Multi-User Detection with Oversampled Large Antenna Arrays and Low-resolution ADCs

by

Zied Jarraya

A Thesis submitted to The Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

University of Manitoba

Winnipeg

March 2024

When you look at yourself from a universal standpoint, something inside always reminds or informs you that there are bigger and better things to worry about.

ALBERT EINSTEIN

Abstract

This thesis investigates the uplink scenario in millimetre-wave (mmWave) massive multiple-input multiple-output (MIMO) communication systems characterized by dense, uniform linear arrays (ULAs) of antenna elements tightly packed within a confined space and equipped with low-resolution Analog-to-Digital Converters (ADCs). The primary focus of our study is to address the critical challenges of power consumption reduction and hardware simplification while simultaneously improving the performance of quantized systems by exploring spatial oversampling. Due to the consideration of subwavelength inter-element spacing in the ULA, extrinsic spatial thermal noise correlations arise due to significant coupling between adjacent antenna terminals. In addition to this correlated noise, the noise figure caused by hardware imperfections profoundly impacts signal recovery and cannot be dismissed as a negligible factor in system performance analysis. To tackle the problem of signal recovery in such high-density ULAs with low-resolution ADCs, we propose a non-linear inference method based on Vector Approximate Message Passing (VAMP) and Belief Propagation. Additionally, we employ a state evolution analysis to investigate the algorithm's asymptotic behaviour. The main objective of this algorithm is to reconstruct transmitted signals from the quantized measurements obtained by the coupled antennas. In this work, we demonstrate that employing oversampling techniques in the context of low-bit quantized systems can substantially enhance system performance, bringing it closer to the ideal scenario with infinite-resolution ADCs. Remarkably, this performance improvement persists even when the system is oversampled, emphasizing the potential of spatial oversampling as an effective strategy for enhancing the performance of low-resolution ADCs. We also analyze how the noise figure impacts the recovery in the context of

oversampling, highlighting its significance in system design considerations.

Keywords: Quantization, Massive MIMO, subwavelength oversampling, low-resolution analog-digital converters (ADCs), noise correlation, detection, vector approximate message passing (VAMP).

Acknowledgement

First and foremost, I extend my heartfelt gratitude to my esteemed mentors, Professor Amine Mezghani and Professor Faouzi Bellili, for their invaluable guidance and unwavering support throughout this research endeavour. Their mentorship has been instrumental in my academic and personal growth, and I am deeply appreciative of the opportunity to learn under their tutelage

I extend my sincere gratitude to the esteemed committee members, Professor Ekram Hossain and Professor Pradeepa Yahampath, for their valuable feedback to improve the quality of this thesis.

I also wish to express my profound appreciation to my family and friends for their unwavering support and encouragement throughout this journey. Their steadfast belief in me has been a constant source of motivation, and I am immensely grateful for the stability and encouragement they have provided as I pursue this endeavour.

.

Contents

List of Figures

List of Abbreviations

ADC	Analog-to-Digital-Converter
AMP	Approximate Message Passing
BER	Bit-Error-Rate
BP	Belief Propagation
BS	Base station
DFT	Discrete Fourier Transform
GLM	Generalized-Linear-Model
MAP	Maximum a Posteriori
MIMO	Multiple-input multiple-output
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
QPSK	Quadrature Phase Shift Keying
SE	State Evolution
SLM	Standard-Linear-Model
SNR	Signal-to-Noise-Ratio
SP	Sum-Product
ULA	Uniform Linear Array
VAMP	Vector Approximate Message Passing

Chapter 1

Introduction

1.1 Overview and related works

Massive multiple-input multiple-output (MIMO) communication systems are advocated as an integral part of the upcoming generations of wireless networks [2]- [3]. Massive MIMO employs large-scale antenna arrays (i.e., tens/hundreds of antenna elements), thereby allowing for enhanced spatial multiplexing and beamforming capabilities. This enables the base station (BS) to serve a large number of users using the same time-frequency resources, which is particularly attractive for massive connectivity use cases such as massive machine-type communications [4]. There are two competing massive MIMO architectures: hybrid (analog/digital) and fully digital architectures [5]. Fully digital massive is, however, more attractive due to the inherited full flexibility of the digital signal processing at the baseband (e.g., phase shifting in beamforming simplifies to a straightforward complex multiplication in the digital domain).

Nevertheless, a potential barrier to the practical realization of fully digital massive

MIMO systems lies in their elevated power consumption. In forthcoming millimetre-wave (mmWave) massive MIMO-equipped BSs, the deployment involves hundreds of antennas, with each antenna connected to a dedicated radio-frequency (RF) chain. Although significant progress in mmWave chip fabrication has yielded reductions in electronics costs, it is essential to note that the power consumption of Analog-to-Digital Converters (ADCs) along with the required automatic gain control (AGC) circuitry remain the predominant contributor to the overall power consumption of the RF chain. This situation can be attributed to two key factors. Firstly, the power consumption of ADCs exhibits a linear relationship with the sampling rate, a rate that tends to be substantial owing to the extensive bandwidth of mmWave signals. In addition, in a B-bit ADC, the power consumption experiences exponential growth with the increasing resolution B [6]- [7].

The challenge of high power consumption can be mitigated using low-resolution ADCs (1-3 bits). In fact, opting for low-resolution ADCs amounts to trading the (received) signal fidelity for power efficiency, provided that the performance degradation remains within acceptable levels. In recent studies [8,9], it has been shown that massive MIMO systems with low-resolution ADCs can efficiently handle multiple users transmitting high-order modulation symbols despite the implied compromise on performance. In addition, an advantage of using one-bit ADCs is dispensing with the need for automatic gain control units, resulting in low hardware complexity with less power consumption [10, 11]. Furthermore, it was shown in [12] that the power penalty for using 1-bit quantized ADCs is approximately equal to $\frac{2}{\pi}$ (i.e., 1.96 dB) at low signal-to-noise ratio (SNR). For more information on the benefits of using low-resolution data converters in massive MIMO, we refer the reader to the following work on energy efficiency [13–15], capacity analysis [16–18], channel estimation [19–21], and data detection [22–25].

A viable approach to overcome the performance degradation inherent to quantized systems involves leveraging oversampling in both time and space domains. This strategy is motivated by prior investigation [26] into the impact of the non-linearity introduced by the ADCs on the received analog signal. The presence of non-linearity induces a spectral broadening effect on the original signal, necessitating the consideration of higher sampling rates relative to the received signal. In essence, the concept of "oversampling" is pertinent to the analog received signal, aiming to increase the information extracted from the analog signal post-nonlinearity by sampling at higher rates. This statement holds, given the interchangeability between the sampling and quantization operations [27, 28].

In the context of spatial oversampling, studies [29, 30] have been conducted on utilizing dense, uniform arrays characterized by antenna interelement spacing smaller than half a wavelength. However, deploying such arrays gives rise to spatial correlations attributed to the close proximity of antennas [31]. This proximity leads to correlations in both the channel and noise, posing a significant challenge in the digital signal processing part, especially when considering the quantization of very low-resolution ADCs. Previous studies [32–36] have predominantly focused on linear iterative methods, such as spatial sigma-delta, which models the quantizer as an additive noise source. Additionally, the work outlined in reference [37] has approached the quantizer modelling as a linear transformation with additive noise, employing the Bussgang decomposition.

1.2 Contributions

In this work, we consider Multi-User MIMO systems. We explore diverse configurations of Uniform Linear Arrays (ULA) with varying inter-element spacing. To harmonize our study with more realistic physical phenomena, we consider a ray-based model for the

antenna array response and spatial noise correlations as in [37]. This model accounts for antenna coupling effects based on the law of power conservation, considering different antenna types. Additionally, we incorporate the impact of noise figure effects arising from low-noise amplifiers implemented at the receiver. A novel contribution of our work lies in developing a non-linear processing method for data detection. This method utilizes Vector Approximate Message Passing (VAMP) [1], an approximate message passing algorithm grounded in Bayes optimal inference. It yields a minimum mean squared error (MMSE), setting it apart from previous approaches. Importantly, this method handles channel and noise correlations and accommodates general B bits ADCs resolution. Furthermore, we employ the developed detection algorithm to examine the effects of oversampling for low-resolution ADCs. Our analysis includes isotropic and dipole antenna types, comparing them with the linear case with infinite resolution ADCs, emphasizing the 1-bit case. We also investigate the influence of noise figure on the detection algorithm. To gain insights into large system limits, we formulate the state evolution (SE) of the detection algorithm and compare it with empirical results. Our comprehensive studies reveal a noteworthy agreement between the algorithmic outcomes and previous theoretical findings. Notably, oversampling emerges as a crucial factor in significantly enhancing the performance of quantized systems, especially in the 1-bit case, bridging the gap towards the unquantized case.

1.3 Thesis Organization and Notations

The thesis is structured as follows:

- In Chapter 2, we examine the essential mathematical foundations of message-passing algorithms to comprehend the fundamental structure of the VAMP algorithm.

- In Chapter 3, we outline the system model for a ULA antenna configuration, addressing the detection problem in the context of general B-bits ADCs. We introduce a ray-based multi-path model to account for channel and noise correlation. We then present a VAMP-based detection algorithm and its SE tailored for low-resolution quantized systems capable of effectively managing noise correlation. Finally, we assess and compare the algorithmic outcomes with previous theoretical findings.

- Chapter 4 provides a conclusion to the thesis, highlighting potential avenues for future research.

Notations: In our notation, small boldface letters such as $\mathbf{x}$ represent vectors, while capital boldface letters such as $\mathbf{X}$ denote matrices. The symbol $(\cdot)^H$ signifies the Hermitian transpose operation. We use $x_n = [\mathbf{x}]_n$ to denote the n^{th} element of vector $\mathbf{x}$, and $\mathbf{x}_n$ to denote the n^{th} column of matrix $\mathbf{X}$. The operator $\mathrm{Tr}(\cdot)$ calculates the sum of the diagonal elements of a matrix, and $\mathbf{I}$ denotes the identity matrix. Additionally, $\mathcal{N}(\mathbf{x}; \widehat{\mathbf{x}}, \mathbf{R})$ and $\mathcal{CN}(\mathbf{x}; \widehat{\mathbf{x}}, \mathbf{R})$ represent the probability density functions of real multivariate Gaussian and complex multivariate Gaussian distributions, respectively, for any random vector $\mathbf{x}$ with mean $\widehat{\mathbf{x}}$ and covariance matrix $\mathbf{R}$. We use $\sim$ and $\propto$ as shorthand notations for "distributed according to" and "proportional to," respectively. Moreover, $\mathbb{E}[\mathbf{x}|d(\mathbf{x})]$ and $\mathrm{Cov}[\mathbf{x}|d(\mathbf{x})]$ denote the expectation and covariance of $\mathbf{x}$ given the distribution $d(\mathbf{x})$, while $\delta(\cdot)$ refers to the Dirac delta distribution. The functions $\mathfrak{R}(z)$ and $\mathfrak{I}(z)$ extract the real and imaginary parts of any complex variable z, respectively. Finally, $\Phi(x) \triangleq \int_{-\infty}^{x} \phi(t)\mathrm{d}t$ represents the cumulative normal distribution function, where $\phi(t) \triangleq \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ denotes the normal distribution density function.

Chapter 2

Background

In this chapter, we explore the mathematical core of message-passing algorithms, aiming to equip readers with the essential background to grasp the VAMP framework, central to our work. We start by discussing basic factor graphs, the sum-product algorithm, and the approximate message-passing algorithm before diving into the details of the VAMP algorithm.

2.1 Factor graphs

A factor graph is, by definition, a bipartite graph characterizing a factorization of a certain function $g(.)$, which depends on N real variables $(x_i)_{i\in[\![1,N]\!]}$. In other words, having an M factors factorization of a given function $g(x_1, .., x_N)$ (global function) of variables $x_1, .., x_N$.

$$g\left(x_1, x_2, \ldots, x_N\right) = \prod_{j=1}^{M} f_j\left(S_j\right), \tag{2.1}$$

where, f_j, $j \in [\![1, M]\!]$ are the factors of this factorization (local functions), each depends on a non-empty subset of variables $S_j \subset X = \{x_1, ..., x_N\}$. We define then the set $S = \{S_j;\ j \in [\![1, M]\!]\}$. Generally, the graph representing the given function is an ordered pair $G = (V, E)$. Specifically, V denotes the graph's vertices composed of elements from both sets S and X; thus, $V = X \bigcup S$, which conducts two types of nodes: variable and factor nodes. Moreover, E denotes the set of edges relating between the vertices. Nevertheless, a link is considered in this case as the dependence between a variable node and a factor node only; hence, $E = \{\{x_i, S_j\} \mid x_i \in X, S_j \in S; x_i \in S_j\}$. Practically, the graph arguments of a factor graph are formed merely out of the two sets X and S. Accordingly, the factor graph is denoted as $F(X, S)$, which is a sufficient blueprint for a given representation. For instance, considering the ensuing function $g(\cdot)$:

$$g(x_1, x_2, x_3) = f_1(x_1) f_2(x_1, x_2) f_3(x_1, x_2) f_4(x_2, x_3). \tag{2.2}$$

$g(\cdot)$ has four factorization functions and is parameterized with three variables. Therefore, regarding the above formulation, we exhibit the set of variable nodes $X = \{x_1, x_2, x_3\}$ and the set of factor nodes $S = \{\{x_1\}, \{x_1, x_2\}, \{x_1, x_2\}, \{x_2, x_3\}\}$ which leads to its corresponding factor graph $F(X, S)$ presented in Fig. 2.1. However, without loss of validity, a subset S_j is represented by its corresponding function f_j in the factor graph. In Fig. 2.1, variable nodes are designated by empty circles, whereas factor nodes are in dark-shaded squares. Generally, factor graphs are widely used in several fields, such as probability theory and constrained optimization. Still, the representation may differ; for instance, Bayesian networks and Markov random field graphs can be manifested by factor graphs presented with causalities. Thus, applying factor graphs for probabilistic graphical models has demonstrated empirical success in Bayesian inference problems. In the sequel, we consider the most notable algorithm based on factor graphs, generally known as the Sum-Product (SP) algorithm.

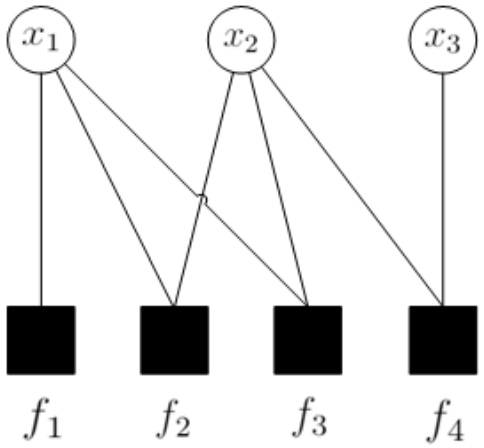


Fig. 2.1: Factor graph example.

2.2 Sum-Product (SP) algorithm

The SP algorithm [38], also known as Belief Propagation (BP), uses factor graphs by presenting the joint distribution factorization of given random variables. Similarly to Bayesian networks, the latter factorization follows the Bayes theorem, which expresses the dependencies of the variables. Furthermore, the algorithm performs inference over the graphical model by calculating the marginal distributions of the unobserved variable nodes, which are conditioned by the observed ones. In the sequel, without loss of generality, we consider an illustrative example by which we will give the basis and the motivation of the SP algorithm.

An illustrative example

Let's assume the probability joint mass function g of five real-valued unobserved variables with the factor factorization in Equation (2.3).

$$\begin{aligned} p(x_1, x_2, x_3, x_4, x_5|y) &= g\left(x_1, x_2, x_3, x_4, x_5\right) \\ &= f_A\left(x_1\right) f_B\left(x_2\right) f_C\left(x_1, x_2, x_3\right) f_D\left(x_3, x_4\right) f_E\left(x_3, x_5\right). \end{aligned} \tag{2.3}$$

As in any inference problem, we require the calculation of the marginal distributions, however, in this example we only consider the marginal of x_1. Promptly, before proceeding, we establish first the summary notation of a multivariable function f for a variable x_i, defined using the operator "$\downarrow$", called "summary operator", as the marginal function as in Equation (2.4).

$$f(X) \downarrow x_i = \sum_{x_j \in X \setminus \{x_i\}} f(X). \tag{2.4}$$

Hence, we write the marginal distribution of x_1 as in Equation (2.5).

$$\begin{aligned} p(x_1 \mid y) &= g(x_1, x_2, x_3, x_4, x_5) \downarrow x_1 \\ &= \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5) \\ &= f_A(x_1) \sum_{x_2} f_B(x_2) \sum_{x_3} f_C(x_1, x_2, x_3) \underbrace{\underbrace{\sum_{x_4} f_D(x_3, x_4)}_{f_D(x_3,x_4)\downarrow x_3} \underbrace{\sum_{x_5} f_E(x_3, x_5)}_{f_E(x_3,x_5)\downarrow x_3}}_{f_{DE}(x_3,x_4,x_5)\downarrow x_3}. \\ &\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{f_{BCDE}(x_1,x_2,x_3,x_4,x_5)\downarrow x_1} \end{aligned} \tag{2.5}$$

Subsequently, by examining the Equation (2.5), we observe that the marginal expression is determined by just knowing $f_A(x_1)$ and $f_{BCDE}(x_1, x_2, x_3, x_4, x_5) \downarrow x_1$. Similarly, the latter factor is immediately obtained by just having $f_B(x_2)$, $f_C(x_1, x_2, x_3)$ and $f_{DE}(x_3, x_4, x_5) \downarrow x_3$. Likewise, $f_{DE}(x_3, x_4, x_5) \downarrow x_3$ is finally yielded by knowing $f_D(x_3, x_4) \downarrow x_3$ and $f_E(x_3, x_5) \downarrow x_3$.

These products can also be represented in a distributed manner within the edges of the issued factor graph for $g(\cdot)$ shown in Fig. 2.2. Hence, inspired by computing systems, each node in the factor graph could be simulated as a processor that performs local computations. Moreover, it communicates with its adjacent neighbour processor by sending messages using a medium exemplified by the edges in the factor graph. Precisely,

these messages describe the summaries of multiple local function products. Eventually, it displays similarities to the message-passing technique within a shared memory system. Given the embodiment presented in Fig. 2.2 and exploring other variable marginals, it

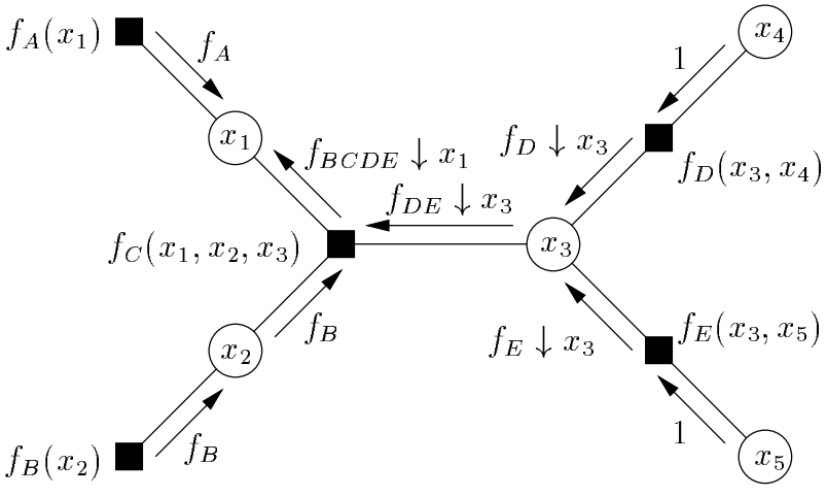


Fig. 2.2: The illustrated example message passing to obtain the marginal of x_1 .

gives us the basis of the SP algorithm that resides in the elementary messages between nodes that depend on the sens. Consequently, it helps considerably avoid unnecessary recomputations and instructions for each marginal. Additionally, the SP algorithm offers update rules which generalize the procedure of determining the flown messages between the nodes.

SP update rule

Forthwith, we present the one fundamental update rule that the SP algorithm follows:

SP Update Rule: The concrete message sent by a node v over an edge e is the product of all message received through all edges except for e multiplied by the local function presented in v, which is a unit function if v is not a factor node, and finally, the whole product is summarized for the corresponding variable on edge e.

Hence, we can derive separately the two mathematical expressions in (2.6) and

(2.7) that presents update rules in both cases; the message sent from a variable x to a local function f and the message sent from a local function f to a variable x. The Fig. 2.3 helps understand the latter expressions.

$$\mu_{x\to f}(x) = \prod_{h\in n(x)\setminus\{f\}} \mu_{h\to x}(x), \tag{2.6}$$

$$\mu_{f\to x}(x) = \left(f\left(X_{n(f)}\right) \prod_{y\in n(f)\setminus\{x\}} \mu_{y\to f}(y) \right) \downarrow x. \tag{2.7}$$

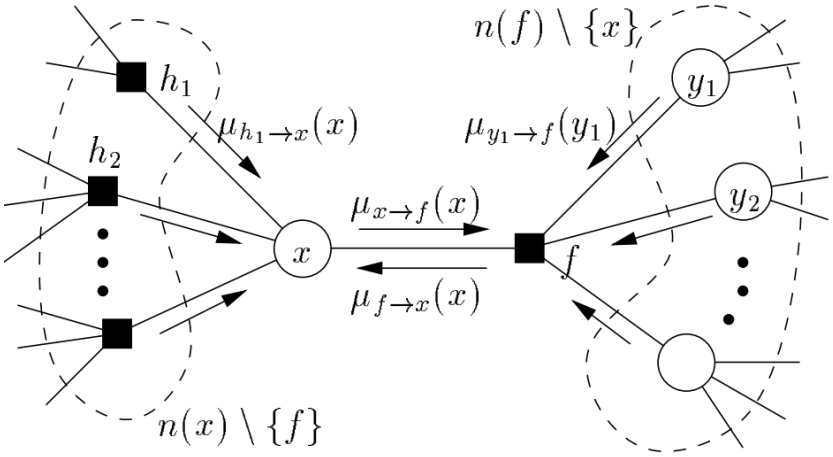


Fig. 2.3: A factor graph fragment that shows the SP algorithm update rules.

2.3 AMP algorithm

Message-passing algorithms, such as the SP algorithm, present drawbacks that make solving high-dimensional problems difficult. Firstly, messages are frequently taken as probability density measures over $\mathbb{R}$; therefore, the computations would be highly impractical. Additionally, these algorithms demand predefined priors that are unavailable for numerous applications. Lastly, factor graphs that display high density have an

immense number of messages that need to be determined, so this procedure presents huge complexity. As a result, justified approximate approaches to such models, like in Approximate Message Passing (AMP) algorithm [39], are required.

2.3.1 Relation to the iterative soft thresholding

In the sequel, we acquaint ourselves with the AMP algorithm derived from message passing on factor graphs, similar to the standard SP algorithm. Yet, in this section, we focus on the standard linear regression problem, where the goal is to recover a sparse vector $\mathbf{x}$ from noisy linear observations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, which is the main application of the AMP algorithm. In this problem, the update rules of the SP algorithm are harder to handle than those of AMP. For instance, if we consider recovering n signals from m observations, the factor graph would require the calculation of $2mn$ messages, which is computationally unbearable for large systems. Nevertheless, AMP offers a low complexity resolution, reconstruction power, and high phase transition. In fact, these characteristics are mostly influenced by the similarities of AMP to the iterative soft thresholding algorithm. To show that, we must exploit first the generic form of the latter algorithm in (2.8) :

$$\begin{aligned} \mathbf{x}^{t+1} &= \eta\left(\mathbf{x}^t + \mathbf{A}^\top \mathbf{z}^t; b_t\right), \\ \mathbf{z}^t &= \mathbf{y} - \mathbf{A}\mathbf{x}^t. \end{aligned} \quad . \tag{2.8}$$

where η is the threshold function defined by its threshold $\mathbf{b}$ as $\eta(\mathbf{x}; \mathbf{b}) = \text{sign}(\mathbf{x})|\mathbf{x} - \mathbf{b}|$ being applied elementwise, t refers to the state or iteration index, $\mathbf{x}^t$ is the estimate at state t and $\mathbf{z}^t$ is the residual error. However, the convergence rate of this algorithm is still slow, even if we consider a variable threshold state. This is precisely where AMP is more advantageous by improving the convergence speed rate of the iterative soft thresholding by using the so-called Onsager correction term inspired by statistical

physics. The algorithmic steps of AMP are summarized as follows:

$$\begin{aligned}
\mathbf{x}^0 &= \mathbf{0}, \\
\mathbf{x}^{t+1} &= \eta\left(\mathbf{x}^t + \mathbf{A}^\top \mathbf{z}^t; (\widehat{\gamma}^{-1})^t\right), \\
\mathbf{z}^t &= \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{n}{m}\mathbf{z}^{t-1}\left\langle \eta'\left(\mathbf{x}^{t-1} + \mathbf{A}^\top \mathbf{z}^{t-1}; (\widehat{\gamma}^{-1})^{t-1}\right)\right\rangle, \\
(\widehat{\gamma}^{-1})^t &= \frac{n}{m}(\widehat{\gamma}^{-1})^{t-1}\left\langle \eta'\left(\mathbf{A}^\top \mathbf{z}^{t-1} + \mathbf{x}^t; (\widehat{\gamma}^{-1})^{t-1}\right)\right\rangle.
\end{aligned} \tag{2.9}$$

2.3.2 AMP derivation steps

The AMP algorithm derivation follows four basic steps:

1. Construction of the graphical model corresponding to the joint-distribution (2.10) over the variables $\mathbf{x} = (x_1, ..., x_n)^\top$ willed to be recovered from m measurements $\mathbf{y} = (y_1, ..., y_m)^\top$ from a linear transformation.

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{i=1}^{n}\exp\left(-\beta\left|x_i\right|\right)\prod_{a=1}^{m}\delta_{\{y_a=(\mathbf{A}\mathbf{x})_a\}}. \tag{2.10}$$

2. Under large systems limit, the messages of the SP belief propagation are approximated by applying the central limit theorem.

3. By taking β to infinity, the proper message passing rules are drawn.

4. The approximated mean of the SP belief of $\mathbf{x}$ is deduced via approximate message passing rules.

The AMP algorithm 1 is generalized then to broader probabilistic models with a denoiser $\mathbf{g}(\cdot;\cdot)$.

Algorithm 1 AMP

Require: Denoiser $\mathbf{g}(.;.)$, the transformation matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, measurements $\mathbf{y} \in \mathbb{R}^m$, and number of iterations K_{it}.

1: Initialization of r_0,$\mathbf{v}_{-1} = 0$, γ_0 and $\alpha_{-1} = 0$.
2: **for** k:=0 **to** K_{it} **do**
3: $\quad \widehat{\mathbf{x}}_k = \mathbf{g}\left(\mathbf{r}_k, \gamma_k\right)$.
4: $\quad \alpha_k = \langle \mathbf{g}'\left(\mathbf{r}_k, \gamma_k\right) \rangle$.
5: $\quad \mathbf{v}_k = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}_k + \frac{n}{m}\alpha_{k-1}\mathbf{v}_{k-1}$.
6: $\quad \mathbf{r}_{k+1} = \widehat{\mathbf{x}}_k + \mathbf{A}^\top \mathbf{v}_k$.
7: $\quad \gamma_{k+1}$ selection.
8: **end for**
9: **Return** $\widehat{\mathbf{x}}_{K_{it}}$.

2.4 SLM-VAMP

2.4.1 Motivating VAMP

What motivates VAMP is the limitations presented in the AMP algorithm. Moreover, AMP restricts only to a narrow class of random matrices $\mathbf{A}$ which are the identically independently distributed (i.i.d) subGaussians. Precisely, those having a non-zero mean or a mildly ill-conditioned matrix, i.e., having a high condition number, may cause the AMP to diverge. Hence, VAMP is considered as an alternative to AMP since applies to a wider class of sensing matrices $\mathbf{A}$ including ill-conditioned and non-zero mean ones. Furthermore, it presents a broader approximation of BP with a vectorized factor graph.

2.4.2 Algorithm derivation

Fundamentally, VAMP presents multi-various deductions leading to its derivation, similarly to AMP. The derivations of AMP and VAMP differ in the graphical model representation mentioned before. More specifically, VAMP relies on vector-valued nodes in a simple linear factor graph instead of a loopy graph as for the case of AMP. Primarily, the derivation of VAMP begins with the following joint distribution factorization.

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x}) \mathcal{N}\left(\mathbf{y}; \mathbf{A}\mathbf{x}, \gamma_w^{-1}\mathbf{I}\right), \tag{2.11}$$

Afterward, simple relaxations are used by splitting the variable $\mathbf{x}$ by using two new variables $\mathbf{x}_1 = \mathbf{x}_2$ that are identical to each other. Hence, the factorization in (2.11) is replaced by:

$$p\left(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2\right) = p\left(\mathbf{x}_1\right) \delta\left(\mathbf{x}_1 - \mathbf{x}_2\right) \mathcal{N}\left(\mathbf{y}; \mathbf{A}\mathbf{x}_2, \gamma_w^{-1}\mathbf{I}\right), \tag{2.12}$$

whose corresponding factor graph is illustrated in Fig 2.4.

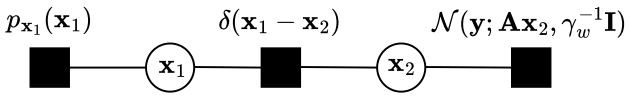


Fig. 2.4: Factor graph of SLM-VAMP [1].

Similarly to AMP, VAMP passes messages according to the following three rules:

- A variable node $\mathbf{x}$ follows the distribution determined by the approximate belief $b_{app}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \widehat{\mathbf{x}}, \eta^{-1}\mathbf{I}\right)$ where $\widehat{\mathbf{x}} = \mathbb{E}\left[\mathbf{x} \mid b_{\text{sp}}\right]$ and $\eta^{-1} = \left\langle \text{diag}\left(\text{Cov}\left[\mathbf{x} \mid b_{\text{sp}}\right]\right)\right\rangle$ are respectively the mean and average variance of the SP belief $b_{\text{sp}}(\mathbf{x}) \propto \prod_i \mu_{f_i \to \mathbf{x}}(\mathbf{x})$ namely the normalized form of all messages' product sent to the node as represented in Fig. 2.5 (a).

- A factor node f_i receives a message from a connected variable node $\mathbf{x}$ as follows:

$$\mu_{\mathbf{x}\to f_i}(\mathbf{x}) \propto \frac{b_{\text{app}}(\mathbf{x})}{\mu_{f_i\to\mathbf{x}}(\mathbf{x})}, \tag{2.13}$$

  Hence, this message is characterized by the filtration of the approximate belief through dividing it by the message recently sent contrariwise. The illustration in Fig. 2.5 (b) shows the passing messages in this case.

- A variable node $\mathbf{x}_i$ receives the following message from an adjacent factor node f:

$$\mu_{f\to\mathbf{x}_i}(\mathbf{x}_i) \propto \int f\left(\mathbf{x}_i, \{\mathbf{x}_j\}_{j\neq i}\right) \prod_{j\neq i} \mu_{\mathbf{x}_j\to f}(\mathbf{x}_j)\, \mathrm{d}\mathbf{x}_j. \tag{2.14}$$

  This equation is illustrated in the factor graph present in Fig. 2.5 (c).

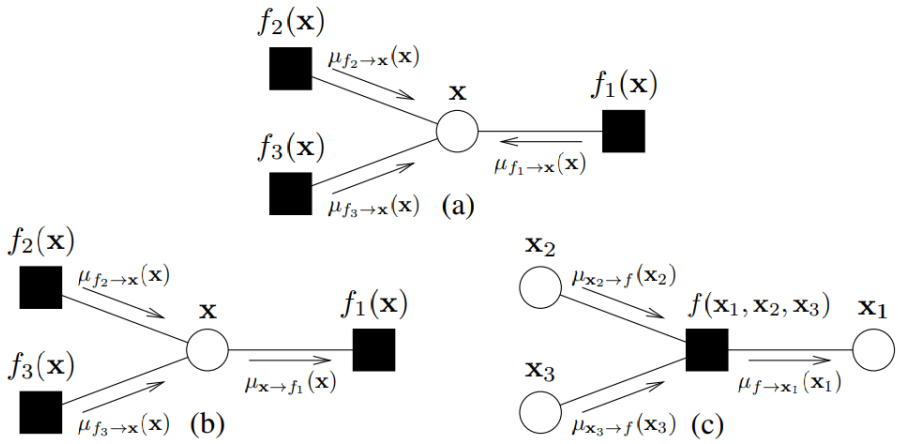


Fig. 2.5: VAMP update rules [1].

These rules manifest the essence of VAMP's derivation which will be detailed subsequently. As a starting point, let K_{it} and n denote the number of iterations and the dimension of the unknown vector $\mathbf{x}$. Afterward, the message $\mu_{\delta\to\mathbf{x}_1}(\mathbf{x}_1) = \mathcal{N}\left(\mathbf{x}_1; \mathbf{r}_{10}, \gamma_{10}^{-1}\mathbf{I}\right)$ is initialized by fixing $\mathbf{r}_{10}$ and γ_{10}. Eventually, the next steps are repeated for $k \in [0, K_{it}]$:

1. From the first update rule, the approximate belief $b_{app}(\mathbf{x}_1) = \mathcal{N}\left(\mathbf{x_1}; \widehat{\mathbf{x}}_1, \eta_1^{-1}\mathbf{I}\right)$ on x_1 is yielded by having its corresponding SP belief $b_{\text{sp}}(\mathbf{x}_1) \propto p\left(\mathbf{x}_1\right)\mathcal{N}\left(\mathbf{x}_1; \mathbf{r}_{1k}, \gamma_{1k}^{-1}\mathbf{I}\right)$ which conducts to the conditional estimator $\mathbf{g}_1\left(\mathbf{r}_{1k}, \gamma_{1k}\right) := \mathbb{E}\left[\mathbf{x}_1 \mid b_{\text{sp}}\left(\mathbf{x}_1\right)\right] = \widehat{\mathbf{x}}_{1k}$. Moreover, an i.i.d prior $p\left(\mathbf{x}_1\right)$, i.e., $p(\mathbf{x}) = \prod_{n=1}^{N} p\left(x_n\right)$, results in a scalar decomposition of this estimator $\left[\widehat{\mathbf{x}}_{1k}\right]_n = g_1\left(r_{1k,n}, \gamma_{1k}\right)$ which allows the following conditional variance calculation $\eta_{1k}^{-1} = \left\langle \text{diag}\left(\text{Cov}\left[\mathbf{x}_1 \mid b_{\text{sp}}\left(\mathbf{x}_1\right)\right]\right)\right\rangle = \gamma_{1k}^{-1} g_1'\left(r_{1k,n}, \gamma_{1k}\right)$.

2. From the second update rule, the message $\mu_{\mathbf{x}_1 \to \delta}\left(\mathbf{x}_1\right)$ is proportional to $\mathcal{N}\left(\mathbf{x_1}; \hat{\mathbf{x}}_1, \eta_{1k}^{-1}\mathbf{I}\right) / \mathcal{N}\left(\mathbf{x_1}; \mathbf{r}_{1k}, \gamma_{1k}^{-1}\mathbf{I}\right) \propto \mathcal{N}\left(\mathbf{x}; \left(\hat{\mathbf{x}}_{1k}\eta_{1k} - \mathbf{r}_{1k}\gamma_{1k}\right)/\left(\eta_{1k} - \gamma_{1k}\right), \left(\eta_{1k} - \gamma_{1k}\right)^{-1}\mathbf{I}\right) \propto \mathcal{N}\left(\mathbf{x}_1; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I}\right)$ where, $\gamma_{2k} = \eta_{1k} - \gamma_{1k}$ and $\mathbf{r}_{2k} = \left(\hat{\mathbf{x}}_{1k}\eta_{1k} - \mathbf{r}_{1k}\gamma_{1k}\right)/\left(\eta_{1k} - \gamma_{1k}\right)$. Then, using the third update rule, this message flows unchanged through the factor node δ which yields the message $\mu_{\delta \to \mathbf{x}_2}\left(\mathbf{x}_2\right) = \mathcal{N}\left(\mathbf{x}_2; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I}\right)$ received at the variable node $\mathbf{x_2}$ from the latter factor node.

3. Once again from the first update rule, the approximate belief $b_{app}(x_2) = \mathcal{N}\left(\mathbf{x_2}; \widehat{\mathbf{x}}_2, \eta_2^{-1}\mathbf{I}\right)$ on x_2 is yielded by having its corresponding SP belief $b_{\text{sp}}(\mathbf{x}_2) \propto \mathcal{N}\left(\mathbf{x}_2; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{y}; \mathbf{A}\mathbf{x}_2, \gamma_w^{-1}\mathbf{I}\right)$ which also conducts to the second conditional estimator $\mathbf{g}_2\left(\mathbf{r}_{2k}, \gamma_{2k}\right) := \mathbb{E}\left[\mathbf{x}_2 \mid b_{\text{sp}}\left(\mathbf{x}_2\right)\right] = \widehat{\mathbf{x}}_{2k}$. Further manipulations shows a more explicit Gaussian expression of the SP belief (2.15):

$$b_{sp}(\mathbf{x}_2) \propto \mathcal{N}\left(\mathbf{x}_2; \left(\gamma_w\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_w\mathbf{A}^\top\mathbf{y} + \gamma_{2k}\mathbf{r}_{2k}\right), \left(\gamma_w\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\right), \tag{2.15}$$

This yields the determination of the mean and average variance of the SP belief :

$$\begin{aligned} \widehat{\mathbf{x}}_{2k} &= \left(\gamma_w\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_w\mathbf{A}^\top\mathbf{y} + \gamma_{2k}\mathbf{r}_{2k}\right) = \mathbf{g}_2\left(\mathbf{r}_{2k}, \gamma_{2k}\right), \\ \eta_{2k}^{-1} &= \left\langle \text{diag}\left(\left(\gamma_w\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\right)\right\rangle = \gamma_{2k}^{-1}\left\langle \text{diag}\left(\mathbf{g}_2'\left(\mathbf{r}_{2k}, \gamma_{2k}\right)\right)\right\rangle. \end{aligned} \tag{2.16}$$

We notice that the estimator $\mathbf{g}_2$, which coincides with the MMSE and MAP estimators of the random vector $\mathbf{x}_2$ under its Gaussian prior and likelihood; $\mathcal{N}\left(\mathbf{x}_2; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I}\right)$ and $\mathcal{N}\left(\mathbf{y}; \mathbf{A}\mathbf{x}_2, \gamma_w^{-1}\mathbf{I}\right)$ respectively, is linear on $\mathbf{r}_{2k}$. Hence, it is named the LMMSE estimator.

4. Likewise, to the second step, the successive application of the second and third update rules yields the message which originally comes from the variable node $\mathbf{x}_2$ and passes through the factor node δ as unchanged , then, arrives to the variable node $\mathbf{x}_1$.

$$\mu_{\delta\to\mathbf{x}_1}\left(\mathbf{x}_1\right) = \mathcal{N}\left(\mathbf{x}; \left(\hat{\mathbf{x}}_{2k}\eta_{2k} - \mathbf{r}_{2k}\gamma_{2k}\right)/(\eta_{2k}-\gamma_{2k}), (\eta_{2k}-\gamma_{2k})^{-1}\mathbf{I}\right) \propto \mu_{\mathbf{x}_2\to\delta}\left(\mathbf{x}_2\right), \tag{2.17}$$

We define then $\mathbf{r}_{1k+1} = \left(\hat{\mathbf{x}}_{2k}\eta_{2k} - \mathbf{r}_{2k}\gamma_{2k}\right)/(\eta_{2k}-\gamma_{2k})$ and $\gamma_{1k+1} = \eta_{2k} - \gamma_{2k}$.

The above sequence of messaging between nodes is repeated for the K_{it} iterations which conducts to the VAMP Algorithm 2. In this approach, we refer to $\mathbf{r}_1$ and $\mathbf{r}_2$ as the pseudo-measurement and pseudo-prior of $\mathbf{x}$, since, the approximate beliefs $\mathbf{x} \sim \mathbf{r}_{1k} + \mathcal{N}\left(\mathbf{0}, {\gamma_{1k}}^{-1}\mathbf{I}\right)$ and $\mathbf{x} \sim \mathbf{r}_{2k} + \mathcal{N}\left(\mathbf{0}, {\gamma_{2k}}^{-1}\mathbf{I}\right)$ in the LMMSE estimation and denoising steps respectively.

Algorithm 2 SLM VAMP

Require: Denoiser $\mathbf{g}_1(\cdot;\cdot)$, LMMSE estimator $\mathbf{g}_2(\cdot;\cdot)$, and number of iterations K_{it}

Initialization of $\mathbf{r}_{10}$ and γ_{10}.
for k:=0 **to** K_{it} **do**
#Denoising
  $\widehat{\mathbf{x}}_{1k} = \mathbf{g}_1\left(\mathbf{r}_{1k}, \gamma_{1k}\right)$.
  $\alpha_{1k} = \left\langle \mathbf{g}_1'\left(\mathbf{r}_{1k}, \gamma_{1k}\right)\right\rangle$.
  $\eta_{1k} = \gamma_{1k}/\alpha_{1k}$.
  $\gamma_{2k} = \eta_{1k} - \gamma_{1k}$.
  $\mathbf{r}_{2k} = \left(\eta_{1k}\widehat{\mathbf{x}}_{1k} - \gamma_{1k}\mathbf{r}_{1k}\right)/\gamma_{2k}$.
#LMMSE estimation
  $\widehat{\mathbf{x}}_{2k} = \mathbf{g}_2\left(\mathbf{r}_{2k}, \gamma_{2k}\right)$.
  $\alpha_{2k} = \left\langle \mathbf{g}_2'\left(\mathbf{r}_{2k}, \gamma_{2k}\right)\right\rangle$.
  $\eta_{2k} = \gamma_{2k}/\alpha_{2k}$.
  $\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$.
  $\mathbf{r}_{1,k+1} = \left(\eta_{2k}\widehat{\mathbf{x}}_{2k} - \gamma_{2k}\mathbf{r}_{2k}\right)/\gamma_{1,k+1}$.
end for
Return $\widehat{\mathbf{x}}_{1K_{it}}$.

Finally, we can perceive the symmetry displayed in the algorithm, which motivates graphical illustration, namely, the block diagram representation.

2.4.3 Block diagram of VAMP

As described previously, the algorithm presents symmetry in its message passing and message filtering which allows introducing the new block diagram representation in Fig. 2.6.

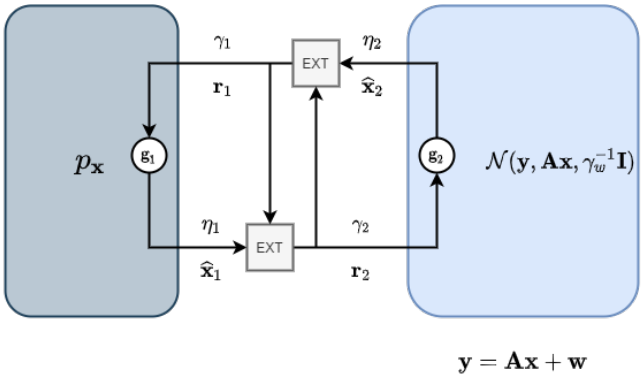


Fig. 2.6: Block diagram of SLM-VAMP.

First, we highlight the extrinsic message notation, which enables a better understanding of the block diagram depicted in Fig. 2.6. An extrinsic message is the exterior set of information received at a specific factor node after removing the information sent by the last-mentioned node. Hence, "EXT" blocks refer to the presence of the $\delta(\cdot)$ factor node. We observe also the loopy message passing aspect in the block diagram, which provides a more considerable algorithm design. Thus, we will capitalize on this representation over the graphs embodiment as it exhibits a straightforward application of the algorithm procedure.

2.5 GLM-VAMP

2.5.1 Motivating GLM

The generalized linear model (GLM), in which we ought to observe a random $\mathbf{x}$ through a noisy nonlinear function $\phi(\cdot)$ or a non-Gaussian posterior function $p_{y|z}(\cdot)$ of the linear transform $\mathbf{z} = \mathbf{Ax}$ in the measurement channel, as shown in Fig. 2.7, issues various purposes in different fields such as statistical regression, image processing, and communications. In the latter, we mention many applications, such as the binary linear

classification, which solves the problem $y_n = \text{sign}(z_n + w_n)$. Moreover, we mention the quantization-based compression, using a general scalar quantizer $Q(\cdot)$. Finally, other applications, such as robust regression, photon imaging, and phase retrieval, need to be mentioned.

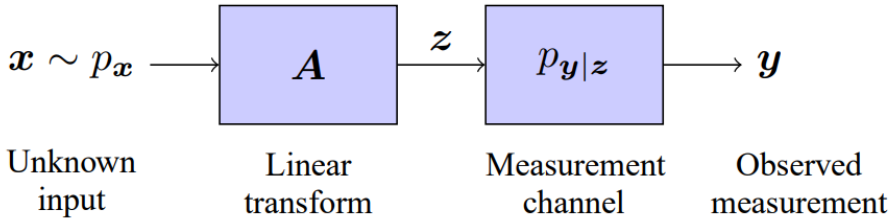


Fig. 2.7: Genrelized Linear Model (GLM).

2.5.2 GLM-VAMP derivation as an extension to SLM-VAMP

In this section, we briefly review the derivation of GLM-VAMP by introducing slight modifications to SLM-VAMP. Indeed, after splitting the two variables $\mathbf{x}$ and $\mathbf{z}$, we obtain the following joint-distribution factorization (3.37), which is also depicted graphically the Fig. 2.8:

$$p\left(\mathbf{y}, \mathbf{z_1}, \mathbf{z_2}, \mathbf{x}_1, \mathbf{x}_2\right) = p\left(\mathbf{x}_1\right) \delta\left(\mathbf{x}_1 - \mathbf{x}_2\right) \delta\left(\mathbf{z_1} - \mathbf{A}\mathbf{x}_2\right) \delta\left(\mathbf{z}_1 - \mathbf{z}_2\right) p\left(\mathbf{y}|\mathbf{z}_2; \gamma_w^{-1}\right). \quad (2.18)$$

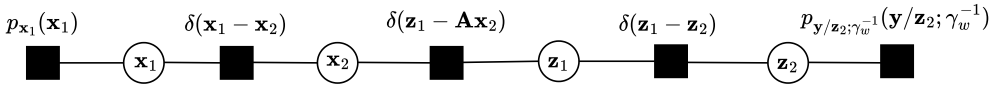


Fig. 2.8: Factor graph of GLM-VAMP.

At both ends of the factor graph in Fig. 2.8, the message-passing rules stand the same as in the standard linear model, which implies the direct construction of the denoising functions $\mathbf{g}_{\mathbf{x}_1}(\cdot)$ and $\mathbf{g}_{\mathbf{z}_2}(\cdot)$ which are also separable. More explicitly, the denoiser $\mathbf{g}_{\mathbf{x}_1}(\cdot)$

is identical to the one in the previous section; nevertheless, the finding of $\mathbf{g}_{\mathbf{z}_2}(\cdot)$ is straightforward after replacing $p(\mathbf{x}_1)$ by $p(\mathbf{y}|\mathbf{z}_2;\gamma_w^{-1})$. Hence, we refer to $\mathbf{p}_{2k}$ and τ_{2k} by the pseudo-measurement and its related precision, which reflect of the SP means and variances in the following update equations:

$$\begin{aligned}
\widehat{\mathbf{x}}_{1k} &= \mathbf{g}_{\mathbf{x}_1}(\mathbf{r}_{1k},\gamma_{1k}), \\
\eta_{1k}^{-1} &= \gamma_{1k}^{-1}\left\langle \operatorname{diag}\left(\mathbf{g}'_{\mathbf{x}_1}(\mathbf{r}_{1k},\gamma_{1k})\right)\right\rangle, \\
\widehat{\mathbf{z}}_{2k} &= \mathbf{g}_{\mathbf{z}_2}(\mathbf{p}_{2k},\tau_{2k}), \\
\xi_{2k}^{-1} &= \tau_{2k}^{-1}\left\langle \operatorname{diag}\left(\mathbf{g}'_{\mathbf{z}_2}(\mathbf{p}_{2k},\tau_{2k})\right)\right\rangle.
\end{aligned} \tag{2.19}$$

As for the LMMSE estimators, we exploit the equivalence relationship:

$$\mathbf{z}=\mathbf{A}\mathbf{x} \quad \Leftrightarrow \quad \mathbf{0}=[\mathbf{A}-\mathbf{I}]\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} \Leftrightarrow \quad \bar{\mathbf{y}}=\bar{\mathbf{A}}\bar{\mathbf{x}}+\bar{\mathbf{w}}, \tag{2.20}$$

where $\bar{\mathbf{y}}=\mathbf{0}$, $\bar{\mathbf{A}}=[\mathbf{A}-\mathbf{I}]$, $\bar{\mathbf{x}}=\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix}$, and $\bar{\mathbf{w}}\sim\mathcal{N}(\mathbf{0},\gamma_x^{-1}\mathbf{I})$ as $\gamma_x\to\infty$.

Subsequently, this modification throws us back to the SLM-VAMP derivation; then, similarly to the previous section and by using the message passing rules, we draw the pseudo-prior of $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}}=\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\mathbf{r}_{2k}\\ \mathbf{p}_{1k}\end{bmatrix},\begin{bmatrix}\gamma_{2k}^{-1}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \tau_{1k}^{-1}\mathbf{I}\end{bmatrix}\right). \tag{2.21}$$

Furthermore, as mentioned before, the LMMSE and the MAP estimates coincide because the prior $p_{\bar{\mathbf{x}}}$ and the likelihood $p_{\bar{\mathbf{y}}|\bar{\mathbf{x}}}$ are Gaussian. This yields to the following optimization problem within the framework of what is known as *maximum a posteriori*

(MAP estimation):

$$\begin{aligned}\arg\max_{\bar{\mathbf{x}}} p(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}) &= \arg\min_{\bar{\mathbf{x}}} \{-\ln p(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}) - \ln p(\bar{\mathbf{x}})\} \\ &= \arg\min_{\mathbf{x},\mathbf{z}} \{\gamma_x \|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2^2 + \gamma_{2k} \|\mathbf{r}_{2k} - \mathbf{x}\|_2^2 + \tau_{1k} \|\mathbf{p}_{1k} - \mathbf{z}\|_2^2\}.\end{aligned} \tag{2.22}$$

Eventually, we zero the gradients of the expression (2.22) w.r.t $\mathbf{x}$ and $\mathbf{z}$ at $\widehat{\mathbf{x}}_{2k}$ and $\widehat{\mathbf{z}}_{1k}$, which yields:

$$\begin{aligned}&\gamma_x \mathbf{A}^\top \left(\mathbf{A}\widehat{\mathbf{x}}_{2k} - \widehat{\mathbf{z}}_{1k}\right) + \gamma_{2k} \left(\widehat{\mathbf{x}}_{2k} - \mathbf{r}_{2k}\right) = \mathbf{0}, \\ &\gamma_x \left(\widehat{\mathbf{z}}_{1k} - \mathbf{A}\widehat{\mathbf{x}}_{2k}\right) + \tau_{1k} \left(\widehat{\mathbf{z}}_{1k} - \mathbf{p}_{1k}\right) = \mathbf{0}.\end{aligned} \tag{2.23}$$

By a simple reformulation, these two identities can be represented in the matrix/vector notation as follows:

$$\begin{bmatrix} \gamma_{2k}\mathbf{r}_{2k} \\ \tau_{1k}\mathbf{p}_{1k} \end{bmatrix} = \begin{bmatrix} \gamma_x \mathbf{A}^\top \mathbf{A} + \gamma_{2k}\mathbf{I} & -\gamma_x \mathbf{A}^\top \\ -\gamma_x \mathbf{A} & (\tau_{1k} + \gamma_x)\,\mathbf{I} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{x}}_{2k} \\ \hat{\mathbf{z}}_{1k} \end{bmatrix}, \tag{2.24}$$

Now, using the Schur's complement (See Appendix A) in (2.24), we find the inverse of the involved block matrices as follows:

$$\mathbf{Q}_k \triangleq \gamma_x \mathbf{A}^\top \mathbf{A} + \gamma_{2k}\mathbf{I} - \frac{\gamma_x^2}{\tau_{1k} + \gamma_x} \mathbf{A}^\top \mathbf{A} = \frac{\gamma_x \tau_{1k}}{\tau_{1k} + \gamma_x} \mathbf{A}^\top \mathbf{A} + \gamma_{2k}\mathbf{I}, \tag{2.25}$$

Thereby leading to:

$$\begin{bmatrix} \widehat{\mathbf{x}}_{2k} \\ \widehat{\mathbf{z}}_{1k} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_k^{-1} & \frac{\gamma_x}{\tau_{1k}+\gamma_x} \mathbf{Q}_k^{-1} \mathbf{A}^\top \\ \frac{\gamma_x}{\tau_{1k}+\gamma_x} \mathbf{A}\mathbf{Q}_k^{-1} & \frac{1}{\tau_{1k}+\gamma_x} \left(\mathbf{I} + \frac{\gamma_x^2}{\tau_{1k}+\gamma_x} \mathbf{A}\mathbf{Q}_k^{-1}\mathbf{A}^\top\right) \end{bmatrix} \begin{bmatrix} \gamma_{2k}\mathbf{r}_{2k} \\ \tau_{1k}\mathbf{p}_{1k} \end{bmatrix}, \tag{2.26}$$

By taking $\gamma_x \to \infty$ which gives $\mathbf{Q}_k = \tau_{1k}\mathbf{A}^\top \mathbf{A} + \gamma_{2k}\mathbf{I}$, one obtains:

$$\begin{aligned}\begin{bmatrix} \widehat{\mathbf{x}}_{2k} \\ \widehat{\mathbf{z}}_{1k} \end{bmatrix} &= \begin{bmatrix} \mathbf{Q}_k^{-1} & \mathbf{Q}_k^{-1}\mathbf{A}^\top \\ \mathbf{A}\mathbf{Q}_k^{-1} & \mathbf{A}\mathbf{Q}_k^{-1}\mathbf{A}^\top \end{bmatrix} \begin{bmatrix} \gamma_{2k}\mathbf{r}_{2k} \\ \tau_{1k}\mathbf{p}_{1k} \end{bmatrix}, \\ &= \begin{bmatrix} \mathbf{I} \\ \mathbf{A} \end{bmatrix} \left(\tau_{1k}\mathbf{A}^\top \mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1} \left(\gamma_{2k}\mathbf{r}_{2k} + \tau_{1k}\mathbf{A}^\top \mathbf{p}_{1k}\right).\end{aligned} \tag{2.27}$$

Accordingly, we draw the means and variances of the SP belief via the LMMSE estimators, which obtained as follows:

$$\begin{aligned}
\widehat{\mathbf{x}}_{2k} &= \mathbf{g}_{\mathbf{x}_2}\left(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}\right) = \left(\tau_{1k}\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_{2k}\mathbf{r}_{2k} + \tau_{1k}\mathbf{A}^\top\mathbf{p}_{1k}\right), \\
\eta_{2k}^{-1} &= \gamma_{2k}^{-1}\left\langle \operatorname{diag}\left(\frac{\partial \mathbf{g}_{\mathbf{x}_2}}{\partial \mathbf{r}_{2k}}\left(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}\right)\right)\right\rangle = \left\langle \operatorname{diag}\left(\left(\tau_{1k}\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\right)\right\rangle, \\
\widehat{\mathbf{z}}_{1k} &= \mathbf{g}_{\mathbf{z}_1}\left(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}\right) = \mathbf{A}.\mathbf{g}_{\mathbf{x}_2}\left(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}\right), \\
\xi_{1k}^{-1} &= \tau_{1k}^{-1}\left\langle \operatorname{diag}\left(\frac{\partial \mathbf{g}_{\mathbf{z}_1}}{\partial \mathbf{p}_{1k}}\left(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}\right)\right)\right\rangle = \left\langle \operatorname{diag}\left(\mathbf{A}\left(\tau_{1k}\mathbf{A}^\top\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1}\mathbf{A}^\top\right)\right\rangle.
\end{aligned} \tag{2.28}$$

Consequently, the whole procedure of GLM-VAMP (summarized in Algorithm 3) can be illustrated through the block diagram of Fig 2.9.

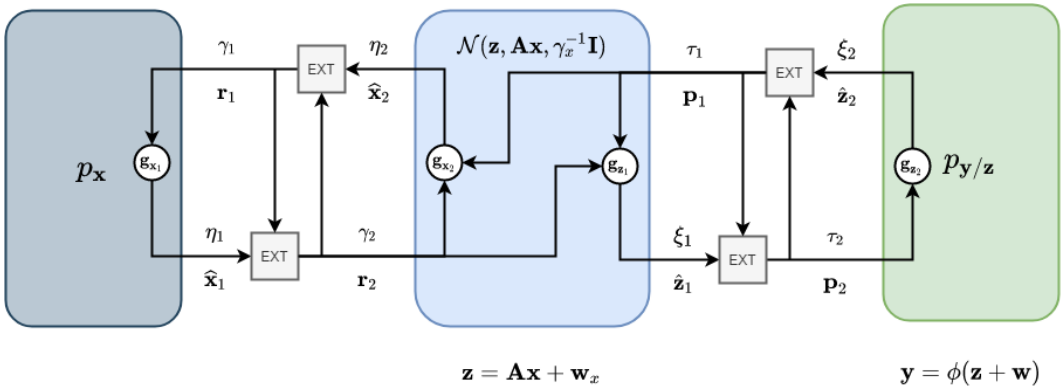


Fig. 2.9: Block diagram of GLM-VAMP.

Algorithm 3 GLM-VAMP

Require: Denoisers $\mathbf{g}_{\mathbf{x}_1}(.;.)$ and $\mathbf{g}_{\mathbf{z}_2}(.;.)$, LMMSE estimators $\mathbf{g}_{\mathbf{x}_2}(.;.)$ and $\mathbf{g}_{\mathbf{z}_1}(.;.)$, and number of iterations K_{it}.

Initialization of $\mathbf{r}_{10}$, $\mathbf{p}_{20}$,γ_{10} and τ_{20}.

for k:=0 **to** K_{it} **do**

#Denoising x

$\widehat{\mathbf{x}}_{1k} = \mathbf{g}_{\mathbf{x}_1}(\mathbf{r}_{1k}, \gamma_{1k})$.

$\alpha_{1k} = \left\langle \mathbf{g}'_{\mathbf{x}_1}(\mathbf{r}_{1k}, \gamma_{1k}) \right\rangle$.

$\eta_{1k} = \gamma_{1k}/\alpha_{1k}$.

$\gamma_{2k} = \eta_{1k} - \gamma_{1k}$.

$\mathbf{r}_{2k} = (\eta_{1k}\widehat{\mathbf{x}}_{1k} - \gamma_{1k}\mathbf{r}_{1k})/\gamma_{2k}$.

#Denoising z

$\widehat{\mathbf{z}}_{2k} = \mathbf{g}_{\mathbf{z}_2}(\mathbf{p}_{2k}, \tau_{2k})$.

$\beta_{2k} = \left\langle \mathbf{g}'_{\mathbf{z}_2}(\mathbf{p}_{2k}, \tau_{2k}) \right\rangle$.

$\xi_{2k} = \tau_{2k}/\beta_{2k}$.

$\tau_{1k} = \xi_{1k} - \gamma_{1k}$.

$\mathbf{p}_{1k} = (\xi_{2k}\widehat{\mathbf{z}}_{2k} - \tau_{2k}\mathbf{p}_{2k})/\tau_{1k}$.

#LMMSE estimation of x

$\widehat{\mathbf{x}}_{2k} = \mathbf{g}_{\mathbf{x}_2}(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k})$.

$\alpha_{2k} = \left\langle \mathbf{g}'_{\mathbf{x}_2}(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}) \right\rangle$.

$\eta_{2k} = \gamma_{2k}/\alpha_{2k}$.

$\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$.

$\mathbf{r}_{1,k+1} = (\eta_{2k}\widehat{\mathbf{x}}_{2k} - \gamma_{2k}\mathbf{r}_{2k})/\gamma_{1,k+1}$.

#LMMSE estimation of z

$\widehat{\mathbf{z}}_{1k} = \mathbf{g}_{\mathbf{z}_1}(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k})$.

$\beta_{1k} = \left\langle \mathbf{g}'_{\mathbf{z}_1}(\mathbf{r}_{2k}, \mathbf{p}_{1k}, \gamma_{2k}, \tau_{1k}) \right\rangle$.

$\xi_{1k} = \tau_{1k}/\beta_{1k}$.

$\tau_{2,k+1} = \xi_{1k} - \tau_{1k}$.

$\mathbf{p}_{2,k+1} = (\xi_{1k}\widehat{\mathbf{z}}_{1k} - \tau_{1k}\mathbf{p}_{1k})/\tau_{2,k+1}$.

end for

Return $\widehat{\mathbf{x}}_{1K_{it}}$.

Chapter 3

Multi-user Detection using VAMP

This chapter presents the system model for a ULA configuration in a MU setting. This model addresses the detection problem for general B-bits ADCs, assuming perfect channel state information. We introduce a ray-based multi-path model to account for channel and noise correlation effects. Subsequently, we introduce a detection algorithm based on VAMP, specifically tailored for low-resolution quantized systems. This algorithm is designed to effectively manage noise correlation, which is especially crucial in oversampling scenarios. We also conduct a large dimension analysis to investigate the algorithm's asymptotic behaviour through SE. Finally, we assess and compare the performance of our algorithm with existing theoretical findings, providing insights into its effectiveness and practical applicability.

3.1 System Model

Consider a BS with N antennas, each equipped by a B-bits ADC, arranged in a ULA configuration with an inter-element spacing d and fixed aperture length D. The BS

serves a network composed of K users. In the uplink scenario, the users transmit symbols denoted as $\mathbf{x} \in \mathcal{C}^{K\times 1}$, wherein $\mathcal{C}$ signifies the set of constellation points characterizing the modulation scheme employed by the transmitters, which is assumed uniform with zero mean and unit energy. The transmitted signals traverse a communication channel that is both noisy and correlated. This channel's characteristics are contingent upon the antenna pattern and are encapsulated by the matrix $\mathbf{H} \in \mathbb{C}^{N\times K}$. On the receiver side, the analog received signal, subsequent to spatial sampling, is thus given by $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w}$, wherein, $\mathbf{w} \in \mathbb{C}^{N\times 1}$, depending on the spacing d, is a complex correlated noise which follows the circularly-symmetric complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \mathbf{R}_{\text{noise}})$. Additionally, the analog signal is digitized using ADCs before being sent to the digital signal processing unit for signal recovery. As a continuation, assuming perfect CSI, the task of the BS is to detect users' signals under the hypothesis of given quantized measurements, denoted by $\mathbf{y}$, whereby

$$\mathbf{y} = \mathfrak{Q}_B(\mathbf{z}), \tag{3.1}$$

Herein, $\mathfrak{Q}_B(\cdot)$ is an element-wise complex-valued quantizer defined for any complex number $z \in \mathbb{C}$ by $\mathfrak{Q}_B(z) \triangleq \mathcal{Q}_B(\Re(z)) + j\ \mathcal{Q}_B(\Im(z))\ )$. The real-valued quantizer, denoted as $\mathcal{Q}_B(\cdot)$, is a mapping operator that discretizes a continuous real input into one of 2^B discrete intervals. These intervals are characterized by a collection of 2^B-1 distinct thresholds, namely $\{r_1, r_2, ...,r_{2^B-1}\}$, arranged in a manner such that $-\infty < r_1 < r_2 < ... < r_{2^B-1} < +\infty$. To simplify notation, we augment the set of thresholds with the conventional values, $r_0 = -\infty$ and $r_{2^B} = +\infty$. In the case of a uniform quantizer, the thresholds are defined by the step size Δ and are given by for $b = 1, ..., 2^B - 1$: $r_b = (-2^{B-1} + b)\Delta$. Henceforth, the quantization input encompassed within the b-th bin, denoted by the interval $A_b = (r_{b-1}, r_b]$, is assigned the discrete value $s_b = r_b - \frac{\Delta}{2}$, where b ranges from 1 to $2^B - 1$. Furthermore, for quantization inputs falling within the interval $A_{2^B} = (r_{2^B-1}, r_{2^B})$, the corresponding quantized value is designated as

$s_{2^B} = r_{2^B-1} + \frac{\Delta}{2}$. As a result, the quantization function is explicitly given by for any real value $\mu \in \mathbb{R} = \bigsqcup_{b=1}^{2^B} A_b$:

$$\mathcal{Q}_B(\mu) = \sum_{b=1}^{2^B} s_b \cdot \chi_{A_b}(\mu), \tag{3.2}$$

where for any arbitrary subset $A \subseteq \mathbb{R}$, $\chi_A(.)$ denotes the indicator function. Fig. 3.1 illustrates an exemplar of the 3-bit uniform quantizer. The performance of a quantization

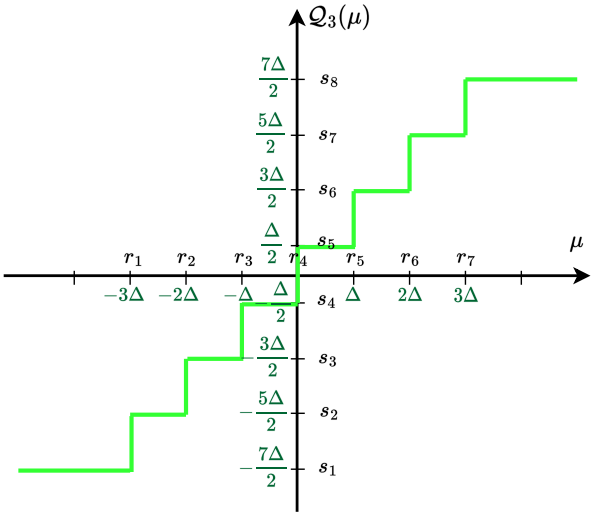


Fig. 3.1: Example of 3-bit quantizer.

system is typically influenced by the choice of step size Δ, which is crucial for adequately covering the support of the received analog signal. However, the step size becomes inconsequential when dealing with a 1-bit quantizer. This phenomenon is evident through the independence of the thresholds, defined as $r_0 = -\infty, r_1 = 0, r_2 = +\infty$. This underscores the unique behaviour of a 1-bit quantizer in contrast to quantizers with larger bit depths, highlighting its unconventional properties in signal processing. Hence, by choosing $\Delta = 2$, the one-bit quantization function is thereby simplified to the two-level

sign function:

$$\mathcal{Q}_1(\mu) = \text{sign}(\mu) := \begin{cases} -1 & \text{if } \mu \leq 0, \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

Nevertheless, in the hypothetical case of infinite resolution ADCs with infinitesimal step size, we achieve the perfect measure of the arriving analog noisy signals thereby the signal recovery becomes a detection problem under the hypothesis of given linear measurements $\mathbf{y}$:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}. \tag{3.4}$$

In the sequel, we will present the ray-based models of the channel and noise that we used in this work.

3.1.1 Channel Model

This work operates on the premise that all wave sources adhere to the far-field approximation. This implies that incident waves upon the linear array plane are construed as planar waves, characterized by elevation angle θ and azimuth angle φ. Subsequently, for a ULA with with the previously stated configuration and as depicted in Fig. 3.2, we proceed to define the far-field linear array steering vector of an incident wave, which encapsulates the wave's arrival direction and the corresponding set of phase delays it undergoes:

$$\mathbf{a}(\theta, \lambda, d) = \begin{bmatrix} 1 \\ e^{-2\pi j \frac{d}{\lambda} cos\theta} \\ \vdots \\ e^{-2\pi j \frac{d}{\lambda}(N-1) cos\theta} \end{bmatrix}. \tag{3.5}$$

with N number of antennas, inter-element spacing d and a wavelength λ.

We consider a multipath channel model in which the individual narrowband channels

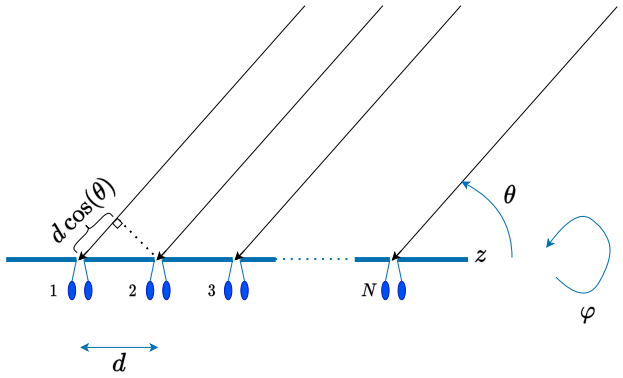


Fig. 3.2: ULA incident wave with an arriving angle θ.

for each user are essentially a superposition of L multipath components, primarily comprised of scatterers. At the ULA end, each antenna element is defined by an aperture denoted by the effective area $A_{\text{eff}}(\theta, d, \lambda)$. For the k^{th} user $(k = 1, .., K)$, the individual channel can be expressed as:

$$\mathbf{h}_k = \frac{1}{R} \sum_{\ell=1}^{L} c_{k,\ell} \cdot \sqrt{\frac{P_t}{4\pi B_W}} \sqrt{A_{\text{eff}}\left(\theta_{k,\ell}, \lambda, d\right)} \mathbf{a}\left(\theta_{k,\ell}, \lambda, d\right), \tag{3.6}$$

here, $c_{k,\ell} \sim \mathcal{CN}(0, 1)$ represents the path-loss and phase-shift weight of the ℓ^{th} incident ray coming from the k^{th} user, R denotes the distance between the transmitter and receiver, P_t stands for the transmission power, B_W represents the bandwidth, and $\theta_{k,\ell}$ signifies the incident angle. Subsequently, we define the pseudo-SNR (obtained with a single isotropic receiving antenna) as the ratio of the carrier received-signal power $P_r = P_t \frac{\lambda^2}{(4\pi R)^2}$ to the noise power $P_N = k_B T B_W$:

$$\text{SNR} = \frac{P_r}{P_N} = \frac{P_t}{k_B T B_W} \frac{\lambda^2}{(4\pi R)^2}. \tag{3.7}$$

Therefore, the variable we will employ to modulate the SNR is the radius R.

3.1.2 Noise Model

In the domain of extremely high frequencies, such as the millimetre wave (mmWave) spectrum, thermal noise emerges as a primary limiting factor that significantly hinders the process of signal recovery. Noteworthy, this thermal noise arises from environmental sources, classifying it as an extrinsic noise. Consequently, the present study deliberately confines its scope by disregarding alternative noise categories. Subsequently, our investigation delves into the formulation of a thermal noise model. This model finds its origins in the principles of thermal radiation density in Equation (3.8), as elucidated by Max Planck's law.

$$B_{\mathrm{f}}(\lambda, T) = \frac{hc}{\lambda^3} \frac{1}{e^{hc/k_B T\lambda} - 1} \simeq \frac{k_B T}{\lambda^2}, \quad \text{for} \quad \frac{hc}{\lambda} \ll k_B T, \tag{3.8}$$

where h is the Planck's constant, c is the speed of light, k_B is the Boltzmann's constant, T is the environmental temperature and λ is the wavelength. Notably, this model considers an isotropic radiation pattern characterized by a singular polarization. The extrinsic spatial noise wave correlation matrix, represented by Equation (3.9), is subsequently established through integration across all possible angles of arrival, which are uniformly distributed, for the spatially auto-correlated power densities absorbed by the array elements.

$$\mathbf{R}_{\text{noise;extrinsic}}(\lambda, d) = B_{\mathrm{f}}(\lambda, T) \int_0^{\pi} \int_0^{2\pi} A_{\text{eff}}(\theta, d, \lambda)\, \mathbf{a}(\theta, \lambda, d) \mathbf{a}(\theta, \lambda, d)^{\mathrm{H}} \sin\theta \, d\varphi \, d\theta, \tag{3.9}$$

This equation elucidates the spatial coupling between antennas presented within the array configuration, which is influenced by the wave phase delays. Additionally, the signal may be degraded by an intrinsic noise, i.e., the noise figure, due to the low noise amplifiers implemented in the receiver. Thus, the total noise covariance matrix is yielded by:

$$\mathbf{R}_{\text{noise}} = \mathbf{R}_{\text{noise;extrinsic}} + k_B T (N_f - 1) \cdot \mathbf{I}, \tag{3.10}$$

with N_f the noise figure factor. In conclusion, the thermal noise model can be described as a circularly symmetric complex Gaussian noise with covariance matrix $\mathbf{R}_{\text{noise}}$, i.e.:

$$\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\text{noise}}). \tag{3.11}$$

3.2 Signal recovery

We utilize a Bayesian inference approach to address the detection problem under the hypothesis of quantized measurements given by Equation (3.2). This approach lets us obtain optimal mean squared error (MSE) estimations. However, employing Bayesian inference typically involves computationally intensive operations. Fortunately, the VAMP algorithm, originally proposed by S. Rangan et al., presents a computationally efficient technique to achieve Bayes-optimal performance. The VAMP algorithm has demonstrated excellent results in various contexts, including both standard and generalized linear models with Additive White Gaussian Noise in Compressed Sensing. In our work, we extend the capabilities of the VAMP algorithm to handle noise correlation in the complex domain. This adaptation applies to scenarios involving general B-bits resolution ADCs.

3.2.1 Signal recovery for B-bit ADCs

As with VAMP, the derivation of this adapted algorithm is based on an expectation propagation approach for approximate Bayesian Inference. We start from the joint-pdf factorization of the detection problem from Equation (3.1)

$$\begin{aligned} p_{\mathbf{y},\mathbf{z},\mathbf{x}}(\mathbf{y},\mathbf{z},\mathbf{x}|\mathbf{H}) &= p(\mathbf{x})\, p(\mathbf{z}|\mathbf{x};\mathbf{R}_{\text{noise}})\, p(\mathbf{y}|\mathbf{z}) \\ &= p(\mathbf{x})\, \mathcal{CN}(\mathbf{z};\mathbf{Hx},\mathbf{R}_{\text{noise}})\, p(\mathbf{y}|\mathbf{z}), \end{aligned} \tag{3.12}$$

In this context, we consider two fundamental probability distributions, denoted as $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{z})$. The first distribution, $p(\mathbf{x})$, represents the prior distribution of the vector

$\mathbf{x}$, while the second distribution, $p(\mathbf{y}|\mathbf{z})$, characterizes the likelihood distribution of $\mathbf{y} = \mathcal{Q}_B(\mathbf{z})$ conditioned on the value of $\mathbf{z}$. These two distribution densities are considered separable, both in the vector space and in the real and imaginary components of the complex space. The separability in the prior distribution arises from the assumption of user independence. In contrast, for the likelihood distribution, the separability is obtained through the conditional independence of the elements of $\mathbf{y}$ given $\mathbf{z}$. Thereafter, we split both variables $\mathbf{x}$ and $\mathbf{z}$ into two identical variables $\mathbf{x}^+ = \mathbf{x}^-$ and $\mathbf{z}^- = \mathbf{z}^+$ resulting in the following joint-pdf:

$$\begin{aligned} p\left(\mathbf{y}, \mathbf{z}^-, \mathbf{z}^+, \mathbf{x}^+, \mathbf{x}^-|\mathbf{H}\right) &= p\left(\mathbf{x}^+\right) p\left(\mathbf{x}^-|\mathbf{x}^+\right) p\left(\mathbf{z}^-|\mathbf{x}^-\right) p\left(\mathbf{z}^+|\mathbf{z}^-\right) p\left(\mathbf{y}|\mathbf{z}^+\right) \\ &= p\left(\mathbf{x}^+\right) \delta_{\mathbf{x}}\left(\mathbf{x}^+ - \mathbf{x}^-\right) \mathcal{CN}\left(\mathbf{z}^-; \mathbf{H}\mathbf{x}^-, \mathbf{R}_{\text{noise}}\right) \delta_{\mathbf{z}}\left(\mathbf{z}^- - \mathbf{z}^+\right) \times \\ &\qquad p\left(\mathbf{y}|\mathbf{z}^+\right), \end{aligned} \tag{3.13}$$

where $\delta_{\mathbf{x}}(\cdot)$ and $\delta_{\mathbf{z}}(\cdot)$ are Dirac delta probability density functions. Accordingly, as illustrated in the factor graph, we represent the factorization in Equation (3.13). As depicted in Fig. 3.3, VAMP relies on vector-valued nodes in an acyclic factor graph instead of a complete-bipartite graph, as for the case of AMP and MP algorithms. Moreover, each belief of a variable node or a message passed by a variable node is approximated by tractable multi-variant complex Gaussian distributions determined by a vector mean and an averaged scalar variance.

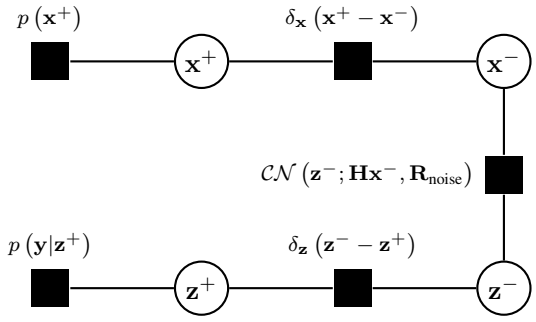


Fig. 3.3: Factor graph.

Consequently, when applying the VAMP message passing rules, one must obtain the Algorithm 4. To offer a more vivid illustration, Fig. 3.4 presents the block diagram of the Algorithm, providing a visual representation of its various constituent blocks. These blocks include the different denoisers, which interact with the LMMSE module. It is worth noting that the iteration number, denoted as t, is omitted in this block diagram. However, this diagram effectively captures the sequential nature of the algorithm's iterations.

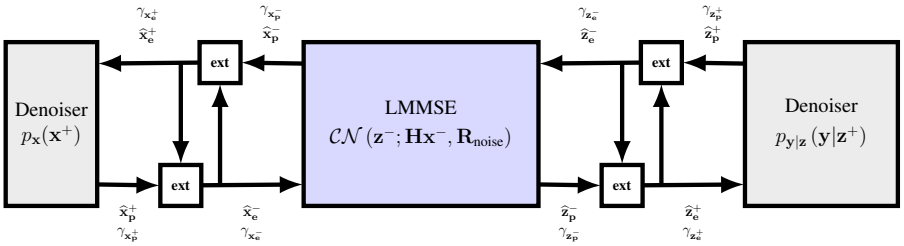


Fig. 3.4: Block diagram .

Denoising $\mathbf{x}^+$

The denoising of $\mathbf{x}^+$ starts by the received message coming from $\delta_{\mathbf{x}}(\cdot)$ factor node with mean $\widehat{\mathbf{x}}_{\mathbf{e}}^+$ and an averaged scalar variance $\gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}$ which are initialized at the start of the algorithm. The approximate belief of the variable node $\mathbf{x}^+$ is then reduced to a complex Gaussian function ${}^{\text{app}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^+}(\mathbf{x}^+) = \mathcal{CN}\left(\mathbf{x}^+; \widehat{\mathbf{x}}_{\mathbf{p}}^+, \gamma_{\mathbf{x}_{\mathbf{p}}^+}^{-1}\mathbf{I}\right)$ where $\widehat{\mathbf{x}}_{\mathbf{p}}^+$ is the estimate of $\mathbf{x}^+$ based on its SP belief and $\gamma_{\mathbf{x}_{\mathbf{p}}^+}^{-1}$ is the estimate averaged error.

$$\begin{aligned} \widehat{\mathbf{x}}_{\mathbf{p}}^+ &= \mathbb{E}\left[\mathbf{x}^+ | {}^{\text{sp}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^+}(\mathbf{x}^+)\right] = \mathbf{g}_{\mathbf{x}^+}(\widehat{\mathbf{x}}_{\mathbf{e}}^+; \gamma_{\mathbf{x}_{\mathbf{e}}^+}) \\ \gamma_{\mathbf{x}_{\mathbf{p}}^+}^{-1} &= \tfrac{1}{K}\,\text{Tr}\left(\text{Cov}[\mathbf{x}^+ | {}^{\text{sp}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^+}(\mathbf{x}^+)]\right) \end{aligned} \tag{3.14}$$

herein

$${}^{\text{sp}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^+}(\mathbf{x}^+) = \frac{p_{\mathbf{x}^+}(\mathbf{x}^+)\mathcal{CN}\left(\mathbf{x}^+; \widehat{\mathbf{x}}_{\mathbf{e}}^+, \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}\mathbf{I}\right)}{\sum\limits_{\mathbf{x}^+ \in \mathcal{C}^K} p_{\mathbf{x}^+}(\mathbf{x}^+)\mathcal{CN}\left(\mathbf{x}^+; \widehat{\mathbf{x}}_{\mathbf{e}}^+, \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}\mathbf{I}\right)} \tag{3.15}$$

Algorithm 4 adapted VAMP for B-bits signal recovery

Require: Denoisers $\mathbf{g}_{\mathbf{x}^+}(\cdot;\cdot)$ and $\mathbf{g}_{\mathbf{z}^+}(\cdot;\cdot)$, LMMSE estimators $\mathbf{g}_{\mathbf{x}^-}(\cdot,\cdot;\cdot,\cdot)$ and $\mathbf{g}_{\mathbf{z}^-}(\cdot,\cdot;\cdot,\cdot)$, and maximum number of iterations T_{max}.

Initialization of $\widehat{\mathbf{x}}^+_{\mathbf{e},1}$, $\widehat{\mathbf{z}}^+_{\mathbf{e},1}$,$\gamma_{\mathbf{x}^+_{\mathbf{e}},1}$ and $\gamma_{\mathbf{z}^+_{\mathbf{e}},1}$.

For t:=1 **to** T_{max}

#Denoising x

$\widehat{\mathbf{x}}^+_{\mathbf{p},t} = \mathbf{g}_{\mathbf{x}^+}(\widehat{\mathbf{x}}^+_{\mathbf{e},t};\gamma_{\mathbf{x}^+_{\mathbf{e}},t})$.

$\gamma_{\mathbf{x}^+_{\mathbf{p}},t} = \frac{\gamma_{\mathbf{x}^+_{\mathbf{e}},t}}{\frac{1}{2K}\mathrm{Tr}\left(\frac{\partial\Re(\mathbf{g}_{\mathbf{x}^+})}{\partial\Re(\widehat{\mathbf{x}}^+_{\mathbf{e}})}\left(\Re(\widehat{\mathbf{x}}^+_{\mathbf{e},t});\gamma_{\mathbf{x}^+_{\mathbf{e}},t}\right)+\frac{\partial\Im(\mathbf{g}_{\mathbf{x}^+})}{\partial\Im(\widehat{\mathbf{x}}^+_{\mathbf{e}})}\left(\Im(\widehat{\mathbf{x}}^+_{\mathbf{e},t});\gamma_{\mathbf{x}^+_{\mathbf{e}},t}\right)\right)}$.

$\gamma_{\mathbf{x}^-_{\mathbf{e}},t} = \gamma_{\mathbf{x}^+_{\mathbf{p}},t} - \gamma_{\mathbf{x}^+_{\mathbf{e}},t}$.

$\widehat{\mathbf{x}}^-_{\mathbf{e},t} = \left(\gamma_{\mathbf{x}^+_{\mathbf{p}},t}\widehat{\mathbf{x}}^+_{\mathbf{p},t} - \gamma_{\mathbf{x}^+_{\mathbf{e}},t}\widehat{\mathbf{x}}^+_{\mathbf{e},t}\right)/\gamma_{\mathbf{x}^-_{\mathbf{e}},t}$.

#Denoising z

$\widehat{\mathbf{z}}^+_{\mathbf{p},t} = \mathbf{g}_{\mathbf{z}^+}(\widehat{\mathbf{z}}^+_{\mathbf{e},t},\mathbf{y};\gamma_{\mathbf{z}^+_{\mathbf{e}},t})$.

$\gamma_{\mathbf{z}^+_{\mathbf{p}},t} = \frac{\gamma_{\mathbf{z}^+_{\mathbf{e}},t}}{\frac{1}{2N}\mathrm{Tr}\left(\frac{\partial\Re(\mathbf{g}_{\mathbf{z}^+})}{\partial\Re(\widehat{\mathbf{z}}^+_{\mathbf{e}})}\left(\Re(\widehat{\mathbf{z}}^+_{\mathbf{e},t}),\Re(\mathbf{y});\gamma_{\mathbf{z}^+_{\mathbf{e}},t}\right)+\frac{\partial\Im(\mathbf{g}_{\mathbf{z}^+})}{\partial\Im(\widehat{\mathbf{z}}^+_{\mathbf{e}})}\left(\Im(\widehat{\mathbf{z}}^+_{\mathbf{e},t}),\Im(\mathbf{y});\gamma_{\mathbf{z}^+_{\mathbf{e}},t}\right)\right)}$.

$\gamma_{\mathbf{z}^-_{\mathbf{e}},t} = \gamma_{\mathbf{z}^+_{\mathbf{p}},t} - \gamma_{\mathbf{z}^+_{\mathbf{e}},t}$.

$\widehat{\mathbf{z}}^-_{\mathbf{e},t} = \left(\gamma_{\mathbf{z}^+_{\mathbf{p}},t}\widehat{\mathbf{z}}^+_{\mathbf{p},t} - \gamma_{\mathbf{z}^+_{\mathbf{e}},t}\widehat{\mathbf{z}}^+_{\mathbf{e},t}\right)/\gamma_{\mathbf{z}^-_{\mathbf{e}},t}$.

#LMMSE estimation of x

$\widehat{\mathbf{x}}^-_{\mathbf{p},t} = \left(\mathbf{H}^H\left(\mathbf{R}_{\text{noise}}+\gamma^{-1}_{\mathbf{z}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\mathbf{H}+\gamma_{\mathbf{x}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\left(\mathbf{H}^H\left(\mathbf{R}_{\text{noise}}+\gamma^{-1}_{\mathbf{z}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\widehat{\mathbf{z}}^-_{\mathbf{e},t}+\gamma_{\mathbf{x}^-_{\mathbf{e}},t}\widehat{\mathbf{x}}^-_{\mathbf{e},t}\right)$.

$\gamma_{\mathbf{x}^-_{\mathbf{p}},t} = \frac{1}{\frac{1}{K}\mathrm{Tr}\left(\left(\mathbf{H}^H\left(\mathbf{R}_{\text{noise}}+\gamma^{-1}_{\mathbf{z}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\mathbf{H}+\gamma_{\mathbf{x}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\right)}$.

$\gamma_{\mathbf{x}^+_{\mathbf{e}},(t+1)} = \gamma_{\mathbf{x}^-_{\mathbf{p}},t} - \gamma_{\mathbf{x}^-_{\mathbf{e}},t}$.

$\widehat{\mathbf{x}}^+_{\mathbf{e},(t+1)} = \left(\gamma_{\mathbf{x}^-_{\mathbf{p}},t}\widehat{\mathbf{x}}^-_{\mathbf{p},t} - \gamma_{\mathbf{x}^-_{\mathbf{e}},t}\widehat{\mathbf{x}}^-_{\mathbf{e},t}\right)/\gamma_{\mathbf{x}^+_{\mathbf{e}},(t+1)}$.

#LMMSE estimation of z

$\widehat{\mathbf{z}}^-_{\mathbf{p},t} = \left(\mathbf{R}^{-1}_{\text{noise}}+\gamma_{\mathbf{z}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\left(\mathbf{R}^{-1}_{\text{noise}}\mathbf{H}\widehat{\mathbf{x}}^-_{\mathbf{p},t}+\gamma_{\mathbf{z}^-_{\mathbf{e}},t}\widehat{\mathbf{z}}^-_{\mathbf{e},t}\right)$.

$\gamma_{\mathbf{z}^-_{\mathbf{p}},t} = \frac{1}{\frac{1}{N}\mathrm{Tr}\left(\left(\left(\mathbf{R}_{\text{noise}}+\gamma^{-1}_{\mathbf{x}^-_{\mathbf{e}},t}\mathbf{H}\mathbf{H}^H\right)^{-1}+\gamma_{\mathbf{z}^-_{\mathbf{e}},t}\mathbf{I}\right)^{-1}\right)}$.

$\gamma_{\mathbf{z}^+_{\mathbf{e}},(t+1)} = \gamma_{\mathbf{z}^-_{\mathbf{p}},t} - \gamma_{\mathbf{z}^-_{\mathbf{e}},t}$.

$\widehat{\mathbf{z}}^+_{\mathbf{e},(t+1)} = \left(\gamma_{\mathbf{z}^-_{\mathbf{p}},t}\widehat{\mathbf{z}}^-_{\mathbf{p},t} - \gamma_{\mathbf{z}^-_{\mathbf{e}},t}\widehat{\mathbf{z}}^-_{\mathbf{e},t}\right)/\gamma_{\mathbf{z}^+_{\mathbf{e}},(t+1)}$.

endfor

Return $\widehat{\mathbf{x}}^+_{\mathbf{p},(T_{max})}$.

is the SP belief, i.e, the normalized product of all approximate messages impinging on the vector node $\mu_{p_{\mathbf{x}}\to\mathbf{x}^+}(\mathbf{x}^+) = p_{\mathbf{x}^+}(\mathbf{x}^+)$ and $\mu_{\delta_{\mathbf{x}}\to\mathbf{x}^+}(\mathbf{x}^+) = \mathcal{CN}\left(\mathbf{x}^+; \widehat{\mathbf{x}}_{\mathbf{e}}^+, \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}\mathbf{I}\right)$ from the $\delta_{\mathbf{x}}(.)$ and $p_{\mathbf{x}}(.)$. Given the separability of the prior, the SP belief can be factorized:

$$\begin{aligned} {}^{\text{sp}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^+}(\mathbf{x}^+) &= \prod_{k=1}^{K} {}^{\text{sp}}b_{\widehat{x}_{\mathbf{e},k}^+}(x_k^+) \\ &= \prod_{k=1}^{K} {}^{\text{sp}}b_{\Re\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Re\left(x_k^+\right)\right) {}^{\text{sp}}b_{\Im\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Im\left(x_k^+\right)\right), \end{aligned} \tag{3.16}$$

where the real and imaginary SP-beliefs of the k^{th} vector node scalar component are given by the following expressions:

$$\begin{aligned} {}^{\text{sp}}b_{\Re\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Re\left(x_k^+\right)\right) &= \frac{p_{\Re\left(x_k^+\right)}\left(\Re\left(x_k^+\right)\right)\mathcal{N}\left(\Re\left(x_k^+\right); \Re\left(\widehat{x}_{\mathbf{e},k}^+\right), \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}/2\right)}{\sum\limits_{\Re\left(x_k^+\right)\in\mathcal{C}_s} p_{\Re\left(x_k^+\right)}\left(\Re\left(x_k^+\right)\right)\mathcal{N}\left(\Re\left(x_k^+\right); \Re\left(\widehat{x}_{\mathbf{e},k}^+\right), \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}/2\right)} \\ {}^{\text{sp}}b_{\Im\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Im\left(x_k^+\right)\right) &= \frac{p_{\Im\left(x_k^+\right)}\left(\Im\left(x_k^+\right)\right)\mathcal{N}\left(\Im\left(x_k^+\right); \Im\left(\widehat{x}_{\mathbf{e},k}^+\right), \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}/2\right)}{\sum\limits_{\Im\left(x_k^+\right)\in\mathcal{C}_s} p_{\Im\left(x_k^+\right)}\left(\Im\left(x_k^+\right)\right)\mathcal{N}\left(\Im\left(x_k^+\right); \Im\left(\widehat{x}_{\mathbf{e},k}^+\right), \gamma_{\mathbf{x}_{\mathbf{e}}^+}^{-1}/2\right)} \end{aligned} \tag{3.17}$$

$\mathcal{C}_s$ here denotes the set of possible values for both the real and imaginary parts of the employed constellation. Subsequently, the estimator $\mathbf{g}_{\mathbf{x}^+}(\cdot;\gamma_{\mathbf{x}_{\mathbf{e}}^+})$ in line (2) is reduced to a complex scalar estimator $g_{x^+}(\cdot;\gamma_{\mathbf{x}_{\mathbf{e}}^+})$ which is uniformly applied to all incoming message mean components. Specifically, for the k^{th} component, the relationship can be expressed as follows:

$$[\mathbf{g}_{\mathbf{x}^+}(\widehat{\mathbf{x}}_{\mathbf{e}}^+;\gamma_{\mathbf{x}_{\mathbf{e}}^+})]_k = \mathbb{E}\left[x_k^+|{}^{\text{sp}}b_{\widehat{x}_{\mathbf{e},k}^+}(x_k^+)\right] = g_{x^+}(\widehat{x}_{\mathbf{e},k}^+;\gamma_{\mathbf{x}_{\mathbf{e}}^+}) \tag{3.18}$$

Similarly, the complex scalar estimator can be reduced to a real scalar estimator $f_{x^+}\left(\cdot;\gamma_{\mathbf{x}_{\mathbf{e}}^+}\right)$ identically defined for both real and imaginary parts.

$$\begin{aligned} g_{x^+}(\widehat{x}_{\mathbf{e},k}^+;\gamma_{\mathbf{x}_{\mathbf{e}}^+}) &= \mathbb{E}\left[x_k^+|{}^{\text{sp}}b_{\widehat{x}_{\mathbf{e},k}^+}(x_k^+)\right] \\ &= \mathbb{E}\left[\Re\left(x_k^+\right)|{}^{\text{sp}}b_{\Re\left(\widehat{x}_{\mathbf{e},k}^+\right)}(\Re\left(x_k^+\right))\right] + j\mathbb{E}\left[\Im\left(x_k^+\right)|{}^{\text{sp}}b_{\Im\left(\widehat{x}_{\mathbf{e},k}^+\right)}(\Im\left(x_k^+\right))\right] \\ &= f_{x^+}\left(\Re(\widehat{x}_{\mathbf{e},k}^+);\gamma_{\mathbf{x}_{\mathbf{e}}^+}\right) + jf_{x^+}\left(\Im(\widehat{x}_{\mathbf{e},k}^+);\gamma_{\mathbf{x}_{\mathbf{e}}^+}\right) \end{aligned} \tag{3.19}$$

Consequently, the estimated average error can be determined through the differentiation of the estimator concerning the mean input. Equation (3.18) provides the derivation process, commencing with the definition stated in Equation (3.14) and resulting in the expression in line (3).

$$\begin{aligned}
\gamma_{\mathbf{x}_\mathbf{p}^+}^{-1} &= \frac{1}{K}\sum_{k=1}^{K}\mathrm{Var}\left[x_k^+|^{\mathrm{sp}}b_{\widehat{x}_{\mathbf{e},k}^+}(x_k^+)\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\mathrm{Var}\left[\Re\left(x_k^+\right)|^{\mathrm{sp}}b_{\Re\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Re\left(x_k^+\right)\right)\right]+\mathrm{Var}\left[\Im\left(x_k^+\right)|^{\mathrm{sp}}b_{\Im\left(\widehat{x}_{\mathbf{e},k}^+\right)}\left(\Im\left(x_k^+\right)\right)\right]\right) \\
&= \frac{\gamma_{\mathbf{x}_\mathbf{e}^+}^{-1}}{2K}\sum_{k=1}^{K}\left(\frac{\partial\Re\left(g_{x^+}\right)}{\partial\Re\left(\widehat{x}_\mathbf{e}^+\right)}\left(\Re\left(\widehat{x}_{\mathbf{e},k}^+\right);\gamma_{\mathbf{x}_\mathbf{e}^+}\right)+\frac{\partial\Im\left(g_{x^+}\right)}{\partial\Im\left(\widehat{x}_\mathbf{e}^+\right)}\left(\Im\left(\widehat{x}_{\mathbf{e},k}^+\right);\gamma_{\mathbf{x}_\mathbf{e}^+}\right)\right) \\
&= \frac{\gamma_{\mathbf{x}_\mathbf{e}^+}^{-1}}{2K}\sum_{k=1}^{K}\left(\frac{\partial f_{x^+}}{\partial\Re\left(\widehat{x}_\mathbf{e}^+\right)}\left(\Re\left(\widehat{x}_{\mathbf{e},k}^+\right);\gamma_{\mathbf{x}_\mathbf{e}^+}\right)+\frac{\partial f_{x^+}}{\partial\Im\left(\widehat{x}_\mathbf{e}^+\right)}\left(\Im\left(\widehat{x}_{\mathbf{e},k}^+\right);\gamma_{\mathbf{x}_\mathbf{e}^+}\right)\right)
\end{aligned} \tag{3.20}$$

In the case of the Quadrature Phase Shift Keying (QPSK) constellation, the real scalar denoiser with its partial derivative is given by:

$$\begin{aligned}
f_{x^+}(\mu_{x^+};\gamma_{x^+}) &= \tanh\left(\sqrt{2}\gamma_{x^+}\mu_{x^+}\right)/\sqrt{2} \\
\tfrac{\partial f_{x^+}}{\partial\mu_{x^+}}(\mu_{x^+};\gamma_{x^+}) &= \frac{4}{\left(\exp\left\{\sqrt{2}\gamma_{x^+}\mu_{x^+}\right\}+\exp\left\{-\sqrt{2}\gamma_{x^+}\mu_{x^+}\right\}\right)^2}
\end{aligned} \tag{3.21}$$

After the computation of the approximate belief, the variable node $\mathbf{x}^+$ proceeds to transmit the approximate message to $\delta_\mathbf{x}(\cdot)$. In conformity with the VAMP message passing rules, this message is proportional to the ratio of the approximate belief and the message previously received from this delta Dirac node. Subsequently, the message traverses through the delta node to reach the variable node $\mathbf{x}^-$, followed by multiplication with the delta factor node function and integration over the variable node $\mathbf{x}^+$. This process results in the derivation of the extrinsic message at the variable node $\mathbf{x}^-$.

In summary, the extrinsic message is given by:

$$\begin{aligned}\mathcal{CN}\left(\mathbf{x}^{-};\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right) &= \mu_{\delta_{\mathbf{x}}\to\mathbf{x}^{-}}(\mathbf{x}^{-}) \\ &\propto \int_{\mathbf{x}^{+}\in\mathbb{C}^{K}} \delta_{\mathbf{x}}(\mathbf{x}^{+}-\mathbf{x}^{-})\frac{{}^{\text{app}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^{+}}(\mathbf{x}^{+})}{\mu_{\delta_{\mathbf{x}}\to\mathbf{x}^{+}}(\mathbf{x}^{+})}\mathrm{d}\mathbf{x}^{+} \\ &= \frac{{}^{\text{app}}b_{\widehat{\mathbf{x}}_{\mathbf{e}}^{+}}(\mathbf{x}^{-})}{\mu_{\delta_{\mathbf{x}}\to\mathbf{x}^{+}}(\mathbf{x}^{-})}\end{aligned} \tag{3.22}$$

Hence, we get lines (4) and (5) by simply identifying them with the normalized Gaussian form of the impinging message.

Denoising $\mathbf{z}^{+}$

The denoising process for $\mathbf{z}^{+}$ begins with the reception of a message originating from the $\delta_{\mathbf{z}}(\cdot)$ factor node. This message is characterized by a mean value of $\widehat{\mathbf{z}}_{\mathbf{e}}^{+}$ and an average scalar variance of $\gamma_{\mathbf{z}_{\mathbf{e}}^{+}}^{-1}$, both of which are initialized at the start of the algorithm, mirroring the initialization for $\mathbf{x}^{+}$. Analogously to what was done in the denoising step of x, the denoising of z has the approximate belief ${}^{\text{app}}b_{\widehat{\mathbf{z}}_{\mathbf{e}}^{+}}(\mathbf{z}^{+}) = \mathcal{CN}\left(\mathbf{z}^{+};\widehat{\mathbf{z}}_{\mathbf{p}}^{+},\gamma_{\mathbf{z}_{\mathbf{p}}^{+}}^{-1}\mathbf{I}\right)$ where

$$\begin{aligned}\widehat{\mathbf{z}}_{\mathbf{p}}^{+} &= \mathbb{E}\left[\mathbf{z}^{+}|{}^{\text{sp}}b_{\widehat{\mathbf{z}}_{\mathbf{e}}^{+}}(\mathbf{z}^{+})\right] = \mathbf{g}_{\mathbf{z}^{+}}(\widehat{\mathbf{z}}_{\mathbf{e}}^{+};\gamma_{\mathbf{z}_{\mathbf{e}}^{+}}) \\ \gamma_{\mathbf{z}_{\mathbf{p}}^{+}}^{-1} &= \tfrac{1}{N}\,\mathrm{Tr}\left(\mathrm{Cov}[\mathbf{z}^{+}|{}^{\text{sp}}b_{\widehat{\mathbf{z}}_{\mathbf{e}}^{+}}(\mathbf{z}^{+})]\right)\end{aligned} \tag{3.23}$$

herein

$${}^{\text{sp}}b_{\widehat{\mathbf{z}}_{\mathbf{e}}^{+}}(\mathbf{z}^{+}) = \frac{\mathcal{CN}\left(\mathbf{z}^{+};\widehat{\mathbf{z}}_{\mathbf{e}}^{+},\gamma_{\mathbf{z}_{\mathbf{e}}^{+}}^{-1}\mathbf{I}\right)p_{\mathbf{y}|\mathbf{z}^{+}}\left(\mathbf{y}|\mathbf{z}^{+}\right)}{\int_{\mathbf{z}^{+}\in\mathbb{C}^{N}}\mathcal{CN}\left(\mathbf{z}^{+};\widehat{\mathbf{z}}_{\mathbf{e}}^{+},\gamma_{\mathbf{z}_{\mathbf{e}}^{+}}^{-1}\mathbf{I}\right)p_{\mathbf{y}|\mathbf{z}^{+}}(\mathbf{y}|\mathbf{z}^{+})\mathrm{d}\mathbf{z}^{+}} \tag{3.24}$$

is the SP belief, i.e., the normalized product of all approximate messages impinging on the vector node. Here we highlight the importance of including the noise in the variable $\mathbf{z}$ which ensures the separability and allows us to deal with the correlation in the LMMSE bloc. Hence, using conditional independence, the SP belief can be factorized:

$$\begin{aligned}{}^{\text{sp}}b_{\widehat{\mathbf{z}}_{\mathbf{e}}^{+}}(\mathbf{z}^{+}) &= \prod_{n=1}^{N}{}^{\text{sp}}b_{\widehat{z}_{\mathbf{e},n}^{+}}(z_{n}^{+}) \\ &= \prod_{n=1}^{N}{}^{\text{sp}}b_{\Re\left(\widehat{z}_{\mathbf{e},n}^{+}\right)}(\Re\left(z_{n}^{+}\right)){}^{\text{sp}}b_{\Im\left(\widehat{z}_{\mathbf{e},n}^{+}\right)}(\Im\left(z_{n}^{+}\right))\end{aligned} \tag{3.25}$$

The subsequent derivation steps closely resemble the denoising process applied to variable $\mathbf{x}^+$, except that the prior information is substituted with the likelihood. The denoising of $\mathbf{z}^+$ is therefore determined by employing a real scalar estimator $f_{z^+}(\cdot,\cdot;\cdot)$, yielding the results presented in lines (6) and (7). It is essential to note that this estimator in the case of variable z is additionally influenced by the quantization values provided by the measurements. The vector estimator elements are then given by:

$$\begin{aligned}[\mathbf{g}_{\mathbf{z}^+}(\widehat{\mathbf{z}}^+_{\mathbf{e}},\mathbf{y};\gamma_{\mathbf{z}^+_{\mathbf{e}}})]_n &= g_{z^+}(\widehat{z}^+_{\mathbf{e},n},y_n;\gamma_{\mathbf{z}^+_{\mathbf{e}}}) \\ &= f_{z^+}\left(\Re(\widehat{z}^+_{\mathbf{e},n}),\Re(y_n);\gamma_{\mathbf{z}^+_{\mathbf{e}}}\right)+jf_{z^+}\left(\Im(\widehat{z}^+_{\mathbf{e},n}),\Im(y_n);\gamma_{\mathbf{z}^+_{\mathbf{e}}}\right)\end{aligned} \tag{3.26}$$

In the case of a B-bit quantization for a real measurement s_b that falls in the b^{th} bin A_b, the real estimator and its partial derivative with respect to the incoming mean are given by:

$$\begin{aligned}f_{z^+}(\mu_{z^+},s_b;\gamma_{z^+}) &= \mu_{z^+}-\frac{\phi_{(v_b)}-\phi_{(v_{b-1})}}{\Phi_{(v_b)}-\Phi_{(v_b)}} \\ \frac{\partial f_{z^+}}{\partial \mu_{z^+}}(\mu_{z^+},s_b;\gamma_{z^+}) &= 1-\frac{v_b\phi_{(v_b)}-v_{b-1}\phi_{(v_{b-1})}}{\Phi_{(v_b)}-\Phi_{(v_{b-1})}}-\left(\frac{\phi_{(v_b)}-\phi_{(v_{b-1})}}{\Phi_{(v_b)}-\Phi_{(v_{b-1})}}\right)^2\end{aligned} \tag{3.27}$$

where

$$v_b = \sqrt{2\gamma_{z^+}}\,(r_b-\mu_{z^+}) \tag{3.28}$$

Accordingly, we determine the mean and precision, respectively, in lines (8) and (9), of the approximate message sent to the node $\mathbf{z}^-$.

LMMSE of $\mathbf{x}^-$ and $\mathbf{z}^-$

This step starts with the received messages $\mu_{\mathbf{x}^-\to p_{\mathbf{z}|\mathbf{x}}}(\mathbf{x}^-)$ and $\mu_{\mathbf{z}^-\to p_{\mathbf{z}|\mathbf{x}}}(\mathbf{z}^-)$ from the variable nodes $\mathbf{x}^-$ and $\mathbf{z}^-$. These two messages are equal to the extrinsic messages received by the variables from their corresponding delta factor nodes variable nodes. The means and averaged variances in Equations 3.30 of the approximate beliefs of both variables

are determined from their joint SP belief which is proportional to the product of the impinging messages and the factor node function $p_{\mathbf{z}|\mathbf{x}}\left(\mathbf{z}^{-}|\mathbf{x}^{-}\right)=\mathcal{CN}\left(\mathbf{z}^{-};\mathbf{H}\mathbf{x}^{-},\mathbf{R}_{\text{noise}}\right)$, i.e.,:

$$
{}^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})\propto\mathcal{CN}\left(\mathbf{x}^{-};\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right)\mathcal{CN}\left(\mathbf{z}^{-};\mathbf{H}\mathbf{x}^{-},\mathbf{R}_{\text{noise}}\right)\mathcal{CN}\left(\mathbf{z}^{-};\widehat{\mathbf{z}}_{\mathbf{e}}^{-},\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right)
\tag{3.29}
$$

$$
\begin{aligned}
\widehat{\mathbf{x}}_{\mathbf{p}}^{-}&=\mathbb{E}\left[\mathbf{x}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})\right]\\
\widehat{\mathbf{z}}_{\mathbf{p}}^{-}&=\mathbb{E}\left[\mathbf{z}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})\right]\\
\gamma_{\mathbf{x}_{\mathbf{p}}^{-}}^{-1}&=\tfrac{1}{K}\,\text{Tr}\left(\text{Cov}[\mathbf{x}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})]\right)\\
\gamma_{\mathbf{z}_{\mathbf{p}}^{-}}^{-1}&=\tfrac{1}{N}\,\text{Tr}\left(\text{Cov}[\mathbf{z}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})]\right)
\end{aligned}
\tag{3.30}
$$

Calculating LMMSE estimators and means in Equation (3.25) yields lines (10)-(11) and (14)-(15).

$$
\begin{aligned}
\widehat{\mathbf{x}}_{\mathbf{p}}^{-}&=\left(\mathbf{H}^{\mathsf{H}}\left(\mathbf{R}_{\text{noise}}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right)^{-1}\mathbf{H}+\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}\mathbf{I}\right)^{-1}\left(\mathbf{H}^{H}\left(\mathbf{R}_{\text{noise}}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right)^{-1}\widehat{\mathbf{z}}_{\mathbf{e}}^{-}+\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}\widehat{\mathbf{x}}_{\mathbf{e}}^{-}\right)\\
\widehat{\mathbf{z}}_{\mathbf{p}}^{-}&=\left(\mathbf{R}_{\text{noise}}^{-1}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}\mathbf{I}\right)^{-1}\left(\mathbf{R}_{\text{noise}}^{-1}\mathbf{H}\widehat{\mathbf{x}}_{\mathbf{p}}^{-}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)\\
\text{Cov}[\mathbf{x}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})]&=\left(\mathbf{H}^{H}\left(\mathbf{R}_{\text{noise}}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}^{-1}\mathbf{I}\right)^{-1}\mathbf{H}+\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}\mathbf{I}\right)^{-1}\\
\text{Cov}[\mathbf{z}^{-}|^{\text{sp}}b_{\left(\widehat{\mathbf{x}}_{\mathbf{e}}^{-},\widehat{\mathbf{z}}_{\mathbf{e}}^{-}\right)}(\mathbf{x}^{-},\mathbf{z}^{-})]&=\left(\left(\mathbf{R}_{\text{noise}}+\gamma_{\mathbf{x}_{\mathbf{e}}^{-}}^{-1}\mathbf{H}\mathbf{H}^{H}\right)^{-1}+\gamma_{\mathbf{z}_{\mathbf{e}}^{-}}\mathbf{I}\right)^{-1}
\end{aligned}
\tag{3.31}
$$

From the equation above, it becomes evident that the estimators exhibit linearity with respect to the incoming message means, a characteristic arising from the SP belief formulated as a product of complex Gaussian densities. This observation leads to an LMMSE estimator. Finally, we obtain the extrinsic messages transmitted to both denoiser blocks, as indicated in lines (12)-(13) and (16)-(17).

3.3 State Evolution

In this part, we analyze the algorithm's performance by predicting its behaviour in the asymptotic regime, as the problem size tends to infinity. First, we denote N_0 as the reference number of antennas for a fixed-length ULA with an interelement spacing of $d_0 = \lambda/2$. Subsequently, we consider an array with the same length and interelement spacing d and denote the number of antennas as N. To facilitate comparison, we introduce the parameter $\alpha = 2d/\lambda = \frac{N_0}{N}$, which represents the oversampling ratio, i.e., having $\alpha = 1$ means we sample at Nyquist rate. This parameter will be instrumental in our subsequent analysis. In the subsequent sections, we introduce a new tractable channel model for SE (SE) analysis before exploring the algorithm's large system limits.

3.3.1 SE Channel model

In the following section, we explore a channel scenario characterized by abundant scattering phenomena close to the users. In this context, the scatterers are situated a considerable distance from the base station, allowing for the application of the far-field approximation. Additionally, the channel exhibits many multipath components, contributing to the complexity and richness of the received signals. The mathematical channel model is introduced in the following equation:

$$\mathbf{H} = \sqrt{\frac{\mathrm{SNR}}{N_0}}\mathbf{F}\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{S}\mathbf{V}^H \tag{3.32}$$

where $\mathbf{F}, \boldsymbol{\Omega} \in \mathbb{C}^{N\times N}$ are respectively the eigen-vector and diagonal eigen-value matrices of the Toepleitz matrix $\mathbf{R}_{\text{noise;extrinsic}}$ such that the nonzero eigen-values are the first entries of the diagonal components of $\boldsymbol{\Omega}$ with some random permutation. Hence, we have the relationship $\mathbf{R}_{\text{noise;extrinsic}} = \mathbf{F}\boldsymbol{\Omega}\mathbf{F}^H$. $\mathbf{F}\boldsymbol{\Omega}^{\frac{1}{2}}$ captures the array antenna pattern and channel correlation. Matrix $\mathbf{V} \in \mathbb{C}^{K\times K}$ follows a uniform distribution within the set of unitary matrices. In contrast, matrix $\mathbf{S} \in \mathbb{C}^{N\times K}$ takes the form of a rectangular

diagonal matrix, with its diagonal elements representing the singular values derived from a iid complex Gaussian matrix of size N_0 by K, with no specific ordering of the singular values. $\mathbf{S}\mathbf{V}^H$ captures the rich scattering in the vicinity of the users.

3.3.2 Large dimension analysis

Here, the asymptotic regime refers to the case where $K, N, N_0 \to +\infty$ with $\frac{K}{N_0} \to \beta = \mathcal{O}(1)$ and $\frac{N_0}{N} \to \alpha = \mathcal{O}(1)$ for some fixed ratios $\beta \leq 1$ and $\alpha \leq 1$ The SE analysis in approximate message passing relies on applying the concentration of measure principle to the precision variables in the asymptotic regime:

$$\begin{aligned}
\lim_{K,N\to\infty} \left(\gamma_{\mathbf{x}_\mathbf{e}^+}, \gamma_{\mathbf{x}_\mathbf{p}^+}\right) &= \left(\overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}, \overline{\gamma}_{\mathbf{x}_\mathbf{p}^+}\right), \\
\lim_{K,N\to\infty} \left(\gamma_{\mathbf{z}_\mathbf{e}^+}, \gamma_{\mathbf{z}_\mathbf{p}^+}\right) &= \left(\overline{\gamma}_{\mathbf{z}_\mathbf{e}^+}, \overline{\gamma}_{\mathbf{z}_\mathbf{p}^+}\right), \\
\lim_{K,N\to\infty} \left(\gamma_{\mathbf{x}_\mathbf{e}^-}, \gamma_{\mathbf{x}_\mathbf{p}^-}\right) &= \left(\overline{\gamma}_{\mathbf{x}_\mathbf{e}^-}, \overline{\gamma}_{\mathbf{x}_\mathbf{p}^-}\right), \\
\lim_{K,N\to\infty} \left(\gamma_{\mathbf{z}_\mathbf{e}^-}, \gamma_{\mathbf{z}_\mathbf{p}^-}\right) &= \left(\overline{\gamma}_{\mathbf{z}_\mathbf{e}^-}, \overline{\gamma}_{\mathbf{z}_\mathbf{p}^-}\right).
\end{aligned} \tag{3.33}$$

First, we compute the error on $\mathbf{x}^+$. The error is yielded by applying the Central limit theorem.

$$\begin{aligned}
\mathcal{E}_{\mathbf{x}^+}(\overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}) &\triangleq \overline{\gamma}_{\mathbf{x}_\mathbf{p}^+}^{-1} \\
&= \lim_{K\to\infty} \frac{\gamma_{\mathbf{x}_\mathbf{e}^+}^{-1}}{2K} \sum_{k=1}^{K} \left(\frac{\partial f_{x^+}}{\partial \Re\left(x_\mathbf{e}^+\right)} \left(\Re\left(x_{\mathbf{e},k}^+\right); \gamma_{\mathbf{x}_\mathbf{e}^+}\right) + \frac{\partial f_{x^+}}{\partial \Im\left(x_\mathbf{e}^+\right)} \left(\Im\left(x_{\mathbf{e},k}^+\right); \gamma_{\mathbf{x}_\mathbf{e}^+}\right) \right) \\
&= \frac{\overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}^{-1}}{2} \mathbb{E}\left[\left(\frac{\partial f_{x^+}}{\partial \Re\left(x_\mathbf{e}^+\right)} \left(\Re\left(x_\mathbf{e}^+\right); \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}\right) + \frac{\partial f_{x^+}}{\partial \Im\left(x_\mathbf{e}^+\right)} \left(\Im\left(x_\mathbf{e}^+\right); \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}\right) \right) | p_{x_\mathbf{e}^+}\left(x_\mathbf{e}^+\right) \right] \\
&= \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}^{-1} \mathbb{E}\left[\frac{\partial f_{x^+}}{\partial \Re\left(x_\mathbf{e}^+\right)} \left(\Re\left(x_\mathbf{e}^+\right); \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}\right) | p_{x_\mathbf{e}^+}\left(x_\mathbf{e}^+\right) \right] \\
&= \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}^{-1} \mathbb{E}\left[\frac{\partial f_{x^+}}{\partial \Im\left(x_\mathbf{e}^+\right)} \left(\Im\left(x_\mathbf{e}^+\right); \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}\right) | p_{x_\mathbf{e}^+}\left(x_\mathbf{e}^+\right) \right]
\end{aligned} \tag{3.34}$$

where $p_{x_\mathbf{e}^+}\left(x_\mathbf{e}^+\right)$ is determined from the relationship $x_\mathbf{e}^+ = x^+ + w_{x^+}$ with $w_{x^+} \sim \mathcal{CN}(\mathbf{0}, \overline{\gamma}_{\mathbf{x}_\mathbf{e}^+}^{-1})$. We obtain then line (2) in the SE algorithm. Next, we compute the error on $\mathbf{z}^+$ by also applying the central limit theorem on the averaged error having the

relationship $z^+ = z^+_{\mathbf{e}} + w_{z^+}$ with $w_{z^+} \sim \mathcal{CN}(\mathbf{0}, \overline{\gamma}^{-1}_{\mathbf{z}^+_{\mathbf{e}}})$.

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}^+}(\overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}}}) &\triangleq \overline{\gamma}^{-1}_{\mathbf{z}^+_{\mathbf{p}}} \\
&= \lim_{N\to\infty} \tfrac{\gamma^{-1}_{\mathbf{z}^+_{\mathbf{e}}}}{2N} \sum_{n=1}^{N} \left(\tfrac{\partial f_{z^+}}{\partial \Re\left(z^+_{\mathbf{e}}\right)} \left(\Re\left(z^+_{\mathbf{e},n}\right); \gamma_{\mathbf{z}^+_{\mathbf{e}}}\right) + \tfrac{\partial f_{z^+}}{\partial \Im\left(z^+_{\mathbf{e}}\right)} \left(\Im\left(z^+_{\mathbf{e},n}\right); \gamma_{\mathbf{z}^+_{\mathbf{e}}}\right) \right) \\
&= \tfrac{\overline{\gamma}^{-1}_{\mathbf{z}^+_{\mathbf{e}}}}{2} \mathbb{E}\left[\left(\tfrac{\partial f_{z^+}}{\partial \Re\left(z^+_{\mathbf{e}}\right)} \left(\Re\left(z^+_{\mathbf{e}}\right); \overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}}}\right) + \tfrac{\partial f_{z^+}}{\partial \Im\left(z^+_{\mathbf{e}}\right)} \left(\Im\left(z^+_{\mathbf{e}}\right); \overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}}}\right) \right) |p_{z^+_{\mathbf{e}}}\left(z^+_{\mathbf{e}}\right) \right] \\
&= \overline{\gamma}^{-1}_{\mathbf{z}^+_{\mathbf{e}}} \mathbb{E}\left[\tfrac{\partial f_{z^+}}{\partial \Re\left(z^+_{\mathbf{e}}\right)} \left(\Re\left(z^+_{\mathbf{e}}\right); \overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}}}\right) |p_{z^+_{\mathbf{e}}}\left(z^+_{\mathbf{e}}\right) \right] \\
&= \overline{\gamma}^{-1}_{\mathbf{z}^+_{\mathbf{e}}} \mathbb{E}\left[\tfrac{\partial f_{z^+}}{\partial \Im\left(z^+_{\mathbf{e}}\right)} \left(\Im\left(z^+_{\mathbf{e}}\right); \overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}}}\right) |p_{z^+_{\mathbf{e}}}\left(z^+_{\mathbf{e}}\right) \right]
\end{aligned} \tag{3.35}$$

This yields line (4) in Algorithm 5. Regarding the errors in the LMMSE estimator, we leverage the Toeplitz structure of the noise correlation. Through the exploration of the asymptotic behavior of such matrices in [40] (See Appendix B) and the empirical distribution of eigenvalues in Gram matrices of iid Gaussian matrices in [41], we ascertain both errors associated with $\mathbf{x}^-$ and $\mathbf{z}^-$.

$$\begin{aligned}
\mathcal{E}_{\mathbf{x}^-}(\overline{\gamma}_{\mathbf{x}^-_{\mathbf{e}}}) &\triangleq \overline{\gamma}^{-1}_{\mathbf{x}^-_{\mathbf{p}}} \\
&= \lim_{K\to\infty} \tfrac{1}{K} \operatorname{Tr}\left\{ \left(\mathbf{H}^H \left(\mathbf{R}_{\text{noise}} + \gamma^{-1}_{\mathbf{z}^-_{\mathbf{e}}} \mathbf{I} \right)^{-1} \mathbf{H} + \gamma_{\mathbf{x}^-_{\mathbf{e}}} \mathbf{I} \right)^{-1} \right\} \\
&= \overline{\gamma}^{-1}_{\mathbf{x}^-_{\mathbf{e}}} \left(1 - \int_0^1 \tfrac{\mathcal{F}(\mathrm{G}(\omega),\beta)}{4\beta \mathrm{G}(\omega)} \mathrm{d}\omega \right)
\end{aligned} \tag{3.36}$$

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}^-}(\overline{\gamma}_{\mathbf{z}^-_{\mathbf{e}}}) &\triangleq \overline{\gamma}^{-1}_{\mathbf{z}^-_{\mathbf{p}}} \\
&= \lim_{N\to\infty} \frac{1}{N} \operatorname{Tr}\left\{ \left(\left(\mathbf{R}_{\text{noise}} + \gamma^{-1}_{\mathbf{x}^-_{\mathbf{e}}} \mathbf{H}^H \mathbf{H} \right)^{-1} + \gamma_{\mathbf{z}^-_{\mathbf{e}}} \mathbf{I} \right)^{-1} \right\} \\
&= \alpha \int_0^1 \left(\rho(\alpha\pi\omega) + \overline{\gamma}^{-1}_{\mathbf{z}^-_{\mathbf{e}}} + k_B T (N_f - 1) \right)^{-1} \left[1 - \frac{\mathcal{F}\left(\mathrm{G}(\omega), \beta\right)}{4\mathrm{G}(\omega)} \right] \mathrm{d}\omega \\
&\quad + \overline{\gamma}^{-1}_{\mathbf{z}^-_{\mathbf{e}}} - (1-\alpha) \left(\overline{\gamma}^{-1}_{\mathbf{z}^-_{\mathbf{e}}} + k_B T (N_f - 1) \right)^{-1}
\end{aligned} \tag{3.37}$$

wherein the functions $\mathcal{F}(\cdot,\cdot)$ and $\mathrm{G}(\cdot)$ are defined as

$$\mathcal{F}(u,v) = \left(\sqrt{u(1+\sqrt{v})^2 + 1} - \sqrt{u(1-\sqrt{v})^2 + 1} \right)^2, \tag{3.38}$$

$$\mathrm{G}(\omega) = \frac{\mathrm{SNR}\rho(\alpha\pi\omega)}{\overline{\gamma}_{\mathbf{x}^-_{\mathbf{e}}} \left(\rho(\alpha\pi\omega) + \overline{\gamma}^{-1}_{\mathbf{z}^-_{\mathbf{e}}} + k_B T (N_f - 1) \right)} \tag{3.39}$$

$\rho(\cdot)$ corresponds to the angular power spectral density of the extrinsic noise correlation, which depends on the antenna radiation pattern.

Algorithm 5 State Evolution

Require: Denoisers $\mathbf{g}_{\mathbf{x}^+}(\cdot;\cdot)$ and $\mathbf{g}_{\mathbf{z}^+}(\cdot;\cdot)$, LMMSE estimators $\mathbf{g}_{\mathbf{x}^-}(\cdot,\cdot;\cdot,\cdot)$ and $\mathbf{g}_{\mathbf{z}^-}(\cdot,\cdot;\cdot,\cdot)$, and maximum number of iterations T_{max}.

Initialization of $\widehat{\mathbf{x}}^+_{\mathbf{e},1}$, $\widehat{\mathbf{z}}^+_{\mathbf{e},1}$,$\gamma_{\mathbf{x}^+_{\mathbf{e}},1}$ and $\gamma_{\mathbf{z}^+_{\mathbf{e}},1}$.

for t:=1 **to** T_{max} **do**

#Denoising x precision

$\overline{\gamma}_{\mathbf{x}^+_{\mathbf{p}},t} = \frac{1}{\mathcal{E}_{\mathbf{x}^+}(\overline{\gamma}_{\mathbf{x}^+_{\mathbf{e}},t})}$.

$\overline{\gamma}_{\mathbf{x}^-_{\mathbf{e}},t} = \overline{\gamma}_{\mathbf{x}^+_{\mathbf{p}},t} - \overline{\gamma}_{\mathbf{x}^+_{\mathbf{e}},t}$.

#Denoising z precision

$\overline{\gamma}_{\mathbf{z}^+_{\mathbf{p}},t} = \frac{1}{\mathcal{E}_{\mathbf{z}^+}(\overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}},t})}$.

$\overline{\gamma}_{\mathbf{z}^-_{\mathbf{e}},t} = \overline{\gamma}_{\mathbf{z}^+_{\mathbf{p}},t} - \overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}},t}$.

#LMMSE precision of x

$\overline{\gamma}_{\mathbf{x}^-_{\mathbf{p}},t} = \frac{1}{\mathcal{E}_{\mathbf{x}^-}(\overline{\gamma}_{\mathbf{x}^-_{\mathbf{e}},t})}$.

$\overline{\gamma}_{\mathbf{x}^+_{\mathbf{e}},(t+1)} = \overline{\gamma}_{\mathbf{x}^-_{\mathbf{p}},t} - \overline{\gamma}_{\mathbf{x}^-_{\mathbf{e}},t}$.

#LMMSE precision of z

$\overline{\gamma}_{\mathbf{z}^-_{\mathbf{p}},t} = \frac{1}{\mathcal{E}_{\mathbf{z}^-}(\overline{\gamma}_{\mathbf{z}^-_{\mathbf{e}},t})}$.

$\overline{\gamma}_{\mathbf{z}^+_{\mathbf{e}},(t+1)} = \overline{\gamma}_{\mathbf{z}^-_{\mathbf{p}},t} - \overline{\gamma}_{\mathbf{z}^-_{\mathbf{e}},t}$.

end for

Return $\overline{\gamma}^{-1}_{\mathbf{x}^+_{\mathbf{p}},(T_{max})}$.

3.4 Numerical Results

3.4.1 Simulation setup

We have implemented our system model and the adapted VAMP algorithm in MATLAB for simulations. Our simulations were based on the assumption of a massive MIMO scenario, where a base station equipped with a large number of antennas ($N > K$) can efficiently recover signals. This assumption aligns with the common practice in massive MIMO technology, where the number of antennas is typically much larger than the number of users. In our study, we focused on a scenario where the ratio $\beta = \frac{K}{N_0} = \frac{1}{2^3}$. We also set the transmission power to $P_t = 250$ mW, carrier frequency to $f = 60$ GHz, speed of light to $c = 2.997924 \times 10^8$ m/s, absolute temperature to $T = 300$ K, bandwidth to $B_W = 1$ GHz, and Boltzmann constant to $k_B = 1.380649 \times 10^{-23}$ J/K. We have used the optimal step size for quantization, which minimizes distortion, as outlined in [28].

In our study, we considered two types of antenna radiation patterns:

- **Isotropic antenna:** This type is characterized by a uniformly distributed radiation pattern that is indifferent to the polar and azimuth angles of the incident wave. The effective area of a single element of the array is given by:

$$A_{\text{eff}}(\lambda, d) = \frac{d\lambda}{2\pi}. \tag{3.40}$$

- **Hertzian dipole:** This type is characterized by omnidirectional radiation in the plane perpendicular to the antenna end. It depends only on the elevation angle θ, and its effective area is given by:

$$A_{\text{eff}}(\theta, \lambda, d) = \frac{3d\lambda \sin^2 \theta}{4\pi}. \tag{3.41}$$

In all simulation scenarios, each user transmits a block of 100 complex symbols gen-

erated from a QPSK constellation to the receiving antennas. The performance metrics we employ are the Bit Error Rate (BER). To ensure the accuracy of BER calculations, we generate 100 channel realizations for each scenario, repeating the recovery process 100 times with different randomly generated channels. We simulate a rich scattering environment with $L = 100$ multipath components for channels generated by multipath scenarios. To make the small-scale channel coefficients comparable with the SE channel model used in later high-dimensional analyses, we normalize them by $\sqrt{L \times N_0}$. In our benchmarking, we implement VAMP for the linear model with infinite-resolution ADCs and compare our results with those obtained using the LMMSE estimator. For BER calculations, we make hard decisions on the estimates for all the aforementioned methods. To explore the impact of oversampling, we vary the ADC resolution from 1 to 3 bits for both isotropic and dipole antennas. We keep the aperture length fixed at $D = 64\lambda$ corresponding to $N_0 = 128$. At $\mathrm{NF}_{dB} = 3$ dB, we systematically adjust the pseudo-SNR, defined in Equation (3.7), ranging from -4 dB to 18 dB in 2 dB steps. Practically, we manipulate the SNR by varying the distance of users from the base station. Finally, we examine three different values for the array interelement spacing, $d = \lambda/2, \lambda/4, \lambda/8$.

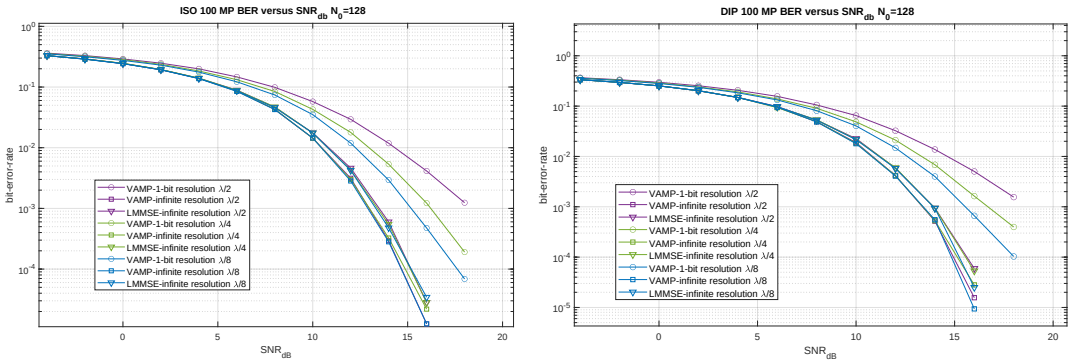
3.4.2 Multi-path channel scenario

In Fig. 3.5, we investigate the impact of oversampling in utilizing 1-bit ADCs, specifically focusing on manipulating the interelement spacing. We examine a total of 100 propagation paths within the Multipath model, employing antennas of both isotropic and dipole types. The 1-bit quantized system exhibits a noteworthy improvement in performance, in contrast to the unquantized system, where observable enhancements are minimal. These results align with expectations, as oversampling beyond the Nyquist rate does not yield benefits in the context of infinite-resolution ADCs. Conversely, in

the case of one-bit ADCs, the positive impact of oversampling is evident due to the induced spectral broadening arising from nonlinearity. Consequently, higher sampling rates prove favourable to enhancing system performance. At low SNR, the observed enhancement is marginal; nonetheless, there is an evident loss of approximately 2dB between 1-bit and infinite quantization, aligning with previous research findings. Conversely, a noticeable improvement becomes evident at moderate and high SNR levels. Notably, the magnitude of improvement from sampling with an interelement spacing of $\lambda/2$ (utilizing 128 antennas) to sampling with an interelement spacing of $\lambda/4$ (employing 256 antennas) appears more pronounced compared to the enhancement observed when oversampling transitions from $\lambda/4$ to $\lambda/8$.In reality, this phenomenon aligns with the principle of diminishing returns. Specifically, the presence of high frequencies (which surpass the Nyquist rate) within the spectrum of quantization noise exhibits components that gradually decrease in importance as sampling frequencies increase. We will delve deeper into this behavior through a thorough analysis in the context of larger dimensions at a later stage.

In terms of benchmarking in the unquantized case, it becomes apparent that for QPSK inputs, the inclusion of prior input information enables VAMP to marginally outperform the LMMSE estimator by employing hard decisions. This outcome is anticipated, as the LMMSE estimator typically yields suboptimal outcomes when subjected to hard decision-making processes in scenarios involving QPSK inputs.

To examine the impact of Noise Figures across various ULA configurations, we plotted Fig. 3.6a and Fig. 3.6b to illustrate the BER vs SNR performance for both 1-bit and infinite resolution quantizations. These figures represent configurations with interelement spacings of $\lambda/2$ and $\lambda/4$, respectively. We systematically adjusted the Noise Figure to 3dB, 5dB, and 7 dB within each configuration. In this investigation, we utilized



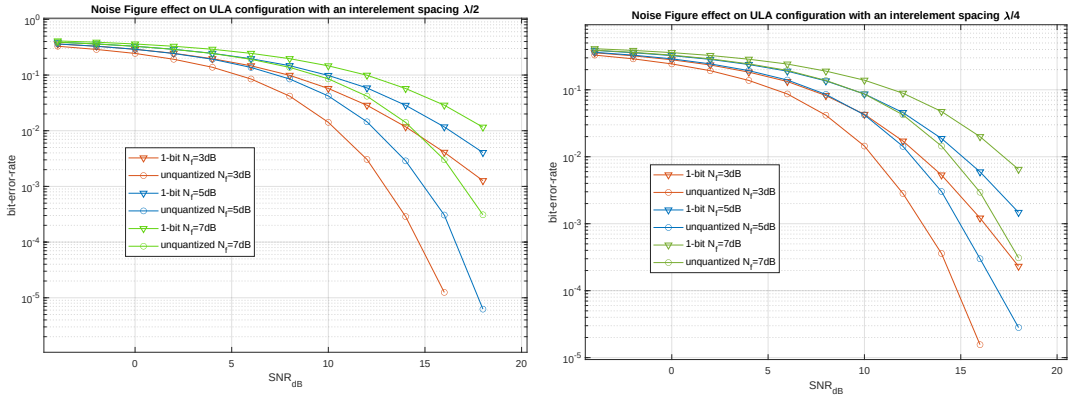
(a) Multi Path channel results using isotropic antennas

(b) Multi Path channel results using dipole antennas

Fig. 3.5: Oversampling performance results when considering Multi Path channel model

isotropic antennas to elucidate the disparity between employing the True Nyquist rate corresponding to $\lambda/2$ spacing and the oversampling rate corresponding to $\lambda/4$ spacing. Examining Fig. 3.6a, associated with the configuration featuring an interelement spacing of $\lambda/2$, we find ourselves in a scenario representative of sampling at the Nyquist rate. This particular setup denotes a special case characterized by the absence of correlation between antenna elements in both the channel and the noise. Our analysis of the outcomes illustrates that the consequence of increasing the noise figure by a certain factor is a corresponding reduction in performance, reflected identically with the same factor across the logarithmic scale. For instance, transitioning from a 3dB to a 5dB noise figure incurred a penalty of 2dB for both the quantized and unquantized systems. This outcome concurs with theoretical expectations, particularly in scenarios featuring uncorrelated channels and uncorrelated noise. Under these conditions, the complexity of the two detection problems referenced in Equations 3.1 and 3.4 is simplified, relying solely on the ratio SNR/N_f. In the scenario involving oversampling with an interelement spacing of $\lambda/4$ as depicted in Fig. 3.6b, we observe that the effect of the noise figure on the unquantized system remains relatively consistent with that observed in the previous

configuration. This consistency persists even when correlations are present, as the fundamental linear problem can always be reframed to depend solely on the SNR/N_f ratio. Nevertheless, in the context of 1-bit quantization, we observe a slight increase in the penalty incurred by introducing noise figures. For instance, the performance experiences a penalty of approximately 2.15 dB when comparing the outcomes associated with 3 dB and 5 dB noise figures or those associated with 5 dB and 7 dB noise figures. Thus, the noise figure affects the performance of the quantized case during oversampling, serving as a significant limitation that must be considered. However, fortunately, oversampling still outperforms the scenario of sampling at the Nyquist rate, particularly at high SNR levels where we can observe a more pronounced downward curvature. We will delve deeper into understanding the noise figure's effect in analyzing larger dimensions.



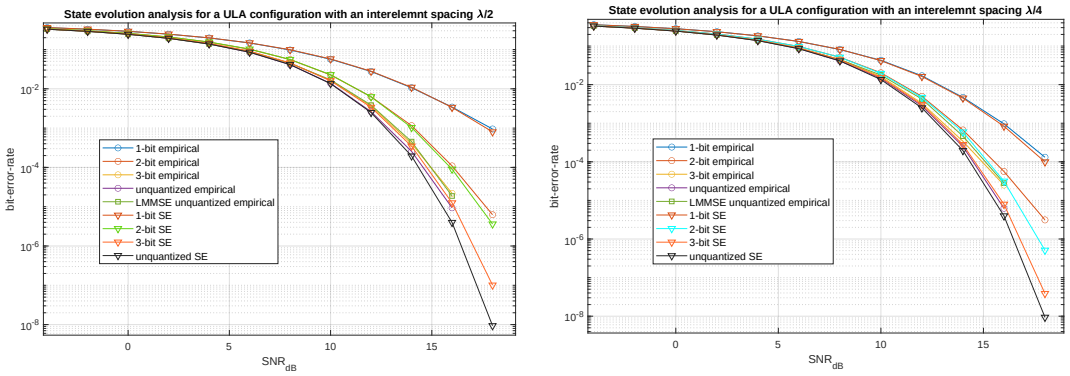
(a) ULA with interelement spacing $\lambda/2$ **(b)** ULA with interelement spacing $\lambda/4$

Fig. 3.6: Noise Figure effect on the performance when considering the Multi-Path model

3.4.3 SE Analysis

In our study on SE, we utilized simulations illustrated in Fig. 3.7 to observe the performance of the SE channel with isotropic antennas. We compared these results with the behavior of the SE algorithm when applied to large systems. We considered two

configurations: the first one, shown in Fig. 3.7a, involved sampling at the Nyquist rate, while the second involved oversampling with an interelement spacing of $\lambda/4$. Additionally, we investigated the performance of ADCs with resolutions ranging from 1 to 3 bits and compared their outcomes with those derived from the unquantized scenario. Upon initial observation of both sets of results and comparing them to the findings in Fig. 3.5a of the MP channel, it appears that the outcomes are comparable for both channel models, this similarity is particularly evident when a large number of multipath components are considered. As both channels demonstrate the same first-order statistics, the SE channel presents a practical alternative to investigate SE. Both configurations show that empirical results approach SE, particularly at low SNR. However, a noticeable disparity emerges as SNR increases to medium and high levels, with the SE exhibiting comparatively superior performance. The observed difference can be traced to the limited dimensions utilized in the empirical analysis. In essence, this limitation arises from a failure to adhere to the Shannon sampling theorem, which dictates the necessity of an infinite number of samples for perfect reconstruction. Consequently, the BER results obtained from the SE algorithm serve as a reference point for algorithmic BER outcomes, offering a lower bound. However, the degradation resulting from finite dimensionality is influenced by additional factors. Firstly, the number of bit resolutions plays a significant role; indeed, both figures demonstrate that higher resolutions aggravate the disparity between empirical and SE results. Secondly, the choice of sampling frequency also contributes to this degradation. A comparison of the two plots reveals that, for each bit resolution, the degradation becomes pronounced at medium and high SNR levels. Another noteworthy observation from the figures concerns the relationship between resolution and the enhancement resulting from oversampling. Specifically, it becomes evident that the rate of improvement diminishes significantly as the number of resolution bits increases, to the point where quantization with a 3-bit resolution closely approximates the performance of the unquantized scenario.



(a) ULA with interelement spacing $\lambda/2$ **(b)** ULA with interelement spacing $\lambda/4$

Fig. 3.7: Comparison between SE and the algorithmic empirical outcomes using SE channel model

Continuing the analysis of oversampling performance illustrated in Fig. 3.8, we used SE to investigate the limits as the oversampling ratio approaches zero, specifically in the context of 1-bit quantization. These results were obtained under an isotropic antenna ULA configuration with a noise figure of 3 dB. We systematically varied the oversampling ratio across discrete values of 1, 0.5, 0.25, 0.125, and 0.0625 and observed a gradual decrease in the extent of performance enhancement as the oversampling ratio decreased, ultimately converging towards a specific limit. These results are consistent with the empirical observations depicted in Fig. 3.5, providing a clearer illustration of the diminishing returns principle. When comparing these results with the unquantized system, it becomes apparent that in scenarios involving infinitely dense arrays, where the ratio $\alpha \to 0$, the 1-bit system closely approximates the performance of the unquantized system, albeit with some loss. Prior research [37] has demonstrated that this loss factor is theoretically dependent solely on the noise figure, as outlined below:

$$lossfactor = \frac{N_F}{1 + \frac{\pi}{2}(N_F - 1)}. \tag{3.42}$$

In this section, we investigate the impact of noise figure on the performance of infinitely

dense arrays with an oversampling ratio $\alpha \to 0$. This is shown in Fig. 3.9. We consider noise figure values of 5 dB, 3 dB, 1 dB, and 0 dB and observe a correlation between SE outcomes and theoretical predictions. The loss factor attributed to 1-bit quantization aligns precisely with theoretical expectations outlined in Equation (3.42). The performance of the 1-bit system is dependent on the noise figure factor. As this factor reduces, we see a convergence of the 1-bit case towards the ideal scenario with no quantization, eventually achieving the same performance at the threshold of $N_F = 0$ dB, representing a nearly ideal receiver chain.

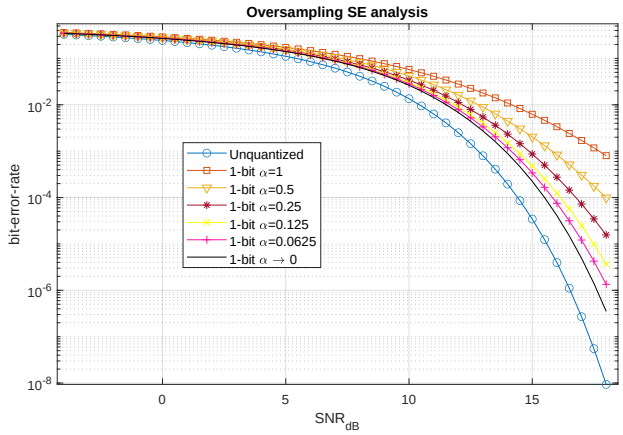


Fig. 3.8: SE oversampling effect when considering different ratios α

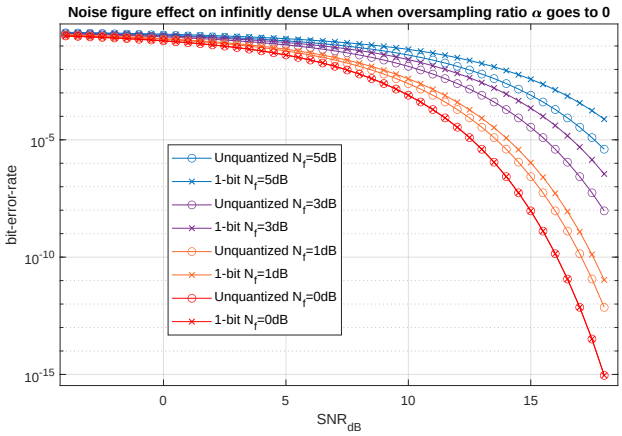


Fig. 3.9: Noise Figure effect when considering infinitely dense arrays with $\alpha \to 0$

Chapter 4

Conclusion

This work introduces a novel non-linear processing method based on the VAMP framework that results to Bayes-optimal solution. This method accommodates B-bits quantized measurements under correlated noise conditions. Furthermore, we incorporate practical channel and noise models, taking into account the ULA configuration parameters, antenna types, and the intrinsic noise induced by the hardware impairments. Through the application of our developed processing method to the proposed model, we have showcased its consistency with prior theoretical findings of the oversampling impact. The analysis has demonstrated the advantages of oversampling, effectively narrowing the performance disparity between low and infinite-resolution ADCs. Moreover, the numerical results revealed a substantial degradation in the performance of low-resolution ADCs in the presence of noise figure, in addition to the typical reduction in SNR. This finding underscores the importance of mitigating the negative impact of noise figure on the system through the implementation of cooling mechanisms.

Looking ahead, future research directions could explore several avenues. First,

investigating physical consistent channel models could provide valuable insights into the power consumption of low-resolution ADCs with spatial oversampling using the proposed framework. Additionally, alternative modulation schemes of the transmitted signal can be exploited using the proposed framework. Lastly, the proposed framework can be also applied to different array geometries (e.g., uniform planar arrays) to optimize the spacing between the antennas.

Appendix A

Schur Complement

Suppose the following bloc matrix $\mathbf{M} \in \mathbb{C}^{(n+m)\times(n+m)}$ with n and m nonnegative integers:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \tag{A.1}$$

where $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{B} \in \mathbb{C}^{n\times m}$, $\mathbf{C} \in \mathbb{C}^{m\times n}$ and $\mathbf{D} \in \mathbb{C}^{m\times m}$.

If $\mathbf{D}$ is invertible, then the **Schur complement** of the block $\mathbf{D}$ of the matrix $\mathbf{M}$ is the $n \times n$ matrix defined by:

$$\mathbf{M}/\mathbf{D} := \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}. \tag{A.2}$$

In addition, if $\mathbf{M}$ is invertible, the inverse of $\mathbf{M}$ may be expressed involving $\mathbf{D}^{-1}$ and the inverse of Schur's complement (if it exists) only as:

$$\mathbf{M}^{-1} = \begin{bmatrix} (\mathbf{M}/\mathbf{D})^{-1} & -(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{M}/\mathbf{D})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}. \tag{A.3}$$

Appendix B

Toeplitz and Circulant matrices

B.1 Asymptotic equivalence

B.1.1 Matrix norms

For a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ we can associate the two follwing norms:

The strong norm $\|.\|$:

$$\|\mathbf{A}\| = \max_{\mathbf{x};\mathbf{x}^{\mathrm{H}}\mathbf{x}=1} \left[\mathbf{x}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A}\mathbf{x}\right]^{1/2} \tag{B.1}$$

The weak norm $|.|$:

$$|\mathbf{A}| = \left(\frac{1}{n}\,\mathrm{Tr}\left[\mathbf{A}^{\mathrm{H}}\mathbf{A}\right]\right)^{1/2} \tag{B.2}$$

B.1.2 Asymptotic equivalence

Two sequences of $n \times n$ complex matrices $\{\mathbf{A}_n\}_{n\in\mathbb{N}^*}$ and $\{\mathbf{B}_n\}_{n\in\mathbb{N}^*}$ are said to be asymptotically equivalent, noted by $\mathbf{A}_n \sim \mathbf{B}_n$, if:

(1) $\mathbf{A}_n$ and $\mathbf{B}_n$ are uniformly bounded in the strong norm:

$$\|\mathbf{A}_n\|, \|\mathbf{B}_n\| < \infty, \forall n \in \mathbb{N}^*$$

(2) $(\mathbf{A}_n - \mathbf{B}_n)$ goes to zero in weak norm as $n \to \infty$:

$$\lim_{n\to\infty} |\mathbf{A}_n - \mathbf{B}_n| = 0$$

B.2 Circulant matrix

B.2.1 Definition

A Circulant matrix $\mathbf{C} \in \mathbb{C}^{n\times n}$ is a Toeplitz matrix defined by the sequence of complex numbers $\{c_i\}_{i\in[\![0,n-1]\!]}$, as follows:

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & c_2 & & \vdots \\ c_{n-2} & c_{n-1} & c_0 & c_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & c_2 \\ \vdots & & \ddots & \ddots & \ddots & c_1 \\ c_1 & \cdots & \cdots & c_{n-2} & c_{n-1} & c_0 \end{bmatrix}, \tag{B.3}$$

where each row is a cyclic shift of the row above it.

B.2.2 Spectral properties

The m-th eigenvalue ρ_m of a circulant matrix is identified by the DFT of its sequence as follows:

$$\rho_m = \sum_{k=0}^{n-1} c_k e^{-2\pi imk/n} \tag{B.4}$$

The corresponding normalized eigenvector $\mathbf{u}_m$ of ρ_m is identified by the columns of the DFT matrix as follows:

$$\mathbf{u}_m = \frac{1}{\sqrt{n}} \left(1, e^{-2\pi i m/n}, \cdots, e^{-2\pi i m(n-1)/n}\right)^\top \tag{B.5}$$

B.3 Toeplitz matrix

B.3.1 Definition

A Toeplitz matrix $\mathbf{T} \in \mathbb{C}^{n\times n}$ is defined, by the sequence of complex numbers $\{t_i\}_{i\in[\![-(n-1),n-1]\!]}$, as follows:

$$\mathbf{T} = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & \cdots & \cdots & t_{-(n-1)} \\ t_1 & t_0 & t_{-1} & \ddots & & \vdots \\ t_2 & t_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & t_{-1} & t_{-2} \\ \vdots & & \ddots & t_1 & t_0 & t_{-1} \\ t_{n-1} & \cdots & \cdots & t_2 & t_1 & t_0 \end{bmatrix}, \tag{B.6}$$

When the dimension of $\mathbf{T}$ tend to infinity the mathematical description of such matrices is still valid by the sequence $\{t_i\}_{i\in\mathbb{Z}}$. Thus, a Toeplitz matrix is called **Wiener Class**, when the sequence $\{t_i\}_{i\in\mathbb{Z}}$ is absolutely summable.

B.3.2 Asymptotic behavior and property

Let $\mathbf{C}$ a circulant matrix defined from the sequence of the Topleitz matrix in Equation B.6 as follows:

$$\mathbf{C} = \begin{bmatrix} t_0 & t_{-1} & \cdots & t_{-m} & & & & & t_m & \cdots & t_1 \\ t_1 & & & & & & & & & \ddots & \vdots \\ & & & & & & & & & & t_m \\ \vdots & & & & \ddots & & & & & & \\ t_m & & & & & & & & 0 & & \\ & \ddots & & & & & & & & & \\ & & t_m & \cdots & t_1 & t_0 & t_{-1} & \cdots & t_{-m} & & \\ & & & \ddots & & & & & & \ddots & \\ & & & 0 & & & & & & & t_{-m} \\ t_{-m} & & & & & & & & & & \\ \vdots & \ddots & & & & & & & & & \vdots \\ & & & & & & & & & t_0 & t_{-1} \\ t_{-1} & \cdots & & t_{-m} & & & & t_m & & \cdots\ t_1 & t_0 \end{bmatrix} \tag{B.7}$$

It has been demonstrated in [40] that the Toeplitz matrix $\mathbf{T}$ is asymptotically equivalent to the circulant matrix defined previously. Consequently, in high-dimensional scenarios, the spectral properties, including eigenvalues and eigenvectors, of a Wiener-class Toeplitz matrix are equivalent to those of the corresponding circulant matrix derived from the sequence as described in Equation B.7.

B.3.3 Power density function and theorem

The power density function $f(.)$ of a Topleitz matrix is defined as the Discrete Fourier Transform of its corresponding sequence:

$$f(\lambda) \quad = \quad \sum_{k=-\infty}^{\infty} t_k e^{ik\lambda}; \lambda \in [0, 2\pi] \tag{B.8}$$

Theorem [40]: Under the assumptions of Szego theorem for a function $F(.)$ that is continuous in the range of the spectral power density $f(.)$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} F(\tau_{n,k}) = \frac{1}{2\pi} \int_0^{2\pi} F(f(\lambda)) d\lambda \tag{B.9}$$

where $\tau_{n,k}$ for $k \in [\![0, n-1]\!]$ are the eigenvalues of the Topleitz matrix of size n by n defined from the sequence $\{t_i\}_{i\in[\![-(n-1),n-1]\!]}$.

Bibliography

[1] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

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

[3] 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, 2014.

[4] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5g: physical and mac-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[5] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek, "A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect csi," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 484–498, 2018.

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

[7] 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. [Online]. Available: https://doi.org/10.1109%2Fmcom.2018.1600731

[8] S. Wang, Y. Li, and J. Wang, "Multiuser detection in massive spatial modulation MIMO with low-resolution ADCs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2156–2168, 2015.

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

[10] I. O'Donnell and R. Brodersen, "An ultra-wideband transceiver architecture for low power, low rate, wireless systems," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 5, pp. 1623–1631, 2005.

[11] 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, 2005.

[12] 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*, 2007, pp. 1286–1289.

[13] 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), 2015, pp. 191–198.

[14] Y. Li, C. Tao, A. Mezghani, A. L. Swindlehurst, G. Seco-Granados, and L. Liu, "Optimal design of energy and spectral efficiency tradeoff in one-bit massive MIMO systems," 2017.

[15] Q. Bai and J. A. Nossek, "Energy efficiency maximization for 5g multi-antenna receivers," *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 1, pp. 3–14, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.2892

[16] 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, 2015.

[17] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Communications Letters*, vol. 19, no. 12, p. 2186–2189, Dec. 2015. [Online]. Available: http://dx.doi.org/10.1109/LCOMM.2015.2494600

[18] A. Mezghani and J. A. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:163162025

[19] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," 2014.

[20] 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, 2017.

[21] J. Mo, P. Schniter, and R. W. Heath, "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1141–1154, 2018.

[22] A. Mezghani and J. A. Nossek, "Belief propagation based MIMO detection operating on quantized channel output," in *2010 IEEE International Symposium on Information Theory*, 2010, pp. 2113–2117.

[23] S. Wang, Y. Li, and J. Wang, "Multiuser detection in massive spatial modulation MIMO with low-resolution ADCs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2156–2168, 2015.

[24] 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, 2016.

[25] 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, 2016.

[26] W. R. Bennett, "Spectra of quantized signals," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.

[27] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*, 06 2008.

[28] A. Mezghani, "Information-theoretic analysis and signal processing techniques for quantized MIMO communications," Ph.D. dissertation, Technische Universität München, 2015.

[29] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Large-scale MIMO transmitters in fixed physical spaces: The effect of transmit correlation and mutual coupling," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2794–2804, 2013.

[30] S. Biswas, C. Masouros, and T. Ratnarajah, "Performance analysis of large multiuser MIMO systems with space-constrained 2-d antenna arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3492–3505, 2016.

[31] P. Kyritsi, D. Cox, R. Valenzuela, and P. Wolniansky, "Correlation analysis based on MIMO channel measurements in an indoor environment," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 713–720, 2003.

[32] D. S. Palguna, D. J. Love, T. A. Thomas, and A. Ghosh, "Millimeter wave receiver design using low precision quantization and parallel $\delta\sigma$ architecture," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6556–6569, 2016.

[33] H. Pirzadeh, G. Seco-Granados, A. L. Swindlehurst, and J. A. Nossek, "On the effect of mutual coupling in one-bit spatial sigma-delta massive MIMO systems," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[34] D. Scholnik, J. Coleman, D. Bowling, and M. Neel, "Spatio-temporal delta-sigma modulation for shared wideband transmit arrays," in *Proceedings of the 2004 IEEE Radar Conference (IEEE Cat. No.04CH37509)*, 2004, pp. 85–90.

[35] J. D. Krieger, C.-P. Yeang, and G. W. Wornell, "Dense delta-sigma phased arrays," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 1825–1837, 2013.

[36] M. Shao, W.-K. Ma, Q. Li, and A. L. Swindlehurst, "One-bit sigma-delta MIMO

precoding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 1046–1061, 2019.

[37] A. Mezghani, F. Bellili, and J. a. Robert W. Heath, "Massive MIMO with dense arrays and 1-bit data converters," 2020.

[38] F. R. Kschischang, B. J. Frey, and H. . Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[39] A. Maleki and A. Montanari, "Analysis of approximate message passing algorithm," pp. 1–7, 2010.

[40] R. M. Gray, 2006.

[41] A. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.