J.-B. Dore, S. Erkucuk, M. Fuentes, K. Guan, Y. Hu, X. Huang, J. Hulkkonen, J. M. Jornet, M. Katz, R. Nilsson, E. Panayirci, K. Rabie, N. Rajapaksha, M. Salehi, H. Sarieddeen, T. Svensson, O. Tervo, A. Tolli, Q. Wu, and W. Xu, "White Paper on Broadband Connectivity in 6G," Apr. 2020. [Online]. Available: http://arxiv.org/abs/2004.14247

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," *IEEE Access*, vol. 7, pp. 78 729–78 757, 2019.

[6] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, May 2020.

[7] A. I. Perez-Neira, M. A. Vazquez, M. B. Shankar, S. Maleki, and S. Chatzinotas, "Signal Processing for High-Throughput Satellites: Challenges in New Interference-Limited Scenarios," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 112–131, Jul. 2019.

[8] M. Giordani and M. Zorzi, "Non-Terrestrial Networks in the 6G Era: Challenges and Opportunities," *IEEE Network*, vol. 35, no. 2, pp. 244–251, Mar. 2021.

[9] M. Shinagawa, Y. Akazawa, and T. Wakimoto, "Jitter analysis of high-speed sampling systems," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 1, pp. 220–224, Feb. 1990.

[10] M. Shimanouchi, "An approach to consistent jitter modeling for various jitter aspects and measurement methods," in *Proceedings International Test Conference 2001 (Cat. No.01CH37260)*. Baltimore, MD, USA: IEEE, 2001, pp. 848–857.

[11] W. Kester, "Aperture Time, Aperture Jitter, Aperture Delay Time— Removing the Confusion," Analog Devices, Tech. Rep. MT-007 Tutorial, Oct. 2008. [Online]. Available: https://www.analog.com/media/en/training-seminars/tutorials/MT-007.pdf

[12] M. Löhning and G. Fettweis, "The effects of aperture jitter and clock jitter in wideband ADCs," *Computer Standards & Interfaces*, vol. 29, no. 1, pp. 11–18, Jan. 2007.

[13] Z. J. Towfic, Shang-Kee Ting, and A. H. Sayed, "Clock Jitter Compensation in High-Rate ADC Circuits," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5738–5753, Nov. 2012.

[14] R. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, Apr. 1999.

[15] B. Murmann, "The Race for the Extra Decibel: A Brief Review of Current ADC Performance Trajectories," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 3, pp. 58–66, 2015.

[16] B. Widrow, I. Kollar, and Ming-Chang Liu, "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353–361, Apr. 1996.

[17] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[18] T. Schenk and J.-P. Linnartz, *RF imperfections in high-rate wireless systems: impact and digital compensation*. Dordrecht: Springer, 2008.

[19] A. Moschitta, J. Schoukens, and P. Carbone, "Information and Statistical Efficiency When Quantizing Noisy DC Values," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 308–317, Feb. 2015.

[20] M. R. Khanzadi, *Phase noise in communication systems: modeling, compensation, and performance analysis.* Göteborg: Chalmers University of Technology, 2015.

[21] S. Bicais and J.-B. Dore, "Design of Digital Communications for Strong Phase Noise Channels," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 227–243, 2020.

[22] A. Katz, "Linearization: reducing distortion in power amplifiers," *IEEE Microwave Magazine*, vol. 2, no. 4, pp. 37–49, Dec. 2001.

[23] M. Schetzen, "Theory of pth-order inverses of nonlinear systems," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 5, pp. 285–291, May 1976.

[24] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems*, vol. 32, no. 11, pp. 1150–1161, Nov. 1985.

[25] S. Benedetto and E. Biglieri, *Principles of Digital Transmission*, ser. Information Technology: Transmission, Processing, and Storage. Boston: Kluwer Academic Publishers, 2002.

[26] R. Raich and G. Zhou, "On the modeling of memory nonlinear effects of power amplifiers for communication applications," in *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop.* Pine Mountain, GA, USA: IEEE, 2002, pp. 7–10.

[27] F. M. Ghannouchi, O. Hammi, and M. Helaoui, *Behavioral Modeling and Predistortion of Wideband Wireless Transmitters*, 1st ed. Wiley, Jun. 2015.

[28] M. De Sanctis, E. Cianca, T. Rossi, C. Sacchi, L. Mucchi, and R. Prasad, "Waveform design solutions for EHF broadband satellite communications," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 18–23, Mar. 2015.

[29] I. Peruga Nasarre, T. Levanen, and M. Valkama, "Constrained PSK: Energy-Efficient Modulation for Sub-THz Systems," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops).* Dublin, Ireland: IEEE, Jun. 2020, pp. 1–7.

[30] A. Katz, J. Wood, and D. Chokola, "The Evolution of PA Linearization: From Classic Feedforward and Feedback Through Analog and

Digital Predistortion," *IEEE Microwave Magazine*, vol. 17, no. 2, pp. 32–40, Feb. 2016.

[31] S. Benedetto and E. Biglieri, "Nonlinear Equalization of Digital Satellite Channels," *IEEE Journal on Selected Areas in Communications*, vol. 1, no. 1, pp. 57–62, Jan. 1983.

[32] B. F. Beidas, "Intermodulation Distortion in Multicarrier Satellite Systems: Analysis and Turbo Volterra Equalization," *IEEE Transactions on Communications*, vol. 59, no. 6, pp. 1580–1590, Jun. 2011.

[33] A. Piemontese, A. Modenini, G. Colavolpe, and N. S. Alagha, "Improving the Spectral Efficiency of Nonlinear Satellite Systems through Time-Frequency Packing and Advanced Receiver Processing," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3404–3412, Aug. 2013.

[34] O. B. Usman and A. Knopp, "Digital Predistortion in High Throughput Satellites: Architectures and Performance," *IEEE Access*, vol. 9, pp. 42 291–42 304, 2021.

[35] J. Tsimbinos and K. Lever, "Applications of higher-order statistics to modelling, identification and cancellation of nonlinear distortion in high-speed samplers and analogue-to-digital converters using the Volterra and Wiener models," in *[1993 Proceedings] IEEE Signal Processing Workshop on Higher-Order Statistics*. South Lake Tahoe, CA, USA: IEEE, 1993, pp. 379–383.

[36] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," Apr. 2014, arXiv:1404.7736 [cs, math].

[37] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized Precoding for Massive MU-MIMO," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[38] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On Low-Resolution ADCs in Practical 5G Millimeter-Wave Massive MIMO Systems," *IEEE Communications Magazine*, vol. 56, no. 7, pp. 205–211, Jul. 2018.

[39] J. Liu, Z. Luo, and X. Xiong, "Low-Resolution ADCs for Wireless Communication: A Comprehensive Survey," *IEEE Access*, vol. 7, pp. 91 291–91 324, 2019.

[40] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "Linear and Deep Neural Network-Based Receivers for Massive MIMO Systems With One-Bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7333–7345, Nov. 2021.

[41] K. Xenoulis and N. Kalouptsidis, "Achievable Rates for Nonlinear Volterra Channels," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1237–1248, Mar. 2011.

[42] J. Qi and S. Aissa, "On the Power Amplifier Nonlinearity in MIMO Transmit Beamforming Systems," *IEEE Transactions on Communications*, vol. 60, no. 3, pp. 876–887, Mar. 2012.

[43] M. Fozooni, M. Matthaiou, E. Bjornson, and T. Q. Duong, "Performance Limits of MIMO Systems with Nonlinear Power Amplifiers," in *2015 IEEE Global Communications Conference (GLOBECOM)*. San Diego, CA, USA: IEEE, Dec. 2015, pp. 1–7.

[44] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Transactions on Communications*, vol. 57, no. 12, pp. 3629–3639, Dec. 2009.

[45] J. Mo and R. W. Heath, "Capacity Analysis of One-Bit Quantized MIMO Systems With Transmitter Channel State Information," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[46] 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*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.

[47] O. Kanhere, H. Poddar, Y. Xing, D. Shakya, S. Ju, and T. S. Rappaport, "A Power Efficiency Metric for Comparing Energy Consumption in Future Wireless Networks in the Millimeter-Wave and Terahertz Bands," *IEEE Wireless Communications*, vol. 29, no. 6, pp. 56–63, Dec. 2022.

[48] S. Wesemann, J. Du, and H. Viswanathan, "Energy Efficient Extreme MIMO: Design Goals and Directions," *IEEE Communications Magazine*, vol. 61, no. 10, pp. 132–138, Oct. 2023.

[49] A. Lozano and S. Rangan, "Spectral vs Energy Efficiency in 6G: Impact of the Receiver Front-End," Oct. 2023, arXiv:2310.02622

[cs, eess, math]. [Online]. Available: http://arxiv.org/abs/2310.02622

[50] Ickhyun Song, Jongwook Jeon, Hee-Sauk Jhon, Junsoo Kim, Byung-Gook Park, Jong Duk Lee, and Hyungcheol Shin, "A Simple Figure of Merit of RF MOSFET for Low-Noise Amplifier Design," *IEEE Electron Device Letters*, vol. 29, no. 12, pp. 1380–1382, Dec. 2008.

[51] Bin Le, T. Rondeau, J. Reed, and C. Bostian, "Analog-to-digital converters," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 69–77, Nov. 2005.

[52] B. Murmann, "ADC Performance Survey 1997-2024." [Online]. Available: https://github.com/bmurmann/ADC-survey.

[53] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[54] S. Hoyos, B. Sadler, and G. Arce, "Monobit digital receivers for ultrawideband communications," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1337–1344, Jul. 2005.

[55] A. Mezghani and J. A. Nossek, "On Ultra-Wideband MIMO Systems with 1-bit Quantized Outputs: Performance Analysis and Input Optimization," in *2007 IEEE International Symposium on Information Theory*. Nice: IEEE, Jun. 2007, pp. 1286–1289.

[56] O. Dabeer, J. Singh, and U. Madhow, "On the Limits of Communication Performance with One-Bit Analog-To-Digital Conversion," in *2006 IEEE 7th Workshop on Signal Processing Advances in Wireless Communications*. Cannes, France: IEEE, Jul. 2006, pp. 1–5.

[57] A. Mezghani and J. A. Nossek, "Analysis of Rayleigh-fading channels with 1-bit quantized output," in *2008 IEEE International Symposium on Information Theory*. Toronto, ON, Canada: IEEE, Jul. 2008, pp. 260–264.

[58] E. Masry, "The reconstruction of analog signals from the sign of their noisy samples," *IEEE Transactions on Information Theory*, vol. 27, no. 6, pp. 735–745, Nov. 1981.

[59] H. Papadopoulos, G. Wornell, and A. Oppenheim, "Sequential signal encoding from noisy measurements using quantizers with dynamic bias control," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 978–1002, Mar. 2001.

[60] O. Dabeer and E. Masry, "Multivariate Signal Parameter Estimation Under Dependent Noise From 1-Bit Dithered Quantized Data," *IEEE Transactions on Information Theory*, vol. 54, no. 4, pp. 1637–1654, Apr. 2008.

[61] R. Gray and T. Stockham, "Dithered quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[62] J. Rapp, R. M. A. Dawson, and V. K. Goyal, "Estimation From Quantized Gaussian Measurements: When and How to Use Dither," *IEEE Transactions on Signal Processing*, vol. 67, no. 13, pp. 3424–3438, Jul. 2019.

[63] T. Lok and V.-W. Wei, "Channel estimation with quantized observations," in *Proceedings. 1998 IEEE International Symposium on Information Theory.* Cambridge, MA, USA: IEEE, 1998, p. 333.

[64] M. T. Ivrlac˘ and J. A. Nossek, "On MIMO Channel Estimation with Single-Bit Signal-Quantization," in *Proc. ITG Workshop Smart Antennas*, Feb. 2007.

[65] T. Koch and A. Lapidoth, "At Low SNR, Asymmetric Quantizers are Better," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, Sep. 2013.

[66] A. Mezghani and J. A. Nossek, "Capacity Lower Bound of MIMO Channels with Output Quantization and Correlated Noise," in *Proc. IEEE International Symposium on Information Theory*, 2012.

[67] A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, "Low SNR Asymptotic Rates of Vector Channels With One-Bit Outputs," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7615–7634, Dec. 2020.

[68] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel Estimation and Performance Analysis of One-Bit Massive MIMO Systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[69] K. Gao, J. N. Laneman, N. J. Estes, J. Chisum, and B. Hochwald, "Training for Channel Estimation in Nonlinear Multi-Antenna Transceivers," Dec. 2019, arXiv:1912.06924 [cs, math]. [Online]. Available: http://arxiv.org/abs/1912.06924

[70] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-Optimal Joint Channel-and-Data Estimation for Massive MIMO With Low-Precision ADCs," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[71] J. Choi, J. Mo, and R. W. Heath, "Near Maximum-Likelihood Detector and Channel Estimator for Uplink Multiuser Massive MIMO Systems With One-Bit ADCs," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2005–2018, May 2016.

[72] C. Studer and G. Durisi, "Quantized Massive MU-MIMO-OFDM Uplink," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[73] 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*, vol. 16, no. 1, pp. 87–100, Jan. 2017.

[74] Q. Wan, J. Fang, H. Duan, Z. Chen, and H. Li, "Generalized Bussgang LMMSE Channel Estimation for One-Bit Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4234–4246, Jun. 2020.

[75] I. Atzeni and A. Tolli, "Channel Estimation and Data Detection Analysis of Massive MIMO With 1-Bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3850–3867, Jun. 2022.

[76] H. Chen, P. K. Varshney, S. M. Kay, and J. H. Michels, "Theory of the Stochastic Resonance Effect in Signal Detection: Part I—Fixed Detectors," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3172–3184, Jul. 2007.

[77] Y. Dong, H. Wang, and Y.-D. Yao, "Channel Estimation for One-Bit Multiuser Massive MIMO Using Conditional GAN," *IEEE Communications Letters*, vol. 25, no. 3, pp. 854–858, Mar. 2021.

[78] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "Model-Inspired Deep Detection with Low-Resolution Receivers,"

in *2021 IEEE International Symposium on Information Theory (ISIT)*. Melbourne, Australia: IEEE, Jul. 2021, pp. 3349–3354.

[79] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "SVM-Based Channel Estimation and Data Detection for One-Bit Massive MIMO Systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2086–2099, 2021.

[80] Y.-S. Jeon, D. Kim, S.-N. Hong, N. Lee, and R. W. Heath, "Artificial Intelligence for Physical-Layer Design of MIMO Communications with One-Bit ADCs," *IEEE Communications Magazine*, vol. 60, no. 7, pp. 76–81, Jul. 2022.

[81] Y.-S. Jeon, N. Lee, S.-N. Hong, and R. W. Heath, "One-Bit Sphere Decoding for Uplink Massive MIMO Systems With One-Bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4509–4521, Jul. 2018.

[82] J. Park, S. Park, A. Yazdan, and R. W. Heath, "Optimization of Mixed-ADC Multi-Antenna Systems for Cloud-RAN Deployments," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3962–3975, Sep. 2017.

[83] A. B. Ucuncu and A. O. Yilmaz, "Oversampling in One-Bit Quantized Massive MIMO Systems and Performance Analysis," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7952–7964, Dec. 2018.

[84] R. Deng, J. Zhou, and W. Zhang, "Bandlimited Communication With One-Bit Quantization and Oversampling: Transceiver Design and Performance Evaluation," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 845–862, Feb. 2021.

[85] A. Lozano, "1-Bit MIMO for Terahertz Channels," Sep. 2021, arXiv:2109.04390 [cs, math]. [Online]. Available: http://arxiv.org/abs/2109.04390

[86] R. Combes and S. Yang, "An Approximate ML Detector for MIMO Channels Corrupted by Phase Noise," *IEEE Transactions on Communications*, vol. 66, no. 3, pp. 1176–1189, Mar. 2018.

[87] Y.-S. Jeon, S.-N. Hong, and N. Lee, "Supervised-Learning-Aided Communication Framework for MIMO Systems With Low-Resolution ADCs," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7299–7313, Aug. 2018.

[88] N. Chopin and J. Ridgway, "Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation," *Statistical Science*, vol. 32, no. 1, Feb. 2017.

[89] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "MIMO Networks With One-Bit ADCs: Receiver Design and Communication Strategies," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1580–1594, Mar. 2022.

[90] S. Rini, L. Barletta, Y. C. Eldar, and E. Erkip, "A general framework for MIMO receivers with low-resolution quantization," in *2017 IEEE Information Theory Workshop (ITW)*. Kaohsiung, Taiwan: IEEE, Nov. 2017, pp. 599–603.

[91] A. Wadhwa and U. Madhow, "Blind phase/frequency synchronization with low-precision ADC: A Bayesian approach," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Monticello, IL: IEEE, Oct. 2013, pp. 181–188.

[92] A. C. Ulusoy, S. Krone, G. Liu, A. Trasser, F. Guderian, B. Almeroth, A. Barghouthi, M. Hellfeld, S. Schumann, C. Carta, C. Estan, K. Dombrowski, V. Brankovic, D. Radovic, F. Ellinger, G. Fettweis, and H. Schumacher, "A 60 GHz multi-Gb/s system demonstrator utilizing analog synchronization and 1-bit data conversion," in *2013 IEEE 13th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems*. Austin, TX: IEEE, Jan. 2013, pp. 99–101.

[93] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[94] A. Li, C. Masouros, A. L. Swindlehurst, and W. Yu, "1-Bit Massive MIMO Transmission: Embracing Interference with Symbol-Level Precoding," *IEEE Communications Magazine*, vol. 59, no. 5, pp. 121–127, May 2021.

[95] J. J. Bussgang, "Bussgang_1952_cross-correlation functions of amplitude-distorted Gaussian inputs.pdf," Research Laboratory of Electronics, Massachusetts Institute of Technology, Tech. Rep., Mar. 1952, http://hdl.handle.net/1721.1/4847.

[96] A. Mezghani, M.-S. Khoufi, and J. A. Nossek, "A Modified MMSE Receiver for Quantized MIMO Systems," in *IEEE Workshop on Smart Antennas*. Vienna, Austria: IEEE, Jan. 2007.

[97] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," in *2015 Information Theory and Applications Workshop (ITA)*. San Diego, CA, USA: IEEE, Feb. 2015, pp. 191–198.

[98] O. T. Demir and E. Bjornson, "The Bussgang Decomposition of Nonlinear Systems: Basic Theory and MIMO Extensions [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 131–136, Jan. 2021.

[99] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IEEE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, Jun. 1958.

[100] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust Predictive Quantization: Analysis and Design Via Convex Optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 618–632, Dec. 2007.

[101] J. Max, "Quantizing for minimum distortion," *IEEE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[102] J. Bucklew and N. Gallagher, "A note on optimal quantization (Corresp.)," *IEEE Transactions on Information Theory*, vol. 25, no. 3, pp. 365–366, May 1979.

[103] ——, "Some properties of uniform step size quantizers (Corresp.)," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 610–613, Sep. 1980.

[104] L. Liu, Y. Ma, and N. Yi, "Hermite Expansion Model and LMMSE Analysis for Low-Resolution Quantized MIMO Detection," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5313–5328, 2021.

[105] S. Wang, Y. Li, and J. Wang, "Convex optimization based multiuser detection for uplink large-scale MIMO under low-resolution quantization," in *2014 IEEE International Conference on Communications (ICC)*. Sydney, NSW: IEEE, Jun. 2014, pp. 4789–4794.

[106] M. Damen, H. El Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2389–2402, Oct. 2003.

[107] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[108] U. Fincke and M. Pohst, "Improved Methods for Calculating Vectors of Short Length in a Lattice, Including a Complexity Analysis," *Math. Comp.*, vol. 44, pp. 463–471, Apr. 1985.

[109] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 7, pp. 1566–1577, Jul. 2005.

[110] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.

[111] L. You, X. Gao, X.-G. Xia, N. Ma, and Y. Peng, "Pilot Reuse for Massive MIMO Transmission over Spatially Correlated Rayleigh Fading Channels," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3352–3366, Jun. 2015.

[112] B. Hochwald and T. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.

[113] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[114] C. Thrampoulidis and W. Xu, "The Performance Of Box-Relaxation Decoding In Massive MIMO With Low-Resolution ADCS," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*. Freiburg im Breisgau, Germany: IEEE, Jun. 2018, pp. 821–825.

[115] A. Ganti, A. Lapidoth, and I. Telatar, "Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[116] G. Taricco and E. Biglieri, "Space-time decoding with imperfect channel estimation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1874–1888, Jul. 2005.

[117] Z. I. Botev, "The Normal Law Under Linear Restrictions: Simulation and Estimation via Minimax Tilting," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 1, pp. 125–148, Jan. 2017.

[118] K.-H. Ngo and S. Yang, "A Generalized Gaussian Model for Wireless Communications," in *2021 IEEE International Symposium on Information Theory (ISIT)*. Melbourne, Australia: IEEE, Jul. 2021, pp. 3237–3242.

[119] D. Durante, "Conjugate Bayes for probit regression via unified skew-normal distributions," *Biometrika*, vol. 106, no. 4, pp. 765–779, Dec. 2019, arXiv:1802.09565 [stat]. [Online]. Available: http://arxiv.org/abs/1802.09565

[120] K. Safa, R. De Lacerda, and S. Yang, "Channel Estimation and Data Detection in MIMO channels with 1-bit ADC using Probit Regression," in *2023 IEEE Information Theory Workshop (ITW)*. Saint-Malo, France: IEEE, Apr. 2023, pp. 457–461.

[121] B. Sun, X. Tang, Y. Zhou, Z. Pan, and H. Lin, "High Order QAM Modulation in Massive MIMO Systems With Asymmetrically Quantized 1-Bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 22, no. 9, pp. 6369–6382, Sep. 2023.

[122] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[123] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. Wiley, Sep. 2005.

[124] I. Abou-Faycal, M. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[125] E. Telatar, "Capacity of Multi-antenna Gaussian Channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[126] S. Krone and G. Fettweis, "Energy-Efficient A/D Conversion in Wideband Communications Receivers," in *2011 IEEE Vehicular Technology Conference (VTC Fall)*. San Francisco, CA, USA: IEEE, Sep. 2011, pp. 1–5.

[127] B. Qing, M. Amine, and J. A. Nossek, "On the Optimization of ADC Resolution in Multi-antenna Systems," in *2013 The Tenth*

International Symposium on Wireless Communication Systems. VDE, 2013, pp. 1–5.

[128] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[129] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37–60, Aug. 1994.

[130] S. Yang and R. Combes, "Asymptotic Capacity of Non-Coherent One-Bit MIMO Channels with Block Fading," in *2024 IEEE International Symposium on Information Theory (ISIT)*. Athens, Greece: IEEE, Jul. 2024, pp. 2359–2364.

[131] ——, "Asymptotic Capacity of 1-Bit MIMO Fading Channels," Jul. 2024, arXiv:2407.16242 [cs, math]. [Online]. Available: http://arxiv.org/abs/2407.16242

[132] F. Nielsen, "A Simple Approximation Method for the Fisher–Rao Distance between Multivariate Normal Distributions," *Entropy*, vol. 25, no. 4, p. 654, Apr. 2023.

[133] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 0th ed. Chapman and Hall/CRC, Nov. 2013.

[134] E. Gilbert, "Increased information rate by oversampling," *IEEE Transactions on Information Theory*, vol. 39, no. 6, pp. 1973–1976, Nov. 1993.

[135] S. Shamai, "Information rates by oversampling the sign of a bandlimited process," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1230–1236, Jul. 1994.

[136] T. Koch and A. Lapidoth, "Increased Capacity per Unit-Cost by Oversampling," Sep. 2010, arXiv:1008.5393 [cs, math]. [Online]. Available: http://arxiv.org/abs/1008.5393

[137] L. T. N. Landau, M. Dorpinghaus, R. C. De Lamare, and G. P. Fettweis, "Achievable Rate With 1-Bit Quantization and Oversampling Using Continuous Phase Modulation-Based Sequences," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 7080–7095, Oct. 2018.

[138] L. V. Nguyen, L. Liu, N. Linh-Trung, and A. L. Swindlehurst, "One-Bit Massive MIMO Precoding for Frequency-Selective Fading Channels," in *2023 IEEE Statistical Signal Processing Workshop (SSP)*. Hanoi, Vietnam: IEEE, Jul. 2023, pp. 324–328.

[139] E. Balevi and J. G. Andrews, "One-Bit OFDM Receivers via Deep Learning," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4326–4336, Jun. 2019.

[140] S. Teodoro, A. Silva, R. Dinis, F. M. Barradas, P. M. Cabral, and A. Gameiro, "Theoretical Analysis of Nonlinear Amplification Effects in Massive MIMO Systems," *IEEE Access*, vol. 7, pp. 172 277–172 289, 2019.

[141] K. Pedersen, P. Mogensen, and B. Fleury, "A stochastic model of the temporal and azimuthal dispersion seen at the base station in outdoor propagation environments," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 2, pp. 437–447, Mar. 2000.