uc3m | Universidad Carlos III de Madrid

DOCTORAL THESIS

Approximate Inference in Massive MIMO Scenarios with Moment Matching Techniques

Author:
Javier Céspedes Martín

Advisors:
Pablo Martínez Olmos
Matilde Sánchez Fernández

Departamento de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

Leganés, 17 January 2017

uc3m | Universidad **Carlos III** de Madrid

Tesis Doctoral:	Approximate Inference in Massive MIMO Scenarios with Moment Matching Techniques
Autor:	Javier Céspedes Martín
Director:	Dr. Pablo Martínez Olmos
Directora:	Dra. Matilde Sánchez Fernández

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores:

Presidente:	Juan José Murillo Fuentes
Vocal:	Isabel Valera Martínez
Secretario:	Gonzalo Vázquez Vilar

acuerda otorgarle la calificación de:

Leganés, 17 de Enero de 2017

Agradecimientos

Me gustaría remontarme a mi proyecto de fin de carrera, aquel PFC me ha llevado hasta donde estoy ahora mismo. Gracias Fernando, tú abriste la puerta de este camino que tanta satisfacción me ha llevado.

Al entrar en el master con aquella beca recuerdo que la asignación del tutor era algo que yo no controlaba. Mi equipo de trabajo inicial, era Mati, pero al poco tiempo se incorporó Pablo. Creo que otros no podrán decir lo mismo, pero si ahora me preguntan, si pudieras volver atrás y elegir ¿Cambiarías el equipo de trabajo?. Mi respuesta no puede ser más rotunda, ¡NO!, no cambiaría nada. Trabajar con vosotros es increíble. Muchas gracias.

A todos y cada uno de los profesores del departamento, cada sabía palabra, cada consejo que te lleva a mejorar día a día, tanto en enseñanza como en docencia, son para tenerlos muy en consideración y seguro que llegará el día de ponerlo en práctica.

Por los momentos pasados, y que pasaremos, con sentimientos encontrados, al fin y al cabo nuestro camino es similar. Agradecer a todos los compañeros de laboratorio. Ana, Özge, Omar, Cecilia, Juanjo, Máximo, Alex, Borja, Alex A., Juan Carlos, Estefanía, Iratxe, Alex B., Kun, Ahmad. Muchas gracias, recorrer el camino a vuestro lado ha sido genial. Como consejo para lo que aún os queda parte de camino, *cuando nos caemos nos levantamos, y seguimos adelante.*

He tenido la suerte de poder hacer dos estancias, en primer lugar Suecia, en la Universidad de Lund. Muy enfocada en la tecnología MIMO, allí conocí a Fredik Rusek y parte del equipo Joao y Pepe, gracias por los momentos tan buenos que allí pasé. Mi vuelta no fue en las mejores condiciones ni cuando debía, pero vosotros me ayudasteis mucho. En segundo lugar Irlanda, en los Nokia Bell Labs. Pude comparar lo que es la investigación realizada bajo una empresa. Una experiencia realmente enriquecedora, pero más si cabe por el increíble grupo de personas con las que allí coincidí. Giovanni, Lorenzo, Andrea, Adrían, Alexandros pero especialmente a David. Muchas gracias a todos allí aprendí y disfruté mucho (salvando un pequeño susto).

Debido a que este es un largo camino, también hay momentos de tensión y estress, yo aficionado al deporte, y debido a las posibilidades de las que disponemos en la Universidad. He podido disfrutar de jugar al tenis, de un gran profesor, gracias Fernando, por los días en los que todas las bolas entraban y por los que no entraba ni una. Gracias a todos y cada uno de mis compañeros, por vosotros monté junto con Alex el ranking de tenis, al final cualquier pique que pasaba en la pista, quedará en la pista, las risas y buenos ratos, esos quedan en mi memoria.

En los dos últimos años, he hecho una incursión en una nueva actividad que en realidad me ha cambiado, he visto que la hermandad en ella es brutal, gracias compañeros. Al lado de Reyes, entrenar cada día produce una satisfacción instantánea, he notado agujetas en partes que no sabía que podría tenerlas, puedo decir que ahora conozco mejor mi cuerpo. Siempre con ganas de ir con expectación a un nuevo entrenamiento. Gracias, porque he cumplido retos, que ni hubiera creído posibles con anterioridad a unirme a tu equipo.

A mis amigos del pueblo, ya que a pesar de no pasar mucho tiempo con ellos, siempre habéis estado ahí. Ir al pueblo para estar con vosotros es un placer, desconectar de una forma de vida, para poder conectar con otra en Daimiel, es genial. Sé que vosotros siempre me apoyáis. Gracias.

Desde el primer año de carrera, en la residencía, conocí a Honorio. El ha realizado este camino delante de mí, y en muchos de los momentos por los que he pasado, me ha apoyado. Gracias por compartir piso tantos años, y por toda la ayuda que me has bridado, realmente no puedo imaginar cómo recompensarlo.

Por último y no menos importante, de hecho es como un postre, no me importaría repetir. Mi familia, mis hermanos Víctor y Jorge, y mi mamá M^{a} Pilar. No voy a repetir el agradecimiento que hubo en el PFC, pero sé que vosotros habéis estado ahí, cada carcajada, cada lágrima, cada momento con vosotros aunque no fuera divertido me ha servido para coger aire fuerte, y poder seguir con mi camino en los momentos más duros. Si los momentos duros llegan sé que con vosotros, todo puede ser superado. No me puedo olvidar de esas personas que ya no están, pero que también me han ayudado de levantarme después de caer, de hecho las vuestras han sido las mayores caidas. Papa, mis mayores logros también son para tí, solo pido una cosa, que siempre me guardes. Abuelo y Abuela, sé que vosotros al igual que mis padres estáis orgullosos. Yo, solo puedo decir que os estimo mucho.

En general, para mi *Familia y Amigos*, los que siempre han estado y siempre me han apoyado.

Contents

Abstract **V**

Acronyms **XV**

Notation **XIX**

I Introduction & Motivation **1**

I.1 MIMO Systems 1

I.2 Symbol Detection in Massive MIMO 3

I.2.1 Hard-Detection System Model 4

I.2.2 Soft-Detection System Model 6

I.3 Contributions 10

II Hard-Detection Methods **13**

II.1 Sphere Detection Methods 13

II.2 Minimum Mean Squared Error 14

II.3 MMSE with Successive Interference Cancellation 15

II.4 Gaussian Tree-Approximation 16

II.5 GTA with Successive Interference Cancellation 17

II.6 Lattice Reduction Techniques 18

II.7 Tabu Search Detection 19

II.8 Comparison of State-of-the-art Methods 19

III Hard-Detection via EP Approximations **21**

III.1 Expectation Propagation Approximate Inference 22

III.2 Expectation Propagation Detection 24

III.3 EP MIMO Detector Updates 25

III.4 EPD complexity 27

III.5 EP combined with LLL 30

III.6 Experimental Results 31

IV Soft-Detection Methods 37

IV.1 MMSE and GTA as Soft Detectors 38

IV.2 CHEMP algorithm 38

IV.3 Markov Chain Monte Carlo Methods 40

IV.4 Comparison of State-of-the-art Methods 41

V Soft-Detection via EC Approximations 45

V.1 Expectation Consistency Approximate Inference 46

V.2 EC MIMO detection 50

V.2.1 EC distributions 50

V.2.2 Single Loop ECD detection 52

V.3 EPD vs ECD 54

V.3.1 Assessing convergence 57

V.4 Experimental Results 60

V.4.1 Calibration Curves 60

V.4.2 A Low Dimensional MIMO System 62

V.4.3 A $m = r = 32$ MIMO system 65

V.4.4 Feedback helps for LDPC block lengths 67

VI Conclusions 71

Abstract

This Thesis explores low-complexity inference probabilistic algorithms in high-dimensional Multiple-Input Multiple-Output (MIMO) systems and high-order M-Quadrature Amplitude Modulation (QAM) constellations. Several modern communications systems are using more and more antennas to maximize spectral efficiency, in a new phenomena call Massive MIMO. However, as the number of antennas and/or the order of the constellation grow several technical issues have to be tackled, one of them is that the symbol detection complexity grows fast exponentially with the system dimension. Nowadays the design of massive MIMO low-complexity receivers is one important research line in MIMO because symbol detection can no longer rely on conventional approaches such as Maximum a Posteriori (MAP) due to its exponential computation complexity. This Thesis proposes two main results. On one hand a hard decision low-complexity MIMO detector based on Expectation Propagation (EP) algorithm which allows to iteratively approximate within polynomial cost the posterior distribution of the transmitted symbols. The receiver is named Expectation Propagation Detector (EPD) and its solution evolves from Minimum Mean Square Error (MMSE) solution and keeps per iteration the MMSE complexity which is dominated by a matrix inversion. Hard decision Symbol Error Rate (SER) performance is shown to remarkably improve state-of-the-art solutions of similar complexity. On the other hand, a soft-inference algorithm, more suitable to modern communication systems with channel codification techniques such as Low-Density Parity-Check (LDPC) codes, is also presented. Modern channel decoding techniques need as input Log-Likehood Ratio (LLR) information for each coded bit. In order to obtain that information, firstly a soft bit inference procedure must be performed. In low-dimensional scenarios, this can be done by marginalization over the symbol posterior distribution. However, this is not feasible at high-dimension. While EPD could provide this probabilistic information, it is shown that its probabilistic estimates are in general poor in the low Signal-to-Noise Ratio (SNR) regime. In order to solve this inconvenience a new algorithm based on the Expectation Consistency (EC) algorithm, which generalizes several algorithms such as Belief

Propagation (BP) and EP itself, was proposed. The proposed algorithm called Expectation Consistency Detector (ECD) maps the inference problem as an optimization over a non convex function. This new approach allows to find stationary points and tradeoffs between accuracy and convergence, which leads to robust update rules. At the same complexity cost than EPD, the new proposal achieves a performance closer to channel capacity at moderate SNR. The result reveals that the probabilistic detection accuracy has a relevant impact in the achievable rate of the overall system. Finally, a modified ECD algorithm is presented, with a Turbo receiver structure where the output of the decoder is fed back to ECD, achieving performance gains in all block lengths simulated.

The document is structured as follows. In Chapter I an introduction to the MIMO scenario is presented, the advantages and challenges are exposed and the two main scenarios of this Thesis are set forth. Finally, the motivation behind this work, and the contributions are revealed. In Chapters II and III the state of the art and our proposal are presented for Hard Detection, whereas in Chapters IV and V are exposed for Soft Inference Detection. Eventually, a conclusion and future lines can be found in Chapter VI.

Resumen

Esta Tesis aborda algoritmos de baja complejidad para la estimación probabilística en sistemas de Multiple-Input Multiple-Output (MIMO) de grandes dimensiones con constelaciones M-Quadrature Amplitude Modulation (QAM) de alta dimensionalidad. Son diversos los sistemas de comunicaciones que en la actualidad están utilizando más y más antenas para maximizar la eficiencia espectral, en un nuevo fenómeno denominado Massive MIMO. Sin embargo los incrementos en el número de antenas y/o orden de la constelación presentan ciertos desafíos tecnológicos que deben ser considerados. Uno de ellos es la detección de los símbolos transmitidos en el sistema debido a que la complejidad aumenta más rápido que las dimensiones del sistema. Por tanto el diseño receptores para sistemas Massive MIMO de baja complejidad es una de las importantes líneas de investigación en la actualidad en MIMO, debido principalmente a que los métodos tradicionales no se pueden implementar en sistemas con decenas de antenas, cuando lo deseable serían centenas, debido a que su coste es exponencial.

Los principales resultados en esta Tesis pueden clasificarse en dos. En primer lugar un receptor MIMO para decisión dura de baja complejidad basado en el algoritmo Expectation Propagation (EP) que permite de manera iterativa, con un coste computacional polinómico por iteración, aproximar la distribución a posteriori de los símbolos transmitidos. El algoritmo, denominado Expectation Propagation Detector (EPD), es inicializado con la solución del algoritmo Minimum Mean Square Error (MMSE) y mantiene el coste de este para todas las iteraciones, dominado por una inversión de matriz. El rendimiento del decisor en probabilidad de error de símbolo muestra ganancias remarcables con respecto a otros métodos en la literatura con una complejidad similar. En segundo lugar, un algoritmo que provee una estimación blanda, información que es más apropiada para los actuales sistemas de comunicaciones que utilizan codificación de canal, como pueden ser códigos Low-Density Parity-Check (LDPC). La información necesaria para estos decodificadores de canal es Log-Likehood Ratio (LLR) para cada uno de los bits codificados.

En escenarios de bajas dimensiones se pueden calcular las marginales

de la distribución a posteriori, pero en escenarios de grandes dimensiones no es viable, aunque EPD puede proporcionar este tipo de información a la entrada del decodificador, dicha información no es la mejor al estar el algoritmo pensado para detección dura, sobre todo se observa este fenómeno en el rango de baja Signal-to-Noise Ratio (SNR). Para solucionar este problema se propone un nuevo algoritmo basado en Expectation Consistency (EC) que engloba diversos algoritmos como pueden ser Belief Propagation (BP) y el algoritmo EP propuesto con anterioridad. El nuevo algoritmo llamado Expectation Consistency Detector (ECD), trata el problema como una optimización de una función no convexa. Esta aproximación permite encontrar los puntos estacionarios y la relación entre precisión y convergencia, que permitirán reglas de actualización más robustas y eficaces. Con la misma compleja que el algoritmo propuesto inicialmente, ECD permite rendimientos más próximos a la capacidad del canal en regímenes moderados de SNR. Los resultados muestran que la precisión tiene un gran efecto en la tasa que alcanza el sistema. Finalmente una versión modificada de ECD es propuesta en una arquitectura típica de los Turbo receptores, en la que la salida del decodificador es la entrada del receptor, y que permite ganancias en el rendimiento en todas las longitudes de código simuladas.

El presente documento está estructurado de la siguiente manera. En el primer Capítulo I, se realiza una introducción a los sistemas MIMO, presentando sus ventajas, desventajas, problemas abiertos. Los modelos que se utilizaran en la tesis y la motivación con la que se inició esta tesis son expuestos en este primer capítulo. En los Capítulos II y III el estado del arte y nuestra propuesta para detección dura son presentados, mientras que en los Capítulos IV y V se presentan para detección suave. Finalmente las conclusiones que pueden obtenerse de esta Tesis y futuras líneas de investigación son expuestas en el Capítulo VI.

List of Tables

II.1 Maximum Number of neighbors for $\Delta = 1$. 19

III.1 Simulated Hard-Detection Scenarios 31

V.1 EPD steps 55
V.2 ECD steps 56

List of System Models

I.1 Hard Detection Scenario 5

I.2 Probabilistic Symbol Detector 7

I.3 Probabilistic Symbol Detector with Channel Coding 8

List of Figures

I.1 Configurations for Multi-Antenna Systems 2
I.2 Spectral Efficiency 3
I.3 Multiple-Input Multiple-Output (MIMO) User Configuration . . . 4
I.4 Achievable Rate in a $m = r = 5$ system with 4-QAM 11

II.2 SER performance of several detectors in a scenario with $m = r = 6$ and a 4-QAM constellation. 20

III.1 Evolution of each component of the EPD mean $\boldsymbol{\mu}^{(\ell)}$ in (III.12) as EPD iterates for a $m = r = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. 29
III.2 Evolution of each component of the EPD covariance $\boldsymbol{\Sigma}^{(\ell)}$ in (III.11) as EPD iterates for a $m = r = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. 29
III.3 SER performance in a $m = r = 6$ system with 4-QAM. 32
III.4 SER performance in a $m = r = 12$ system with 16-QAM. 33
III.5 SER performance in a $m = r = 32$ system with 16-QAM. 33
III.6 SER performance in a $m = r = 64$ system with 16-QAM. 34
III.7 SER performance in a $m = r = 64$ system with 64-QAM. 34
III.8 SER performance in a $m = r = 100$ system with 16-QAM. 35
III.9 SER performance in a $m = r = 250$ system with 16-QAM. 35
III.10SER performance in several systems (increasing $m = r$) and 16-QAM. 36

IV.1 Performance in a $m = r = 5$ scenario with 4-QAM modulatio. (a), achievable transmission rates. (b), BER with a $(3, 6)$-regular Low-Density Parity-Check (LDPC) code and block length $k = 5120$ bits. 42
IV.2 $m = r = 32$ MIMO system with 4-QAM and 16-QAM constellation. A $(3, 6)$-regular LDPC with code block $k = 10240$ bits has been used. 43

V.1 Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 5$ scenario with 4-QAM modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. 58

V.2 Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 32$ scenario with 4-QAM modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. 59

V.3 Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 32$ scenario with 16-QAM modulation at a SNR of 13dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. . 60

V.4 Calibration curves in $m = r = 5$ system with a 4-QAM at SNR = 7dB. 61

V.5 Calibration curves in $m = r = 5$ system with a 4-QAM at SNR = 13.5dB. 62

V.6 $m = r = 5$ system with 4-QAM modulation, in (a) achievable transmission rates. In (b), with a $(3, 6)$-regular LDPC code with block length $k = 5120$ bits. 64

V.7 $m = r = 32$ MIMO system with 4-QAM and 16-QAM constellation. A $(3, 6)$-regular LDPC with code block $k = 10240$ bits has been used. 65

V.8 Acheivable rate computed for a $m = r = 32$ MIMO system with 4-QAM. 66

V.9 Acheivable rate computed for a $m = r = 32$ MIMO system with 16-QAM. 66

V.10 Acheivable rate computed for a $m = r = 32$ MIMO system with 64-QAM. 67

V.11 $r = 32$ MIMO system with 16-QAM modulation, in (a) achievable transmission rates for different m values. In (b), BER performance when a $(3, 6)$-regular LDPC with code block $k = 5120$ bits. 68

V.12 BER performance of a $m = r = 32$ system with 16-QAM, using the irregular (dashed lines), $R = 1/2$ LDPC with code block $k30720$ bits and a $(3, 6)$-regular (solid lines) LDPCC with the same block and coding rate 0.479. 69

V.13 ECD performance in a $m = r = 32$ system with 16-QAM using the feedback loop. We consider a $(3, 6)$-regular LDPC code with different k lengths. 70

Acronyms

5G 5^{th} Generation of Mobile Communications.

AMP Approximate Message Passing.

BEC Binary Erasure Channel.

BER Bit Error Rate.

BMSC Binary Memoryless Symmetric Channel.

BP Belief Propagation.

BS Base Station.

CDMA Code Division Multiple Access.

CHEMP Channel Hardening-Exploiting Message Passing.

CSI Channel State Information.

EC Expectation Consistency.

ECD Expectation Consistency Detector.

EP Expectation Propagation.

EPD Expectation Propagation Detector.

EPD-LLL EPD with Lenstra-Lenstra-Lovász.

GTA Gaussian Tree-Approximation.

GTA-SIC GTA with Successive Interference Cancellation.

HSPA High Speed Packet Access.

IEEE Institute of Electrical and Electronics Engineers.

ISI Inter Symbol Interference.

KL Kullback-Leibler.

LD Linear Detector.

LDPC Low-Density Parity-Check.

LDPCC Convolutional Low-Density Parity-Check.

LLL Lenstra-Lenstra-Lovász.

LLR Log-Likehood Ratio.

LR Lattice Reduction.

LTE Long Term Evolution.

MAP Maximum a Posteriori.

MCMC Markov-Chain Monte-Carlo.

MIMO Multiple-Input Multiple-Output.

MISO Multiple-Input Single-Output.

ML Maximum Likehood.

MM moment-matching.

MMSE Minimum Mean Square Error.

MMSE-SIC MMSE with Successive Interference Cancellation.

MU-MIMO Multi-User Multiple-Input Multiple-Output.

OFDM Orthogonal Frequency Division Multiplexing.

pdf probability density function.

QAM Quadrature Amplitude Modulation.

RF Radio-Frequency.

SD Spatial Diversity.

SER Symbol Error Rate.

SIMO Single-Input Multiple-Output.

SISO Single-Input Single-Output.

SM Spatial Multiplexing.

SNR Signal-to-Noise Ratio.

SNR$_\mathbf{c}$ Coded Signal-to-Noise ratio.

SpD Sphere Decoding.

SU-MIMO Single-User Multiple-Input Multiple-Output.

TS Tabu Search.

Notation

a	Scalar.
$\mathbf{a}$	Vector.
a_i	i-th position of the vector $\mathbf{a}$.
$\mathbf{a}_{-i}$	all components in $\mathbf{a}$ except a_i.
$\mathbf{A}$	Matrix.
$\mathbf{a}_i$	i-th column of the Matrix $\mathbf{A}$.
$\mathbf{A}_{-i}$	all components in $\mathbf{A}$ except $\mathbf{a}_i$.
$\mathbf{I}[\cdot]$	Identity Matrix.
$\mathbb{I}$	Indicator Function.
$\mathcal{N}(\mathbf{a} : \mathbf{b}, \mathbf{C})$	probability density function of a Normal distribution over $\mathbf{a}$ with mean $\mathbf{b}$ and covariance matrix $\mathbf{C}$.
$\mathcal{O}(\cdot)$	Computational cost in operations.
$\top$	Transpose.
$*$	Hermitian.
$\dagger$	pseudo-inverse.
D_{KL}	Kullback-Leibler Divergence.
$\mathrm{diag}(\mathbf{x})$	returns a square diagonal matrix with diagonal given by $\mathbf{x}$.
$\mathrm{diag}(\mathbf{X})$	over a square matrix denotes its diagonal vector.
$\{n\}$	denotes the set $\{1, 2, \ldots, n\}$.
$\mathcal{R}(\cdot)$	Real part.
$\mathcal{I}(\cdot)$	Imaginary part.

$\mathbb{E}_{p(\mathbf{u})}$	Expectation with respect to the distribution $p(\mathbf{u})$.
$\mathrm{d}_{\mathrm{H}}\left(\cdot\right)$	Hamming distance.
$\propto$	proportional to.

Chapter I

Introduction & Motivation

I.1 MIMO Systems

In the last years wireless systems have attracted a great deal of interest due to the expansion of mobile communications in detriment of wired systems, which most of the times require higher investements at the deployment process. Radio wireless communications were tradiditonally based on Single-Input Single-Output (SISO) antenna systems, where detection and equalization techniques have affordable compexity [1]. However, the current user demand of higher rates and service reliability is turning the interest back to MIMO systems, airing at increasing the channel capacity and improving spectral efficiency [2–4]. Nowadays, MIMO is at the core of many modern communications systems such as High Speed Packet Access (HSPA) and Long Term Evolution (LTE) [5] in mobile communications. Also, the Institute of Electrical and Electronics Engineers (IEEE) has standarized MIMO techniques, in IEEE802.11n and IEEE802.11ac [6] for WI-FI. Further, at the new mobile communication generation, the so-called 5^{th} Generation of Mobile Communications (5G) [7, 8], MIMO techniques play a relevant role. Several antenna configurations are possible in a wireless communication system. In Figure I.1, different scenarios are shown: SISO, Multiple-Input Single-Output (MISO), Single-Input Multiple-Output (SIMO) and MIMO. Spectral efficiency is maximized in the latter case, and this will be the focus of this Thesis from now on.

Wireless communications are affected by fading, variations on the signal strength and may cause a dramatic degradation on the system's performance. MIMO systems comprise a collection of techniques proposed to enhance the performance of wireless systems by exploiting the scattering environment as the result of having multiple antennas at the transmitter side and the receiver side [9]. The two main characteristics behind MIMO

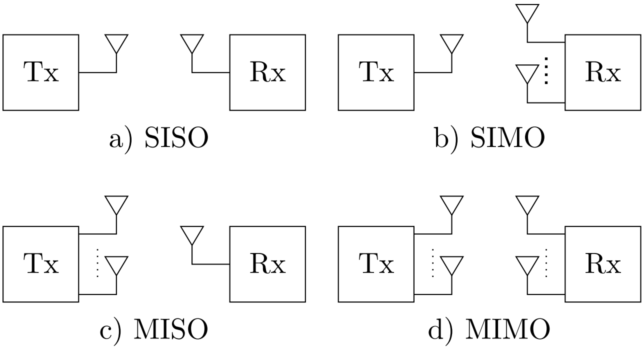


Figure I.1: Configurations for Multi-Antenna Systems

systems are Spatial Diversity (SD) and Spatial Multiplexing (SM). On one hand, SD tries to improve the reliability by combating channel fading using space-time techniques [10]. These techniques exploit the fact that replicas of the transmitted signal (in both time and space) arrive to the receiver affected by different fading coefficients. In a MIMO system with m transmit antennas and r receive antennas, SD is achieved through space-time coding techniques and it is measured by the diversity gain N_d, full SD is obtained when $N_d = mr$ [11]. On the other hand, SM increases throughput by exploiting demultiplexing techniques [12]. In SM several data streams are transmitted over a fading channel exploiting the multipath. Transmit antennas share time and frequency so the efficiency in bits per Hertz is increased. The maximum number of independent streams that can be transmitted is $N_{\text{streams}} = \min(m, r)$ and is called Multiplexing Gain [11].

The overall system rate is determined by both SD and SM, so a proper design is important to reach the desired performance, and an optimal traded-off between both gains can be found [2, 13]. In this Thesis, we do not consider SD techniques and focus on exploiting SM in a MIMO channel, following a V-BLAST architecture [14]. Thus any channel coding technique that we implement does not consider the spatial dimension. In such a case, according to [3], in a scenario with a Rayleigh-distributed channel without Channel State Information (CSI) at the transmitter and perfect CSI at the receiver, the channel spectral efficiency increases linearly with $\min(m, r)$, as shown in Figure I.2 for $m = r$.

Roughly speaking, Massive MIMO appears when we target $m, r \to \infty$, but in the research community has two different meanings. Whereas for some researchers it is a general scenario where a large number of antennas is used (in the range of hundreds), for others basically is the result of using several techniques all together: a celular scenario with multiple users and very large arrays of antennas in the base station [8, 15, 16]. In both

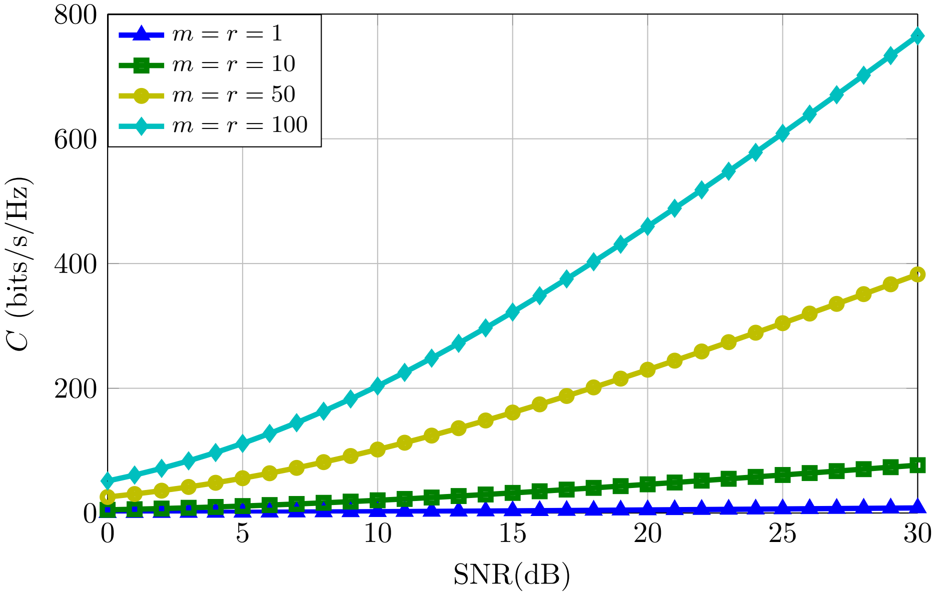


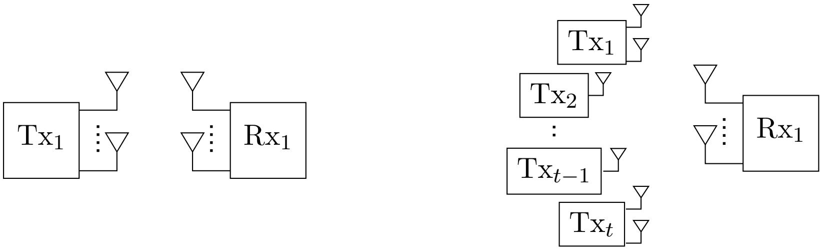
Figure I.2: Spectral Efficiency

scenarios, as the number of antennas is increased, it is possible to find the mentioned MIMO benefits in terms of spectral efficiency, but at the same time complexity becomes an important challenge. The most important technical challenges in massive MIMO are the total antenna array size, the need for deployment of a large number of Radio-Frequency (RF) chains and the signal processing complexity at both the transmit and receive sides [4]. This Thesis explores new techniques in this latter aspect.

When a point-to-point MIMO communications system is used, it is usually called Single-User Multiple-Input Multiple-Output (SU-MIMO), whereas a multipoint-to-point communication system is called Multi-User Multiple-Input Multiple-Output (MU-MIMO). In Fig. I.3 it is possible to observe the different configurations, and it is important to recall that, besides the fact that there are several transmitters in MU-MIMO and, further, that it is an heterogeneous scenario in which users may have different number of antennas, both SU-MIMO and MU-MIMO systems have in common the same uplink receiver configuration.

I.2 Symbol Detection in Massive MIMO

As the transmit antennas in MIMO are sharing time-frequency resources, the symbols arrive to the receiver antennas as a linear superposition [17].



a) SU-MIMO b) MU-MIMO

Figure I.3: MIMO User Configuration

Signal detection is the process by which the receiver estimates the transmitted symbols, dealing with impairments caused by the noise or fading [14]. As the number of antennas m increases, the dimension of the space that contains all possible transmitted vectors grows exponentially fast with m, for that reason optimal detection is completely unaffordable and accurate low-complexity approximate solutions are needed. Actually, this problem still needs a viable solution for a full deployment of these systems, as the gap of state-of-the-art methods with respect to optimal techniques is still very large.

I.2.1 Hard-Detection System Model

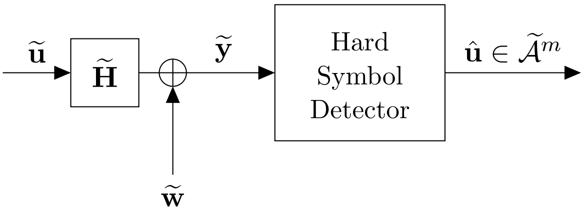
Consider a MIMO system where m transmit antennas communicate to a receiver with r antennas and each antenna transmits a M-Quadrature Amplitude Modulation (QAM) symbol at each channel use. The system model is shown in System Model I.1, where $\widetilde{\mathbf{u}} = \mathbf{a} + j\mathbf{z} \in \widetilde{\mathcal{A}}^m$ is the transmitted vector of QAM symbols, $\widetilde{\mathbf{H}}$ is a $r \times m$ complex matrix representing a memoryless flat-fading complex MIMO channel, and $|\widetilde{\mathcal{A}}| = M$. Each coefficient of $\widetilde{\mathbf{H}}$ is drawn according to a complex zero-mean unit-variance Gaussian distribution, following a channel model without line of sight and large scattering [18]. As perfect CSI is assumed by the receiver, $\widetilde{\mathbf{H}}$ is known at the receiver. The channel output is $\widetilde{\mathbf{y}} \in \mathbb{C}^r$, where

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}\widetilde{\mathbf{u}} + \widetilde{\mathbf{w}}, \tag{I.1}$$

and $\widetilde{\mathbf{w}} \in \mathbb{C}^r$ is an additive white circular-symmetric complex Gaussian noise vector with independent zero-mean components and $\sigma^2_{\widetilde{w}}$-variance. According to this model, the Signal-to-Noise Ratio (SNR) is defined as:

$$\mathrm{SNR(dB)} = 10 \log_{10} \left(m \frac{\widetilde{E_s}}{\sigma^2_{\widetilde{w}}} \right), \tag{I.2}$$

where $\widetilde{E_s}$ is the constellation average energy in Joules.



System Model I.1: Hard Detection Scenario

The channel model in (I.1) can be written in the real domain by considering the real and imaginary parts separately. Defining:

$$\mathbf{u} = \begin{bmatrix}\mathbf{a}^\top & \mathbf{z}^\top\end{bmatrix}^\top \qquad \mathbf{w} = \begin{bmatrix}\mathcal{R}(\widetilde{\mathbf{w}})^\top & \mathcal{I}(\widetilde{\mathbf{w}})^\top\end{bmatrix}^\top$$

$$\mathbf{y} = \begin{bmatrix}\mathcal{R}(\widetilde{\mathbf{y}})^\top & \mathcal{I}(\widetilde{\mathbf{y}})^\top\end{bmatrix}^\top \qquad \mathbf{H} = \begin{bmatrix}\mathcal{R}\left(\widetilde{\mathbf{H}}\right) & -\mathcal{I}\left(\widetilde{\mathbf{H}}\right) \\ \mathcal{I}\left(\widetilde{\mathbf{H}}\right) & \mathcal{R}\left(\widetilde{\mathbf{H}}\right)\end{bmatrix}.$$

The real-valued channel model is

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \mathbf{w}, \tag{I.3}$$

where $\sigma_w^2 = \sigma_{\widetilde{w}}^2/2$ is the variance of the real and imaginary components of the noise, and $\mathcal{A}$ is defined as the new alphabet for the real and imaginary components of the M-QAM constellation, i.e. $\mathbf{u} \in \mathcal{A}^{2m}$, with energy $E_s = \widetilde{E_s}/2$. The real-valued model is assumed without loss of generality in the rest of the Thesis. Also squared QAM constellations are used, as the real and imaginary components can be treated independently, hence $|\mathcal{A}| = \sqrt{M}$.

The MAP Detector

Upon observing $\mathbf{y}$, the optimal detector implements Maximum a Posteriori (MAP) criterion over the joint probability density function (pdf) of the transmitted vector of QAM, $p(\mathbf{u}|\mathbf{y})$,

$$\hat{\mathbf{u}}_{\text{MAP}} = \arg \max_{\mathbf{u} \in \mathcal{A}^{2m}} p(\mathbf{u}|\mathbf{y}), \tag{I.4}$$

and it minimizes the error rate $p(\mathbf{u} \neq \hat{\mathbf{u}}_{\text{MAP}})$. Applying Baye's theorem [19] over (I.4) we have

$$\begin{aligned} p(\mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{u})p(\mathbf{u})}{p(\mathbf{y})} &\propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})\; p(\mathbf{u}) \\ &\propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I}) \prod_{i=1}^{2m} \frac{1}{\sqrt{M}} \mathbb{I}_{u_i \in \mathcal{A}}, \end{aligned} \tag{I.5}$$

where $\mathbb{I}_{u_i \in \mathcal{A}}$ is equal to 1 if $u_i \in \mathcal{A}$ and zero otherwise. $\mathcal{N}(\mathbf{y} : \mathbf{Hu}, \sigma_w^2 \mathbf{I})$ is a probability density function of a Normal distribution over $\mathbf{y}$ with mean $\mathbf{Hu}$ and covariance matrix $\sigma_w^2 \mathbf{I}$. Assuming that the transmitted M-QAM symbols have the same probability, it is possible to observe that the MAP criterion is equivalent to maximize the likelihood $p(\mathbf{y}|\mathbf{u})$. Let $p(\mathbf{y}|\mathbf{u})$ be the likelihood function, the Maximum Likehood (ML) symbol detection is:

$$\hat{\mathbf{u}}_{\text{ML}} = \arg \max_{\mathbf{u} \in \mathcal{A}^{2m}} p(\mathbf{y}|\mathbf{u}) = \arg \min_{\mathbf{u} \in \mathcal{A}^{2m}} ||\mathbf{y} - \mathbf{Hu}||^2. \tag{I.6}$$

As described, ML is a brute-force algorithm which looks among all possible vectors $\mathbf{u} \in \mathcal{A}^{2m}$ and it is not affordable for medium-large systems. Over the scenario proposed, the number of different symbols $\mathbf{u}$ grows exponentially with m and M as M^m.

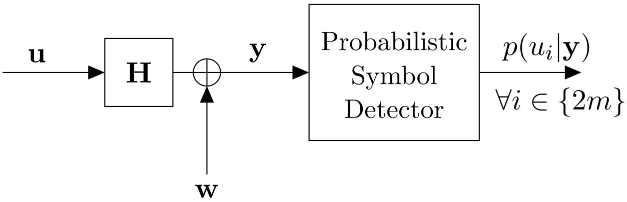
Given this complexity, it is clear that need to find a way to reduce the computational cost of the detectors if a large number of antennas and/or constellation is demanded. Two are the main directions in which MIMO detectors have evolved. On one hand, some proposals reduce the search space in (I.6). On the other hand, some other perform a low complexity approximation to $p(\mathbf{u}|\mathbf{y})$ [20–24]. Methods on both directions will be discussed in the next Chapters.

Performance Metric in Hard-Detection

To evaluate the performance of hard symbol detection, the Symbol Error Rate (SER) is used, defined as $\frac{1}{2m}\sum_{i=1}^{2m} \mathbb{I}\left[u_i \neq \hat{u}_i\right]$. More precisely, the number of QAM symbols wrongly estimated. Note that the former definition of SER is an empirical estimate to $p(u_i \neq \hat{u}_i)$.

I.2.2 Soft-Detection System Model

Coding techniques are used to improve the performance of any communication system, reducing the number of errors at the receiver side. However, it is necessary to remark that the benefits of the coding techniques do not come at zero cost, as more bits must be transmitted, usually called as redundancy bits. This cost is measured by the coding rate R. Modern channel coding techniques such as LDPC [25] codes or Turbo Coding techniques [26] rely on soft inference approximations, so the previous receiver implementation with hard symbol detection is no longer valid, i.e, the output $\hat{u}_i \in \mathcal{A}$ in System Model I.1 needs to be replaced by $p(u_i|\mathbf{u})$ $\forall i \in \{2m\}$. Modern channel decoders update iteratively the probabilistic information provided by the symbol detector during the decoding process until convergence or a stopped criterium is reached. It is well known that the more accurate the inference is made at the symbol detector, the better performance is obtained after the channel decoding [27–29]. Hence, the receiver soft output is $p(u_i|\mathbf{y})\ \forall i \in \{2m\}$. The System Model I.1 is modified as shown in System Model I.2. To incorporate explicitly the channel encoder and decoder over the System Model I.2, a binary information stream is encoded using a block code of rate $R = k/n$, where k is the length of the input-sequence and n is the block length. Let $\mathbf{b} = \left[b_1, b_2, \cdots, b_k\right]^\top$ denote the input information binary vector and $\mathbf{c} = \left[c_1, c_2, \cdots, c_n\right]^\top$ the corresponding codeword. Assume it takes L channel uses to transmit a complete codeword, $L \in \mathbb{Z}_+$. Codeword $\mathbf{c}$ is Gray-mapped and modulated into L vectors of QAM symbols, $\mathbf{U} = [\mathbf{u}[1], \cdots, \mathbf{u}[L]]$. For simplicity, we assume $n = \log_2(M)\, mL$.



System Model I.2: Probabilistic Symbol Detector

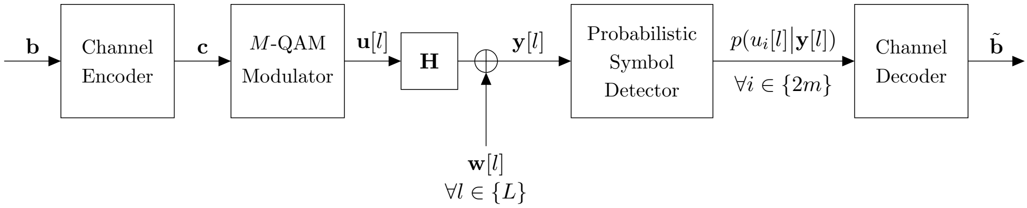
The use of channel coding also introduces delay at the receiver, since $\mathbf{Y} = [\mathbf{y}[1], \cdots, \mathbf{y}[L]]$ needs to be observed before the decoding process can begin. If convolutional codes or turbo codes are used, this delay can be mitigated. However, only block LDPC codes are used for our experiments, and thus the decoder works with the complete vector $\mathbf{Y}$. The received vector is given by:

$$\mathbf{y}[l] = \mathbf{H}[l]\mathbf{u}[l] + \mathbf{w}[l] \quad \forall l \in \{L\}\,. \tag{I.7}$$

Before any decoding process is performed, a marginalization over the posterior distribution I.8 must be carried out, as explained below. The resulting system model is given in System Model I.3.

$$\begin{aligned} p(\mathbf{u}[l]|\mathbf{y}[l]) &= \frac{p(\mathbf{y}[l]|\mathbf{u}[l])p(\mathbf{u}[l])}{p(\mathbf{y}[l])} \\ &\propto \mathcal{N}(\mathbf{y} : \mathbf{H}[l]\mathbf{u}[l], \sigma_w^2\mathbf{I})\ p(\mathbf{u}[l])\ \forall l \in \{L\}\,, \end{aligned} \tag{I.8}$$

Furthermore the SNR (I.2) in the uncoded system differs with the one in the coded system Coded Signal-to-Noise ratio (SNR$_\mathrm{c}$), to take into account



System Model I.3: Probabilistic Symbol Detector with Channel Coding

the decreased symbol energy required to maintain transmitted power:

$$\begin{aligned}\mathrm{SNR_c(dB)} &= 10\log_{10}\left(m\log_2 M\frac{k}{n}\frac{E_b}{\sigma_w^2}\right)\\ &= 10\log_{10}\left(m\log_2 M\frac{E_b}{\sigma_w^2}\right) + 10\log_{10}\left(\frac{k}{n}\right)\\ \mathrm{SNR_c(dB)} &= \mathrm{SNR} + 10\log_{10}(R), \end{aligned} \tag{I.9}$$

where $E_s = \log_2 ME_b$ is the constellation average energy.

Optimal Soft-Detector

Without loss of generality the coded bits are sequentially mapped into M-QAM symbols, so the bit assigned to j-th position of the Gray code at the i-th antenna during the l-th use of the channel is $c_{j+\log_2 M(i-1)+2m\log_2 M(l-1)}$, with $\forall j \in \{\log_2 M\}$, $\forall i \in \{2m\}$ and $\forall l \in \{L\}$. In order to simplify the notation the coded bit is renamed to $c_{ji}[l]$, furthermore in following equations the channel use l is also omitted. The posterior probability of the c_{ji} bit for a given channel observation $\mathbf{y}$ can be computed as follows:

$$p(c_{ji} = c|\mathbf{y}) = \sum_{u_i \in \mathcal{B}_j(c)} p(u_i|\mathbf{y}), \tag{I.10}$$

for $c \in \{0,1\}$, where $\mathcal{B}_j(c) = \big\{u_i \in \mathcal{A}|\mathrm{Gray}_j\,(u) = c\big\}$ and $\mathrm{Gray}_j\,(u_i)$ is the bit in the j-th position of the Gray encoding of symbol $\mathbf{u}$. Extending (I.10) using (I.8), it is possible to obtain:

$$\begin{aligned} p(c_{ji} = c|\mathbf{y}) = \sum_{u_i \in \mathcal{B}_j(c)} p(u_i|\mathbf{y}) &= \sum_{u_i \in \mathcal{B}_j(c)} \sum_{\mathbf{u}_{-i} \in \mathcal{A}^{2m-1}} p(\mathbf{u}|\mathbf{y})\\ &\propto \sum_{u_i \in \mathcal{B}_j(c)} \sum_{\mathbf{u}_{-i} \in \mathcal{A}^{2m-1}} p(\mathbf{y}|\mathbf{u}) \prod_{i=1}^{2m} p(u_i). \end{aligned} \tag{I.11}$$

Despite the detector ignores possible underlying correlations between the coded symbols by assuming the independent prior distribution $p(\mathbf{u})$, computing the symbol posterior probability $p(u_i|\mathbf{y})$ in (I.10) for each antenna still requires $\mathcal{O}(M^m)$ operations. At it was argued before, for a high-dimensional MIMO scenario, where m scales up to hundreds of antennas, the resulting complexity is prohibitive. Under these circumstances, approximate detection methods are needed.

Once the coded bit probabilities in (I.10) are computed, it is possible to obtain the Log-Likehood Ratio (LLR) of those bits, as it is the usual input for the probabilistic channel decoders

$$\mathrm{LLR}(c_{ji}) = \log \frac{p(c_{ji}=1|\mathbf{y})}{p(c_{ji}=0|\mathbf{y})} = \log \frac{\sum_{u_i \in \mathcal{B}_j(1)} p(u_i|\mathbf{y})}{\sum_{u_i \in \mathcal{B}_j(0)} p(u_i|\mathbf{y})}. \tag{I.12}$$

After the decoding process, the output is also a probability for each coded bit. These bit probabilities can be used for final decision $\hat{\mathbf{b}}$, which characterizes open-loop architectures [30, 31], or they can be used to re-initialize the detection stage, resulting in a closed-loop Turbo-like architecture [29, 32].

It is important to recall that the description given in System Model I.3 here applies to a SU-MIMO case. While the receiver structure would be maintained for MU-MIMO case, at the transmitter side each user encode their bit stream into M-QAM constellations.

Performance Metrics in Soft-Detection

Bit Error Rate (BER) after the decoding process is one of the available metrics for the System Model I.3. It is simply given by how many information bits are wrongly estimated after the decoding process $\sum_{i=1}^{k} \mathbb{I}\left[b_i \neq \hat{b}_i\right]$, which is an empirical estimate to $p(b_i \neq \hat{b}_i)$. However, the use of this metric is certainty problematic, as it is necessary to select a particular coding rate, channel code, channel encoder, and channel decoder. A more fundamental performance metric for the System Model I.2 is given in terms of mutual information. Consider a fixed and known channel matrix $\mathbf{H}$. The ergodic channel capacity per transmitted antenna with perfect CSI at the receiver and no CSI at the transmitter is given by:

$$C = \max_{p(\mathbf{u})} \frac{I(\mathbf{u}, \mathbf{y})}{2m} = \frac{\log_2(\det(\mathbf{I}_r + \frac{\mathrm{SNR}}{2m}\mathbf{H}\mathbf{H}^H))}{2m} \tag{I.13}$$

bits per channel use and antenna [33]. Capacity is achieved when $\mathbf{u}$ is Gaussian distributed with zero-mean and covariance matrix equal to identity. When $\mathbf{u}$ is a random vector uniformly distributed in $\mathcal{A}^{2m}$, the system transmission rate degrades and can be far from the capacity limit in (I.13).

The achievable rate per antenna can be computed by evaluating the mutual information between u_i, the transmitted symbol at i-th antenna, $\forall i \in \{2m\}$, and $\hat{u}_i \sim p(u_i|\mathbf{y})$, i.e.,

$$I(u_i, \hat{u}_i) = \mathbb{E}_{p(u_i,\hat{u}_i)} \left[\log_2 \frac{p(\hat{u}_i|u_i)}{p(\hat{u}_i)} \right] \quad \text{(bits/channel use)}. \tag{I.14}$$

However, it is not possible to compute this mutual information in closed-form. A Monte Carlo procedure [34] was followed to estimate that rate in the same channel knowledge scenario as the one assumed in (I.13), namely perfect CSI only at the receiver.

More precisely, at each SNR point $I(u_i, \hat{u}_i)$ for $\forall i \in \{2m\}$ is estimated as follows: firstly, collecting $N \in \mathbb{Z}_+$ samples from the joint distribution of $u_i, \mathbf{y}$ and $\hat{u}_i$. Using this set of samples, we estimate $p(\hat{u}_i)$ and $p(\hat{u}_i|u_i)$ for any u_i, $\hat{u}_i \in \mathcal{A}^{2m}$, and finally, compute a numerical estimate to $I(u_i, \hat{u}_i)$ in (I.14). As $N \to \infty$, this estimate gets accurate. Samples of the joint $(u_i, \mathbf{y}, \hat{u}_i)$ distribution are computed using *ancestral sampling* [35], where each of the N samples is generated following the next steps:

1. Sample $\mathbf{u}$ from a uniform distribution in $\mathcal{A}^{2m}$.
2. Sample $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{u}, \mathbf{H})$.
3. Sample $\hat{u}_i$, $i \in \{2m\}$, from $p(\hat{u}_i|\mathbf{y})$.

In Fig. I.4 an example of the achievable rate by the optimal soft detector is shown for a 4-QAM constellation. It has been computed with $N = 10^6$ samples per SNR point. Also, results have been averaged over 100 realizations of $\mathbf{H}$. Observe that the detector operates close to the limit of $\log_2(M) = 2$ bits/channel use when the SNR is high, but the gap to channel capacity in this regime grows exponentially fast with the SNR. For intermediate SNR the gap to channel capacity is reduced significantly. It is precisely in this regime where we need to operate in order to improve the system efficiency.

I.3 Contributions

MIMO receiver design is one of the most challenging topics in wireless communications, as it was already exposed. The reason is that the complexity may be prohibitive in the Massive MIMO scenario, and thus approximation methods must be explored. There are several low-complexity methods proposed until now to solve the problem (see [24] and the references therein for an in-depth review). Many of them are described in detail in Chapters

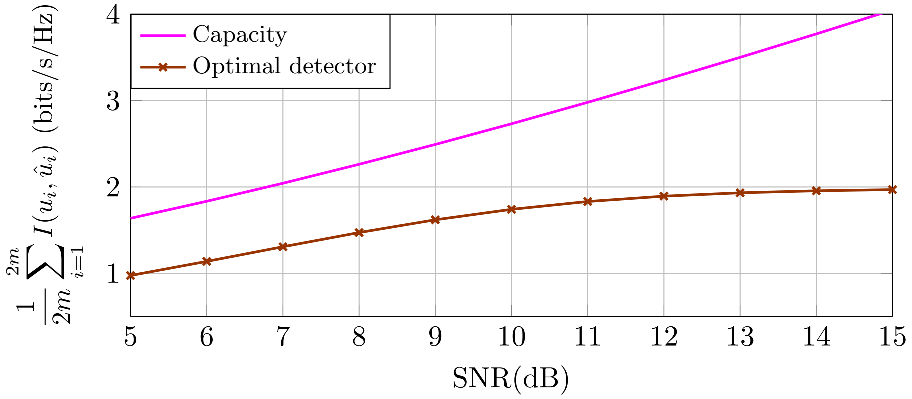


Figure I.4: Achievable Rate in a $m = r = 5$ system with 4-QAM

II (hard-decision algorithms), and IV (soft-decision algorithms). A representative example taken as a baseline, is the Minimum Mean Square Error (MMSE) algorithm, which has cost $\mathcal{O}(m^3)$, and it is able to perform well in the high SNR regime, and also when $r >> m$.

The main contribution of this Thesis is the proposal of $\mathcal{O}(m^3)$-complexity algorithms, able to greatly outperform MMSE and even provide close-to-optimal performance for small m and r. All the proposed methods belong to a class of approximate inference methods, originally proposed in the Machine Learning community, that seek to construct approximations to complex distributions using moment matching as fundamental criterion [35, 36].

The algorithms proposed in this Thesis iteratively approach the posterior distribution with a divide and conquer strategy. The algorithms approach the posterior distribution with $2m$ factors, natural parameters of gaussian distributions, and these are refined at each iteration of the algorithm. Those factors are chosen from a restricted family and then the complexity is determined by the model, as will be clarified in Chapters III and V. These moment propagation algorithms are supported by the Kullback-Leibler (KL) divergence between the two distributions [37, 38]. The larger number of moments which are taken into account, the closer the two distributions would be.

More specifically, the contributions are:

➔ First, Expectation Propagation (EP) is used to obtain a robust hard-symbol MIMO detector, denoted as Expectation Propagation Detector (EPD). The results show that EPD is close to the ML detector in those small scenarios and achieves remarkable gains with respect to other state-of-the-art algorithms in massive MIMO scenarios. These results have been published in a paper in the IEEE Transactions on Communications [39].

➔ The second main contribution is the analysis of the detection problem in terms of Expectation Consistency (EC) approximate inference, which is a generalization of other techniques such as EP itself or Belief Propagation (BP). An exhaustive study of the algorithm is performed, and probabilistic convergence methods are proposed. The EC point of view becomes esencial for this achievement, as EP is simply described as an iterative message-passing algorithm with a lack of information about the fundamental problem that now needs to be solved. A Expectation Consistency Detector (ECD) was used to fed the channel decoder with accurate symbol posterior probabilities, and the results show that in low SNR the proposed scheme achieves a transmission rate much closer to capacity than state-of-the-art methods. This results have been submitted to IEEE Transactions on Vehicular Technology[40].

Chapter II

Hard-Detection Methods

In this chapter a review of state-of-art methods for MIMO hard symbol detection is presented. Although, their approach is very different, all of them try to reduce the computational cost within the closest performance to the optimal receiver. For completeness, we include again System Model II.1.

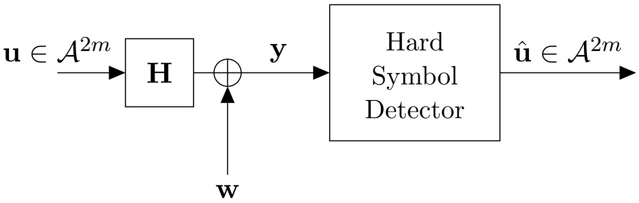

$\mathbf{u} \in \mathcal{A}^{2m}$
$\mathbf{H}$
$\mathbf{y}$
Hard Symbol Detector
$\hat{\mathbf{u}} \in \mathcal{A}^{2m}$
$\mathbf{w}$

Figure II.1: Hard Detection Scenario

II.1 Sphere Detection Methods

As it was previously introduced, the optimal detector is the MAP detector. However, for symbols with uniform prior to be transmitted, the ML detector is equivalent:

$$\begin{aligned}\hat{\mathbf{u}}_{\text{ML}} &= \arg \max_{\mathbf{u} \in \mathcal{A}^{2m}} p(\mathbf{y}|\mathbf{u}) \\ &= \arg \min_{\mathbf{u} \in \mathcal{A}^{2m}} ||\mathbf{y} - \mathbf{H}\mathbf{u}||^2. \qquad \text{(II.1)}\end{aligned}$$

The main target of Sphere Decoding (SpD) is to reduce the complexity inherent to the ML detector [41, 42] in (II.1). This algorithm looks for all lattice points belonging to an sphere with center in the received vector $\mathbf{y}$

and a given radius d. The idea of establishing a limitation on the number of points is clever but the problem is how to choose the radius of that sphere and the necessity of checking *how many?* points are within the sphere. This algorithm usually has a good performance on a medium-high SNR regime given that lattice points are close to the $\mathbf{y}$. The solution of interest belongs to $\mathcal{A}_d^{2m}$, which is the subset of symbols in $\mathcal{A}^{2m}$ that satisfy

$$\mathcal{A}_d^{2m} : \left\{ \mathbf{u} \in \mathcal{A}^{2m} : ||\mathbf{y} - \mathbf{H}\mathbf{u}||^2 \leq d^2 \right\}, \tag{II.2}$$

and, the SpD solution is

$$\hat{\mathbf{u}}_{\text{SD}} = \arg \max_{\mathbf{u} \in \mathcal{A}_d^{2m}} p(\mathbf{u}|\mathbf{y}) = \arg \min_{\mathbf{u} \in \mathcal{A}_d^{2m}} ||\mathbf{y} - \mathbf{H}\mathbf{u}||^2. \tag{II.3}$$

Note that $\hat{\mathbf{u}}_{\text{SD}} \neq \hat{\mathbf{u}}_{\text{ML}}$ may happen if the radius d very is small. A tradeoff between the radius of the sphere and the computational cost, usually depndent on SNR, must be considered for this kind of detector [43, 44]. At high SNR, the radius is taken small. At low SNR, the radius has to grow excesively and the SpD complexity becomes burdersome. There are several heuristics [42], but ultimately the complexity is $\mathcal{O}(M^{\alpha m})$ for $\alpha \in [0, 1]$.

II.2 Minimum Mean Squared Error

The MMSE approach is based on the assumption that $p(\mathbf{u}|\mathbf{y})$ can be approximated by a continuous Gaussian distribution within a quadratic cost function. In signal processing for communications the use of MMSE estimator is not new at all, as it has been used in many other applications such as Code Division Multiple Access (CDMA) [45, 46] CSI estimation in Orthogonal Frequency Division Multiplexing (OFDM) [47].

The MMSE posterior approximation is directly obtained by replacing the discrete uniform prior $p(\mathbf{u})$ in (I.5) by a zero-mean and E_s-variance independent Gaussian distribution:

$$\begin{aligned} p_{\text{MMSE}}(\mathbf{u}|\mathbf{y}) &\propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) p_{\text{MMSE}}(\mathbf{u}) \\ &= \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \prod_i^{2m} \mathcal{N}(u_i : 0, E_s). \end{aligned} \tag{II.4}$$

This approximation, a multivariate Gaussian distribution, is easy to maximize. The MMSE detector [20, 48] first computes the covariance matrix and the mean vector of $p_{\text{MMSE}}(\mathbf{u}|\mathbf{y})$,

$$\boldsymbol{\Sigma}_{\text{MMSE}} = \left(\mathbf{H}^{\top}\mathbf{H} + \frac{\sigma_w^2}{E_s}\mathbf{I}\right)^{-1} \tag{II.5}$$

$$\boldsymbol{\mu}_{\text{MMSE}} = \boldsymbol{\Sigma}_{\text{MMSE}}\mathbf{H}^{\top}\mathbf{y}, \tag{II.6}$$

where the complexity is dominated by the matrix inversion in (II.5), given by $\mathcal{O}(m^3)$. Since $p_{\text{MMSE}}(\mathbf{u}|\mathbf{y})$ is Gaussian distributed the mode and the mean coincide and a simple calculation shows that

$$\hat{\mathbf{u}}_{\text{MMSE}} = \max_{\mathbf{u}\in\mathcal{A}^{2m}} p_{\text{MMSE}}(\mathbf{u}|\mathbf{y}) = \mathbb{E}_{p_{\text{MMSE}}(\mathbf{u}|\mathbf{y})}[\mathbf{u}] = \boldsymbol{\mu}_{\text{MMSE}}. \tag{II.7}$$

Finally a component-wise hard decision is performed by projecting each component of $\boldsymbol{\mu}_{\text{MMSE}}$ into the corresponding QAM constellation:

$$\hat{u}_{i,\text{MMSE}} = \arg\min_{u_j\in\mathcal{A}} |u_j - \mu_{i,\text{MMSE}}|^2 \quad i \in [2m], \tag{II.8}$$

where the complexity of this step is $\mathcal{O}(m)$. The basic steps of the MMSE method are resumed in Algorithm 1, assuming perfect CSI at the receiver.

Algorithm 1 MMSE Algorithm

```
1) Compute $\boldsymbol{\Sigma}_{\text{MMSE}}$ from (II.5)
2) Compute $\boldsymbol{\mu}_{\text{MMSE}}$ from (II.6)
3) Project
$\hat{u}_{i,\text{MMSE}} = \arg\min_{u_j\in\mathcal{A}} |u_j - \mu_{i,\text{MMSE}}|^2 \quad i \in \{2m\}$
```

It should be noted that both SpD and MMSE reduce complexity, but they are based on two complete different approaches. Whereas SpD reduces the search space, MMSE is based on an statistical approximation to $p(\mathbf{u}|\mathbf{y})$.

II.3 MMSE with Successive Interference Cancellation

The MMSE detector does not provide a very good performance, because the multidimensional Gaussian approximation in (II.4) is not a sensible model for large MIMO systems with high-order constellations in which transmission from one antenna can be seen as interference to other antennas due to the superposition. The MMSE performance is significantly improved by successive interference cancellation, yielding the so-called MMSE with Successive Interference Cancellation (MMSE-SIC) [14, 49]. Iteratively, a decision is taken only over the vector symbol component with the smallest diagonal element in the covariance matrix in (II.5). Its effect is afterwards removed at the channel output. After each iteration (ℓ), the received vector is updated

$$\mathbf{y}^{(\ell+1)} = \mathbf{y}^{(\ell)} - \mathbf{h}_i^{(\ell)} \hat{u}_{i,\text{MMSE}}^{(\ell)} \tag{II.9}$$

$$\mathbf{H}^{(\ell+1)} = \mathbf{H}_{-i}^{(\ell)} \tag{II.10}$$

where $\mathbf{h}_i^{(\ell)}$ is the i-th column of $\mathbf{H}^{(\ell)}$ and, from (II.9), we see that the effect of the current decision $\hat{u}_{i,\text{MMSE}}^{(\ell)}$ is removed from the received vector. Afterwards $\mathbf{h}_i^{(\ell)}$ is removed from $\mathbf{H}^{(\ell)}$ obtaining $\mathbf{H}_{-i}^{(\ell)}$. In a nutshell, MMSE-SIC improves the MMSE detector, because only a one-dimensional Gaussian approximation per iteration is involved and the decision is over the component that has more certainty. Despite, MMSE-SIC requires to perform $2m$ times a MMSE matrix inversion similar to that in (II.5), the complexity can be lowered down to $\mathcal{O}(m^3)$ [49], by efficiently computing the matrix inversion at each iteration using the matrix inversion lemma (a rank-one update given the inverted matrix from the previous iteration).

The basic steps of the MMSE-SIC method are summarized in Algorithm 2.

Algorithm 2 MMSE-SIC Algorithm

```
Initialize $\ell = 0$
repeat
    1) Compute $\mathbf{\Sigma}_{\text{MMSE}}^{(\ell)}$ from (II.5)
    2) Compute $\boldsymbol{\mu}_{\text{MMSE}}^{(\ell)}$ from (II.6) using $\mathbf{\Sigma}_{\text{MMSE}}^{(\ell)}$ and $\mathbf{y}^{(\ell)}$
    3) Select $i = \arg\min_{i \in [2m-(\ell)]} \text{diag}(\mathbf{\Sigma}_{\text{MMSE}}^{(\ell)})$
    4) Project $\hat{u}_{i,\text{MMSE}} = \arg\min_{u_j \in \mathcal{A}} |u_j - \mu_{i,\text{MMSE}}|^2$
    5) Removing effect of $\hat{u}_{i,\text{MMSE}}$:
        5.1)    compute $\mathbf{y}^{(\ell+1)}$ and remove $y_i^{(\ell+1)}$
        5.2)    remove $\mathbf{h}_i$ from $\mathbf{H}$
until $\ell = 2m$
```

II.4 Gaussian Tree-Approximation

The Gaussian Tree-Approximation (GTA) algorithm in [50] constructs a tree-factorized approximation to the posterior distribution and relies on BP algorithm for estimating the marginal posterior distribution of the transmitted symbol, i.e. $p(u_i|\mathbf{y})$. GTA was first proposed in [50] as a feasible method to improve the MMSE-SIC solution for MIMO detection. GTA is based on the following idea: given the posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (I.5), the discrete nature of the prior $p(\mathbf{u})$ is first ignored and replaced by a non-informative prior:

$$\begin{aligned} p_{\text{n-i}}(\mathbf{u}|\mathbf{y}) &\propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \\ &= \mathcal{N}(\mathbf{u} : \boldsymbol{z}, \sigma_w^2 (\mathbf{H}^\top \mathbf{H})^{-1}), \end{aligned} \tag{II.11}$$

where $\boldsymbol{z} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{y}$. Consider the family of all possible Gaussian distributions with probability density functions that factorize according to a certain tree graph, i.e. any Gaussian distribution with pdf $g(\mathbf{u})$ such that

$$g(\mathbf{u}) = \prod_i g(u_i | u_{\Pi(i)}), \tag{II.12}$$

where $\Pi(i)$ is the set of parents of u_i and the associated factor graph is cycle-free. Now, GTA finds the distribution in such family that minimizes the KL divergence

$$g_{\text{GTA}}(\mathbf{u}) = \arg\min D_{\text{KL}}(p_{\text{n-i}}(\mathbf{u}|\mathbf{y}) || g(\mathbf{u})).$$

Provided that $p_{\text{n-i}}(\mathbf{u}|\mathbf{y})$ is also Gaussian, $g_{\text{GTA}}(\mathbf{u})$ is known in closed-form [50] and it can be computed at cost $\mathcal{O}(m^2)$. Finally, going back to the original posterior $p(\mathbf{u}|\mathbf{y})$ in (I.8) and replacing the Gaussian likelihood term by the Gaussian tree distribution $g_{\text{GTA}}(\mathbf{u})$ we have

$$p_{\text{GTA}}(\mathbf{u}|\mathbf{y}) \propto \prod_i g_{\text{GTA}}(u_i | u_{p(i)}) \prod_i \mathbb{I}_{u_i \in \mathcal{A}}. \tag{II.13}$$

Since $p_{\text{GTA}}(\mathbf{u}|\mathbf{y})$ is a tree factor graph, it is possible to use the BP algorithm to compute the symbol marginals that are then used for decision. BP over the factor graph $p_{\text{GTA}}(\mathbf{u}|\mathbf{y})$ has a complexity $\mathcal{O}(m^2|\mathcal{A}|^2)$. While the overall complexity is dominated by the matrix inversion $(\mathbf{H}^\top \mathbf{H})^{-1}$, the overhead incurred to compute the tree approximation $g_{\text{GTA}}(\mathbf{u}|\mathbf{y})$ and running BP is not negligible for typical-sized MIMO systems. While the GTA performance is similar to MMSE-SIC for low and medium SNR, GTA outperforms MMSE-SIC for high SNR and it has a significant lower computational complexity [50].

II.5 GTA with Successive Interference Cancellation

Recently, it has been shown in [51, 52] that successive interference cancellation substantially improves the GTA performance, in line with MMSE-SIC improvements. The procedure described before is repeated m times, since per iteration the decision is only over the symbol that has the least uncertainty and its effect is canceled from the system as in (II.9). Evaluating the m matrix inversions during GTA with Successive Interference Cancellation (GTA-SIC), using the techniques proposed in [49] for MMSE-SIC, requires $\mathcal{O}(m^3)$ iterations and performing m times the tree-factorized approximation and running BP have a cost of $\mathcal{O}(\sum_{k=1}^{m} k^2) \approx \mathcal{O}(m^3)$ operations for sufficiently large m. Results reported for GTA-SIC in [51] shows that it is able to outperform the best linear detectors for MIMO detection proposed in the literature in the past years, such as MMSE and MMSE-SIC with lattice reduction using the Lenstra-Lenstra-Lovász (LLL) algorithm [53, 54].

II.6 Lattice Reduction Techniques

Lattice Reduction (LR) techniques seek to construct an orthogonal lattice basis to boost the detection performance. Observe that the MIMO channel model

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \mathbf{w}$$

can be regarded, if we ignore the noise, as a lattice space spanned by $\mathbf{H}$ [55]. Reducing the channel matrix $\mathbf{H}$ using any LR algorithm consist on finding an integer unimodular matrix $\mathbf{T}$ so that $\mathbf{H}\mathbf{T} = \overline{\mathbf{H}}$ results into a nearly-orthogonal basis. Afterwards, detection is performed in the transformed space, namely

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{u} + \mathbf{w} = \overline{\mathbf{H}}\mathbf{T}^{-1}\mathbf{u} + \mathbf{w} \\ &= \overline{\mathbf{H}}\overline{\mathbf{u}} + \mathbf{w} \end{aligned} \tag{II.14}$$

There are several proposal to tune LR techniques, i.e to find the $\mathbf{T}$ matrix, for MIMO receivers [53, 54, 56, 57], even for very large number of antennas [58]. If $\mathbf{W}_{\text{LR}}^{\top}\mathbf{y}$ is the solution of a linear detector in the transformed space, then we get the desired solution as

$$\hat{\mathbf{u}} = \mathbf{T}\mathbf{W}_{\text{LR}}^{\top}\mathbf{y}, \tag{II.15}$$

where $\mathbf{W}_{\text{LR}}$ depends on the linear approach. For example, the LR-MMSE solution is

$$\mathbf{W}_{\text{LR-MMSE}} = (\overline{\mathbf{H}}^{\top}\overline{\mathbf{H}})^{-1}\overline{\mathbf{H}}^{\top} \tag{II.16}$$

$$\hat{\mathbf{u}}_{\text{LR-MMSE}} = \mathbf{T}\mathbf{W}_{\text{LR-MMSE}}^{\top}\mathbf{y}. \tag{II.17}$$

$$\tag{II.18}$$

The enhance in performance for linear MIMO detectors is notorious, also in those implementing SIC [24].

II.7 Tabu Search Detection

Tabu Search (TS) [59–61] is a greedy algorithm that iteratively looks for the point in a certain neighborhood that minimizes $||\mathbf{H}\hat{\mathbf{u}} - \mathbf{y}||^2$, i.e, the ML cost function. TS needs as input an initial point (which can be the solution of any of the MIMO detector presented), and certain parameters that balance the complexity-performance tradeoff, namely the neighborhood distance per iteration or the size of the list of past candidates that we need to store to avoid checking points that were the selected candidate in the past. More precisely, if $\mathbf{u}_p$ is the current point, we have to evaluate

$$f(\mathbf{u}) = ||\mathbf{H}\mathbf{u} - \mathbf{y}||^2, \forall \mathbf{u} \in N(\mathbf{u}_p), \tag{II.19}$$

where $N(\mathbf{u}_p)$ is the set of neighbors to be considered at this iteration. Given a max symbol-Hamming distance Δ, then

$$N(\mathbf{u}_p) : \left\{\mathbf{u} \in \mathcal{A}^{2m} : \mathrm{d_H}(\mathbf{u}_p - \mathbf{u}) \leq \Delta\right\}. \tag{II.20}$$

To keep complexity small, Δ is usually taken to 1. In this case the max number of neighbors to be considered is given in Table II.1.

QAM Costellation	Number of Neighbors
$M = 4$	$2m$
$M \geq 16$	$4m$

Table II.1: Maximum Number of neighbors for $\Delta = 1$.

TS is a hard-detection algorithm, that needs a initial estimation to perform. In the literature, MMSE has been traditionally used to initialize TS. The main steps of the resulting algorithm are summarized in Algorithm 3 Obviously the complexity of the algorithm is reduced respect to ML or SD but heavely depends on the number of iterations and the size of the tabu list. A proper stopping rule along with a good initial solution may reduce the complexity and still keep the close-to-ML performance [59–61] in small scenarios.

II.8 Comparison of State-of-the-art Methods

In Figure II.2 we consider a small scenario in which the optimal detector can be computed. The SER performance is shown for a $m = r = 6$ scenario and a 4-QAM constellation. It is possible to observe that MMSE detector is remarkally improved by the successive interference cancellation technique and also that such a small scenario SpD has a really close performance to the optimal detector.

Algorithm 3 TS MIMO receiver

```
$\ell = 0$, MMSE as initial candidate $\mathbf{u}_p^{(0)}$.
Initialize Tabu list $L = \{\}$
repeat
    1) Compute the neighborhood $N(\mathbf{u}_p^{(\ell)})$ according to (II.20), excluding those in the tabu list $L$.
    2) Compute the cost of the neighborhood $f(\mathbf{u})\ \forall \mathbf{u} \in N(\mathbf{u}_p^{(\ell)}$ according to (II.19).
    3) Choose the next candidate
        $$\mathbf{u}_p^{(\ell+1)} = \arg \min_{\mathbf{u} \in N(\mathbf{u}_p^{(\ell)})} f(\mathbf{u}) \qquad \text{(II.21)}$$
    4) Introduce $\mathbf{u}_p^{(\ell)}$ in the tabu list $L$, if there is no space remove the worst candidate.
until Max $\ell$
```

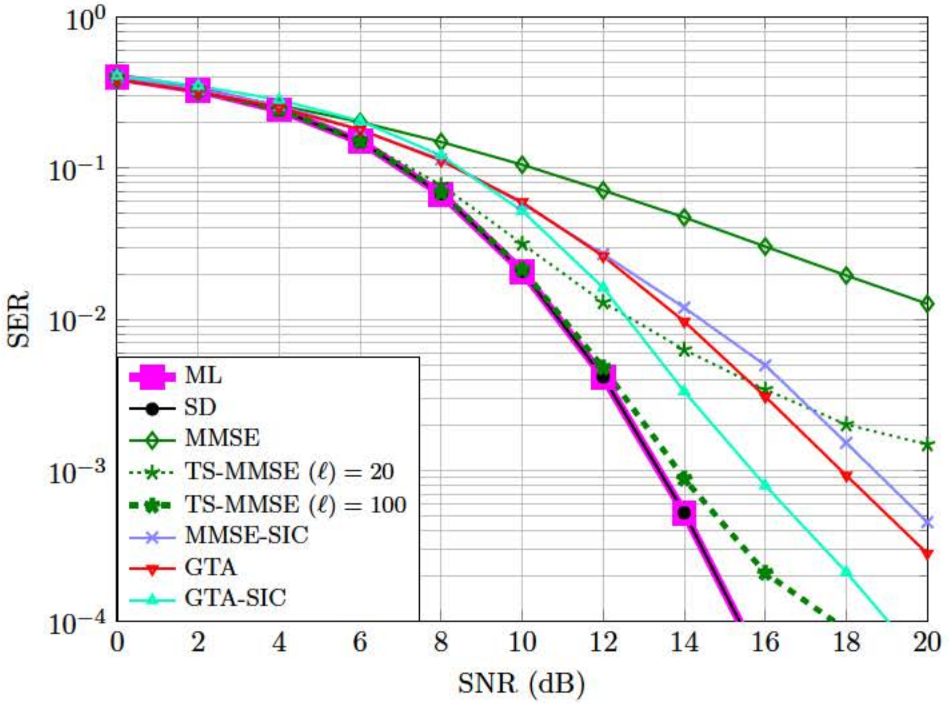


Figure II.2: SER performance of several detectors in a scenario with $m = r = 6$ and a 4-QAM constellation.

Chapter III

Hard-Detection via EP Approximations

In this Chapter, we first present EP approximate inference and we review its application to communication scenarios in the literature. Then, a hard-decision MIMO receiver based on EP is introduced, and its complexity analyzed. Finally, we include an exhaustive comparison with some of the state-of-the-art methods for MIMO hard detection described in the former chapter. EP was first introduced by Minka in his Phd Thesis [37, 38, 62, 63]. It is a technique in Bayesian machine learning to construct tractable approximations to a given probability distribution. EP generalizes and combines two different techniques to construct the approximation. First, the assumed-density filter [64], which is also based on moment-matching approximations but does not take into account the graphical model structure of the true distribution, and second loopy BP [65].

The use of EP inference in communications is not new. For instance, it has been applied to LDPC channel decoding for the Binary Erasure Channel (BEC) in [66], and then for general Binary Memoryless Symmetric Channel (BMSC) in [67]. In this scenario, EP-based algorithms replace the standard BP algorithm at the decoding state, maintaining the same complexity, but enhancing the error correction capability. Another application version of the EP algorithm for communications can be found in [68], in this case as a channel equalizer for single user Inter Symbol Interference (ISI) channels, in which the EP probabilistic output is used to replace the BCJR algorithm, which becomes unfeasiable for high order modulations. The first application of EP for MIMO communications was firstly introduced in [69]. In this work, the posterior distribution of the transmitted symbols is approximated by a fully factorized Gaussian distribution, which neglects correlations imposed by the channel likelihood, and dramatically affects the receiver performance.

The proposal in this Thesis improves and extends this scenario by showing that the EP iterative method can easily tackle with the channel likelihood, obtaining this way accurate approximations that take into account the correlations imposed by the channel output.

The main idea behind the EP strategy for MIMO detection is to approximate the overall distribution by a real-valued Gaussian distribution. Iteratively, discrete symbol priors are introduced in the approximation and its moments are corrected according to such information. The ultimate goal is to converge to a Gaussian distribution whose moments are close to those of the true posterior distribution of the transmitted symbols. Note that for hard detection, only the mode of the approximation is relevant.

III.1 Expectation Propagation Approximate Inference

A brief introduction to EP for graphical models and exponential family distributions[1] is first presented. This description follows essentially [38] and [65]. Suppose we are given some statistical model with δ latent variables, $\mathbf{u} \in \Omega^{\delta}$, that factors in the following way

$$p(\mathbf{u}) \propto f(\mathbf{u}) \prod_{i=1}^{I} t_i(\mathbf{u}), \tag{III.1}$$

where $f(\mathbf{u})$ belongs to an exponential family $\mathcal{F}$ with sufficient statistics $\Phi(\mathbf{u}) = [\phi_1(\mathbf{u}), \phi_2(\mathbf{u}), \ldots, \phi_S(\mathbf{u})]$ and $t_i(\mathbf{u})$, $\forall i \in \{I\}$, are non-negative factors. For instance, if $\mathcal{F}$ is the multivariate Gaussian family, then $\Phi(\mathbf{u}) = \{u_i, u_i u_j\}_{i,j=1}^{\delta}$. Assume now that performing inference over the distribution $p(\mathbf{u})$ in (III.1) is analytically intractable or prohibitively complex. In this scenario, EP provides a general-purpose framework to construct a tractable approximation to $p(\mathbf{u})$ by a distribution $q(\mathbf{u})$ from $\mathcal{F}$. The resemblance between $q(\mathbf{u})$ and $p(\mathbf{u})$ is achieved by designing $q(\mathbf{u})$ such that

$$\mathbb{E}_{q(\mathbf{u})}[\phi_j(\mathbf{u})] = \mathbb{E}_{p(\mathbf{u})}[\phi_j(\mathbf{u})] \qquad \forall j \in \{S\}\,. \tag{III.2}$$

Equation (III.2) is known as the *moment matching* condition. When both $q(\mathbf{u})$ and $p(\mathbf{u})$ are defined over the same support space and measure, the moment matching condition in (III.2) is equivalent to finding $q(\mathbf{u})$ in $\mathcal{F}$ that minimizes the KL divergence with $p(\mathbf{u})$, i.e.

$$q(\mathbf{u}) = \arg \min_{q'(\mathbf{u}) \in \mathcal{F}} D_{\mathrm{KL}}(p(\mathbf{u}) || q'(\mathbf{u})). \tag{III.3}$$

[1] A comprehensive introduction to exponential families and their properties can be found in [65].

One naïve approach to find $q(\mathbf{u})$ would be to first compute the moments $\mathbb{E}_{p(\mathbf{u})}[\phi_j(\mathbf{u})]\ \forall j \in \{S\}$ and second to construct $q(\mathbf{u})$ according to them. However by assumption, this is not a viable option since we cannot do inference over $p(\mathbf{u})$. To overcome this problem, Minka proposed a sequential EP algorithm to iteratively approach the solution in (III.2) at polynomial time complexity [37, 70]. The main idea behind the sequential EP algorithm is the fact that, while performing inference over $p(\mathbf{u})$ in (III.1) is intractable, many times we are able to perform tractable inference over a distribution of the form

$$\hat{p}_i(\mathbf{u}) \propto f(\mathbf{u}) t_i(\mathbf{u}), \tag{III.4}$$

in which there is only present one of the factors $t_i(\mathbf{u})\ \forall i \in \{I\}$ in (III.1) that do not belong to the exponential family $\mathcal{F}$. The sequential EP algorithm is as follows. First, assume the following factorization for $q(\mathbf{u}) \in \mathcal{F}$,

$$q(\mathbf{u}) = f(\mathbf{u}) \prod_{i=1}^{I} \tilde{t}_i(\mathbf{u}), \tag{III.5}$$

where $\tilde{t}_i(\mathbf{u}) \in \mathcal{F}$, $\forall i \in \{I\}$. Note that we have simply replaced each one of the $t_i(\mathbf{u})$ factors in (III.1) by a member $\tilde{t}_i(\mathbf{u})$ of $\mathcal{F}$. Given an initial proposal $q^{(0)}(\mathbf{u})$ and being $q^{(\ell)}(\mathbf{u})$ the approximation to $q(\mathbf{u})$ in (III.3) at iteration ℓ, $q^{(\ell+1)}(\mathbf{u})$ is obtained by updating each one of the $\tilde{t}_i(\mathbf{u})$ factors independently. For all $i \in \{I\}$,

1. Compute the *cavity* distribution

$$q^{(\ell)\backslash i}(\mathbf{u}) \doteq \frac{q^{(\ell)}(\mathbf{u})}{\tilde{t}_i(\mathbf{u})} \in \mathcal{F}. \tag{III.6}$$

2. Compute the distribution $\hat{p}_i(\mathbf{u}) \propto t_i(\mathbf{u}) q^{(\ell)\backslash i}(\mathbf{u})$, and find

$$\mathbb{E}_{\hat{p}_i(\mathbf{u})}[\phi_s(\mathbf{u})]\ \ \forall s \in \{S\}\,. \tag{III.7}$$

3. The refined factor $\tilde{t}_i^{\text{new}}(\mathbf{u})$ is obtained so that

$$\mathbb{E}_{\tilde{t}_i^{\text{new}}(\mathbf{u}) q^{(\ell)\backslash i}(\mathbf{u})}[\phi_s(\mathbf{u})] \tag{III.8}$$

coincides with (III.7) $\forall s \in \{S\}$, i.e,

$$\mathbb{E}_{\tilde{t}_i^{\text{new}}(\mathbf{u}) q^{(\ell)\backslash i}(\mathbf{u})}[\phi_s(\mathbf{u})] = \mathbb{E}_{\hat{p}_i(\mathbf{u})}[\phi_s(\mathbf{u})]. \tag{III.9}$$

The sequential EP algorithm is run until a convergence criterion is met or a maximum number of iterations is reached. As shown in [63], this algorithm can be interpreted as a coordinate gradient descent over the parameter space of the $q(\mathbf{u})$ distribution to find a saddle point of a certain energy function. As such, the convergence to a saddle point is not guaranteed [71]. Nonetheless, sequential EP has been shown to achieve accurate results, typically close to the moment matching solution, in a wide range of applications [37, 62].

As shown in [38, 62], if a factor $t_i(\cdot)$ in (III.1) only depends on a subset $\mathbf{u}_i$ of $\mathbf{u}$ with dimension $\delta_i < \delta$, then the approximate factor $\tilde{t}_i(\mathbf{u}_i)$ in (III.6) is defined over the same domain and its update at each iteration can be alternatively performed over the marginal distribution $q(\mathbf{u}_i)$. An example of this alternative procedure is the EP approximation to the MIMO symbol posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (I.8) that we present in detail in the next section.

III.2 Expectation Propagation Detection

The MMSE approximation to the true posterior distribution in (II.4) replaces the prior over the transmitted symbols by a zero-mean independent component-wise Gaussian whose variance equals the QAM symbol mean energy. Intuitively it might make sense to chose the parameters of the Gaussian prior this way, because it matches the first two moments of the input distribution. However it is certainly not the best choice, as we are interested in matching the posterior distribution to optimally detect the transmitted symbols. The EP approximation to the posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (I.8) is constructed by replacing the prior input distribution by an independent Gaussian distribution:

$$p_{\mathrm{EP}}(\mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2 \mathbf{I}) \prod_{i=1}^{2m} \mathrm{e}^{\gamma_i u_i - \frac{1}{2}\Lambda_i u_i^2}, \tag{III.10}$$

where γ_i and $\Lambda_i > 0$ are real valued parameters that have to be adjusted. Note that, taking $\gamma_i = 0$ and $\Lambda_i = E_s^{-1}$ $\forall i \in \{2m\}$, then (III.10) matchs with the MMSE approximation to the posterior distribution $p(\mathbf{u}|\mathbf{y})$ [20, 27]. For arbitrary vectors $\boldsymbol{\gamma} \in \mathbb{R}^{2m}$ and $\boldsymbol{\Lambda} \in \mathbb{R}_+^{2m}$, $p_{\mathrm{EP}}(\mathbf{u}|\mathbf{y})$ is a Gaussian with mean vector $\boldsymbol{\mu}_{\mathrm{EP}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathrm{EP}}$, where

$$\boldsymbol{\Sigma}_{\mathrm{EP}} = \left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \mathrm{diag}\left(\boldsymbol{\Lambda}\right)\right)^{-1}, \tag{III.11}$$

$$\boldsymbol{\mu}_{\mathrm{EP}} = \boldsymbol{\Sigma}_{\mathrm{EP}}\left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{y} + \boldsymbol{\gamma}\right). \tag{III.12}$$

For each channel observation $\mathbf{y}$, the goal is to choose $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$ so that

$$\boldsymbol{\mu}_{\text{EP}} \to \mathbb{E}_{p(\mathbf{u}|\mathbf{y})}[\mathbf{u}], \tag{III.13}$$

$$\boldsymbol{\Sigma}_{\text{EP}} \to \mathbb{E}_{p(\mathbf{u}|\mathbf{y})}[(\mathbf{u} - \mathbb{E}_{p(\mathbf{u}|\mathbf{y})}[\mathbf{u}])(\mathbf{u} - \mathbb{E}_{p(\mathbf{u}|\mathbf{y})}[\mathbf{u}])^\top]. \tag{III.14}$$

This condition is known as moment-matching (MM). While the direct computation of the $p(\mathbf{u}|\mathbf{y})$ moments requires $|\mathcal{A}|^{2m}$ operations, the EP algorithm iteratively approachs the solution in (III.13) and (III.14) at polynomial-time complexity [37, 70], as we show in the next Section. Once the iterative method has stopped, either by convergence or maximum number of iterations reached, the EPD computes the output independently projecting each component:

$$\hat{u}_{i,\text{EP}} = \arg\min_{u_j \in \mathcal{A}} |u_j - \mu_{i,\text{EP}}|^2 \quad \forall i \in \{2m\}\,. \tag{III.15}$$

III.3 EP MIMO Detector Updates

The formulation of the EP is presented according to the update rules in [38, 62], and this algorithm is denoted by EP MIMO receiver, EPD for short. The EPD iterative method approximates the solution in (III.13) and (III.14) at polynomial complexity by recursively updating the pairs $(\gamma_i, \Lambda_i), \forall i \in \{2m\}$. After initializing $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$ accordingly to the MMSE solution, the pairs $\left(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)}\right)$, $\forall i \in \{2m\}$, are updated in parallel, where ℓ denotes the EPD iteration. Given the i-th marginal of the distribution $p_{\text{EP}}(\mathbf{u}|\mathbf{y})$ in (III.10) at iteration ℓ, namely $p_{\text{EP}}^{(\ell)}(u_i|\mathbf{y}) = \mathcal{N}(u_i : \mu_i^{(\ell)}, \sigma_i^{2(\ell)})$, the pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ is computed as follows:

1. Compute the cavity marginal

$$p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y}) \propto \frac{p_{\text{EP}}^{(\ell)}(u_i|\mathbf{y})}{\exp(\gamma_i^{(\ell)} u_i - \frac{1}{2}\Lambda_i^{(\ell)} u_i^2)} = \mathcal{N}(u_i : t_i^{(\ell)}, h_i^{2(\ell)}), \tag{III.16}$$

   where the mean $t_i^{(\ell)}$ and the variance $h_i^{2(\ell)}$ are computed as follows

$$h_i^{2(\ell)} = \frac{\sigma_i^{(\ell)}}{1 - \sigma_i^{(\ell)}\Lambda_i^{(\ell)}}\ ,\ t_i^{(\ell)} = h_i^{2(\ell)}\left(\frac{\mu_i^{(\ell)}}{\sigma_i^{(\ell)}} - \gamma_i^{(\ell)}\right). \tag{III.17}$$

2. Introduce the true discrete factor $\mathbb{I}_{u_i \in \mathcal{A}}$ in $p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y})$ obtaining $\hat{p}_i(u_i)$

$$\hat{p}_i(u_i) = p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y})\mathbb{I}_{u_i \in \mathcal{A}}, \tag{III.18}$$

and compute $\mathbb{E}_{\hat{p}_i(u_i)}[u_i^2]$ and $\mathbb{E}_{\hat{p}_i(u_i)}[u_i]$

$$\sigma^{2(\ell)}_{\hat{p}_i^{(\ell)}} = \mathbb{E}_{\hat{p}_i(u_i)}[u_i^2]\ ,\ \mu^{(\ell)}_{\hat{p}_i^{(\ell)}} = \mathbb{E}_{\hat{p}_i(u_i)}[u_i]. \tag{III.19}$$

3. Update the pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ so that the following unnormalized Gaussian distribution

$$p^{(\ell)\backslash i}(u_i)\exp(\gamma_i^{(\ell+1)}u_i - \frac{1}{2}\Lambda_i^{(\ell+1)}u_i^2), \tag{III.20}$$

has mean and variance equal to $\mu^{(\ell)}_{\hat{p}_i}$ and $\sigma^{2(\ell)}_{\hat{p}_i}$. A simple calculation shows that the solution is given by:

$$\Lambda_i^{(\ell+1)} = \frac{1}{\sigma^{2(\ell)}_{\hat{p}_i}} - \frac{1}{h_i^{2(\ell)}}, \quad \gamma_i^{(\ell+1)} = \frac{\mu^{(\ell)}_{\hat{p}_i}}{\sigma^{2(\ell)}_{\hat{p}_i}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}}. \tag{III.21}$$

Note that the distribution $\hat{p}_i$ in (III.18) can be seen as a refined approximation to the true marginal, replacing the prior term $\exp(\gamma_i^{(\ell)}u_i - \frac{1}{2}\Lambda_i^{(\ell)}u_i^2)$ by the true one, $\mathbb{I}_{u_i\in\mathcal{A}_i}$. On the other hand, the parameter updated in (III.21) may return a negative $\Lambda_i^{(\ell+1)}$, which should be positive, because it is a precision (inverse variance) term. This result just means that there is no pair $(\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ that places the variance of the Gaussian in (III.20) at $\sigma^{2(\ell)}_{\hat{p}_i}$. In that case, the previous values are kept for those parameters, i.e. $\gamma_i^{(\ell+1)} = \gamma_i^{(\ell)}$ and $\Lambda_i^{(\ell+1)} = \Lambda_i^{(\ell)}$, and update all the other pairs, $(\gamma_j^{(\ell+1)}, \Lambda_j^{(\ell+1)})$ for $j \neq i$. Finally, to improve the robustness of the algorithm, in [38, 72] it is suggested to smooth the parameter update (i.e., a low-pass filter) in (III.21) by a convex combination with the former value, namely

$$\gamma_i^{(\ell+1)} = \beta\left(\frac{\mu^{(\ell)}_{\hat{p}_i}}{\sigma^{2(\ell)}_{\hat{p}_i}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}}\right) + (1-\beta)\gamma_i^{(\ell)}, \tag{III.22}$$

$$\Lambda_i^{(\ell+1)} = \beta\left(\frac{1}{\sigma^{2(\ell)}_{\hat{p}_i}} - \frac{1}{h_i^{2(\ell)}}\right) + (1-\beta)\Lambda_i^{(\ell)}, \tag{III.23}$$

for some damping parameter $\beta \in [0, 1]$. Due to its simplicity, smoothing the parameter updates as in (III.22)-(III.23) is a fairly common technique to stabilize approximate inference iterative algorithms. To ensure numerical stability, a fairly small value of β is chosen ($\beta = 0.2$ in our experiments for hard-detection in the high-SNR regimen). Also, trying to avoid numerical instabilities, a minimum value in the variance per component is set as:

$$\sigma_{\hat{p}_i}^{2(\ell)} = \max(5 \times 10^{-7}, \mathbb{E}_{\hat{p}_i(u_i)}[u_i^2]). \tag{III.24}$$

All the steps of the algorithm are summarized in Algorithm 4. Convergence is assumed when parameters differ less than 10^{-4} between two consecutive iteration. Alternatively, experimental results suggest that a maximum number of $I = 10$ iterations is enough to achieve robust hard-detection performance at high SNR.

Algorithm 4 MIMO EPD

```
Fix a damping factor $\beta$.
Initialize $\ell = 1$, $\boldsymbol{\gamma}^{(0)} = \mathbf{0}$ and $\boldsymbol{\Lambda}^{(0)} = E_s^{-1}$.
repeat
    1) Given $\boldsymbol{\gamma}^{(\ell-1)}, \boldsymbol{\Lambda}^{(\ell-1)}$, compute $\boldsymbol{\Sigma}_{\text{EP}}^{(\ell)}$, $\boldsymbol{\sigma}_{\text{EP}}^{(\ell)} = \text{diag}(\boldsymbol{\Sigma}_{\text{EP}}^{(\ell)})$ and $\boldsymbol{\mu}_{\text{EP}}^{(\ell)}$ following
    (III.11) and (III.12).
    2) For $i \in \{2m\}$ compute the cavity marginals and their moments
        $p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y}) \propto \frac{p_{\text{EP}}^{(\ell)}(u_i|\mathbf{y})}{\exp(\gamma_i^{(\ell)} u_i - \frac{1}{2}\Lambda_i^{(\ell)} u_i^2)} = \mathcal{N}(u_i : t_i^{(\ell)}, h_i^{2(\ell)})$
        $h_i^{2(\ell)} = \frac{\sigma_{i\text{EP}}^{(\ell)}}{1 - \sigma_{i\text{EP}}^{(\ell)} \Lambda_i^{(\ell)}}, \quad t_i^{(\ell)} = h_i^{2(\ell)} \left( \frac{\mu_{i\text{EP}}^{(\ell)}}{\sigma_{i\text{EP}}^{(\ell)}} - \gamma_i^{(\ell)} \right)$
    3) Introduce the true discrete factor $\mathbb{I}_{u_i \in \mathcal{A}}$ in $p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y})$, obtaining $\hat{p}_i^{(\ell)}$ and
    compute $\mathbb{E}_{\hat{p}_i^{(\ell)}(u_i)}[u_i^2]$ and $\mathbb{E}_{\hat{p}_i^{(\ell)}(u_i)}[u_i]$.
    4) Update
        $\Lambda_i^{(\ell+1)} = \beta \left( \frac{1}{\sigma_{\hat{p}_i}^{2(\ell)}} - \frac{1}{h_i^{2(\ell)}} \right) + (1 - \beta)\Lambda_i^{(\ell)}$
        $\gamma_i^{(\ell+1)} = \beta \left( \frac{\mu_{\hat{p}_i}^{(\ell)}}{\sigma_{\hat{p}_i}^{2(\ell)}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}} \right) + (1 - \beta)\gamma_i^{(\ell)}$
    for all $i \in \{2m\}$
    5) $\ell = \ell + 1$
until convergence (or stop criterion)
```

III.4 EPD complexity

The complexity of EPD per iteration is dominated by the computation of the covariance matrix inversion in (III.11) and the mean vector in (III.12). Note that the complexity of this step is identical to the MMSE and GTA posterior covariance matrix computation and mean vector in, respectively, (II.7) and (II.11). Once the marginals moments $\boldsymbol{\mu}_{\text{EP}}$, $\boldsymbol{\sigma}_{\text{EP}}$ have been computed, the parallel update of all pairs $(\gamma_i^\ell, \Lambda_i^\ell) \leftarrow (\gamma_i^{(\ell+1)}, \Lambda_i^{(\ell+1)})$ $\forall i \in \{2m\}$ has a small computational complexity, linear in $m|\mathcal{A}|$. Thus, if EPD is run for I iterations, the final complexity is $\mathcal{O}(I(m^3+m|\mathcal{A}|+m))$. The comparison of the EPD complexity, $\mathcal{O}(I(m^3+m|\mathcal{A}|+m))$, with the complexity of GTA-SIC and MMSE-SIC depends on the channel time varying characteristics:

- In a quasi-static block fading channel model where the fading coefficients do not change within one time symbol, but vary every symbol time [73], the EPD complexity with $I = 10$ iterations is essentially ten times the MMSE complexity and thus comparable with the MMSE-SIC and GTA-SIC complexities.

- In a static block fading channel where the channel matrix $\mathbf{H}$ is constant during T consecutive symbol times, the MMSE-SIC matrix inversion only has to be computed once and thus the complexity of detecting the T blocks of m symbols is given by $\mathcal{O}(m^3+Tm^2)$ [4]. The computation of the tree approximations in (II.13) for GTA-SIC has to be done for each channel observation $\mathbf{y}$ and thus, the complexity to detect the T blocks of m symbols is $\mathcal{O}(Tm^3)$. Similarly, all the EPD processing depends on the channel observation vector and its complexity is $\mathcal{O}(TI(m^3 + m|\mathcal{A}| + m))$, still of the same order that of GTA-SIC.

Experimental results indicate that the typical number of iterations required to converge is $I = 10$, and with this number of iterations EPD guarantees and excellent performance and there is not further improvement by increasing the iterations beyond that level. More importantly, this happens regardless the number of antennas m or constellation order M, which represents a huge save in computational complexity. Indeed, for a low dimension scenario, where the moments of the true posterior $p(\mathbf{u}|\mathbf{y})$ in (I.8) can be computed, it is possible to observe that EPD typically converges to the right moments in a few iterations. In Figure III.1 we show the evolution of the components of the EPD mean vector $\boldsymbol{\mu}^{(\ell)}$ in (III.12) as EPD iterates for a given channel observation $\mathbf{y}$ in a $m = r = 2$ scenario with a 256-QAM constellation and SNR = 15 dB. Note that the MMSE estimate would be the EPD solution at iteration 1. In dashed lines, the mean of the posterior $p(\mathbf{u}|\mathbf{y})$ (real and imaginary parts) are also represented. After 10 iterations, the EPD provides an accurate estimate of the posterior mean of the posterior distribution $p(\mathbf{u}|\mathbf{y})$ in (I.8). For the same scenario, Figure III.2 shows an example of the evolution of the diagonal components of the EPD covariance matrix $\boldsymbol{\Sigma}^{(\ell)}$ in (III.11) as EPD iterates. In dashed lines, the real and

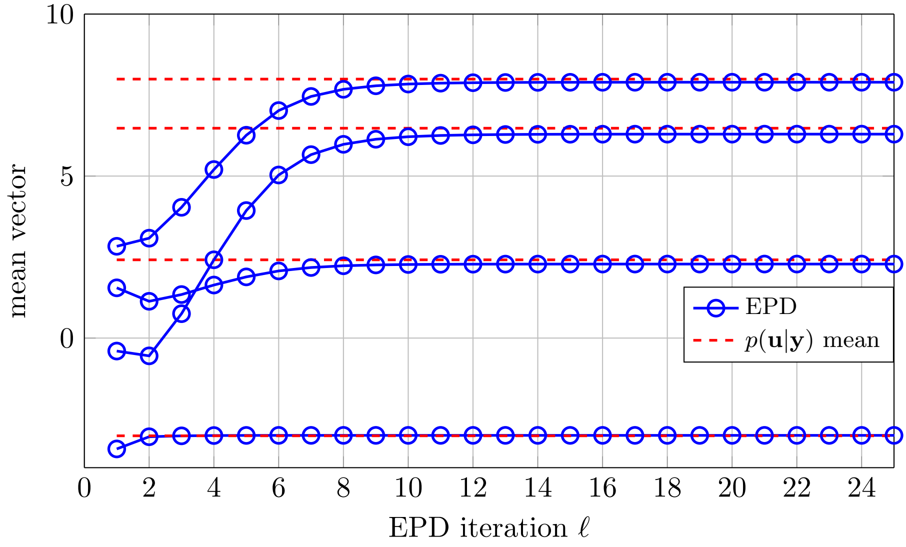


Figure III.1: Evolution of each component of the EPD mean $\boldsymbol{\mu}^{(\ell)}$ in (III.12) as EPD iterates for a $m = r = 2$ scenario with a 256-QAM constellation and SNR $= 15$ dB.

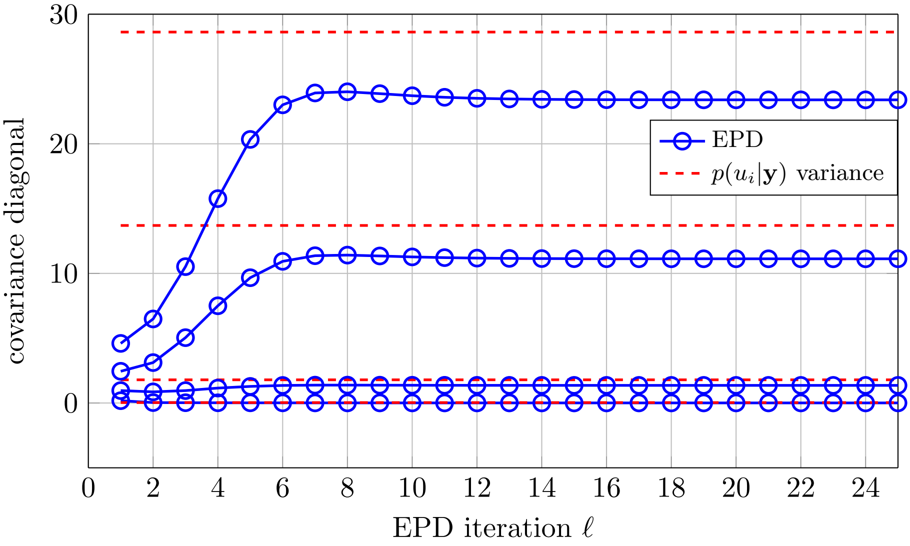


Figure III.2: Evolution of each component of the EPD covariance $\boldsymbol{\Sigma}^{(\ell)}$ in (III.11) as EPD iterates for a $m = r = 2$ scenario with a 256-QAM constellation and SNR $= 15$ dB.

imaginary values of the variance of the marginal symbol posterior $p(u_i|\mathbf{y})$ are shown. EPD, besides accurately matching the posterior mean, provides a reliable measure of the uncertainty per symbol, identifying which symbols can be decided with high grade of confidence and for which ones the risk of error in hard decision is large.

III.5 EP combined with LLL

The inclusion of lattice reduction techniques in EPD algorithm is explored by means of the LLL algorithm [58], denoted by EPD with Lenstra-Lenstra-Lovász (EPD-LLL). It is important to remark that LLL techniques shortens the gap between ML and Linear Detector (LD) performance by means of constructing a more orthogonal basis to the detector. This new basis increases the overlapping between the LD decision region and the ML decision region, thus improves performance [57, 74]. As discussed in Section II.6, this implementation is based on obtaining first a reduced lattice basis for the channel matrix $\mathbf{H}$, $\overline{\mathbf{H}} = \mathbf{H}\mathbf{T}$, where $\mathbf{T}$ is a uni-modular matrix such that all the matrix entries of $\mathbf{T}$ and $\mathbf{T}^{-1}$ are integers.

In order to combine EPD detection with LR techniques, we only have to tailor the EPD formulation provided above to the LR-reduced model

$$\mathbf{y} = \overline{\mathbf{H}}\overline{\mathbf{u}} + \mathbf{w} \tag{III.25}$$

or the equivalently,

$$\overline{\mathbf{y}} = \overline{\mathbf{H}}^{\top}\mathbf{y} = \mathbf{T}^{-1}\mathbf{u} + \overline{\mathbf{H}}\mathbf{w} \tag{III.26}$$

According to (III.26), we obtain similar expressions than the EP formulation in Section III.2

$$p_{\text{EP-LLL}}(\mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\overline{\mathbf{y}} : \mathbf{T}^{-1}\mathbf{u}, \sigma_w^2 \overline{\mathbf{H}}^{\top}\overline{\mathbf{H}}) \prod_{i=1}^{2m} \mathrm{e}^{\gamma_i u_i - \frac{1}{2}\Lambda_i u_i^2}, \tag{III.27}$$

$$\mathbf{\Sigma}_{\text{EP-LLL}} = \left(\sigma_w^{-2}\mathbf{T}^{-1}\left(\overline{\mathbf{H}}^{\top}\overline{\mathbf{H}}\right)^{-1}\mathbf{T}^{-1} + \operatorname{diag}(\mathbf{\Lambda})\right)^{-1}, \tag{III.28}$$

$$\boldsymbol{\mu}_{\text{EP-LLL}} = \mathbf{\Sigma}_{\text{EP-LLL}}\left(\sigma_w^{-2}\mathbf{T}^{-1}\left(\overline{\mathbf{H}}^{\top}\overline{\mathbf{H}}\right)^{-1}\overline{\mathbf{y}} + \gamma\right). \tag{III.29}$$

Under this new signal model, and given the new equations for $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$, the rest of EPD algorithm would formulate similarly. Thus, it is possible to apply LLL to EPD, the same way as it could have been applied to GTA or GTA-SIC.

III.6 Experimental Results

In this section, the performance of the EPD as hard-symbol MIMO detector for several scenarios in high-order high-dimensional scenarios is studied. The results are averaged for $1.5 \cdot 10^4$ realizations of the channel matrix. The considered scenarios are summarized in Table III.1. The EPD parameters are set to $\beta = 0.2$, $I = 10$ iterations, and a minimum value in the variance according to (III.24). The detector performance is shown in terms of SER as a function of SNR, which was defined in (I.2).

Figure	m	r	M-QAM	Detectors
III.3	6	6	4	EPD, ML, SpD, MMSE,TS-MMSE, GTA, GTA-SIC
III.4	12	12	16	EPD, EPD-LLL,TS-EPD, SpD, GTA, GTA-SIC
III.5	32	32	16	EPD, MMSE, GTA, GTA-SIC
III.6	64	64	16	EPD, MMSE, GTA, GTA-SIC
III.7	64	64	64	EPD, MMSE, GTA, GTA-SIC
III.8	100	100	16	EPD, MMSE, GTA, GTA-SIC
III.9	250	250	16	EPD, GTA-SIC

Table III.1: Simulated Hard-Detection Scenarios

Figure III.3 reproduces the scenario already considered in Figure II.2, now including EPD. Observe EPD is very close to ML and SpD for low-SNR regime. The gain with respect to the low complexity receivers such as MMSE or GTA is more than 2dB at 10^{-3} and performs similarly to GTA-SIC. We need to remark the good behavior of TS-MMSE in comparison with MMSE, outperforming more than 8dB at 10^{-2}. The only methods that outperform our proposal at 10^{-3} are ML and SpD by 1.75dB, and TS-MMSE by 1dB. However, their complexity grows fast with both M and m.

Figure III.4 considers a more complex scenario, where more antennas are used at both transmitter and receiver ($m = r = 12$), and also the order of the constellation is raised to 16-QAM. Both EPD and EPD-LLL are considered. As we can observe, both algorithms perform equally, thus indicating that the use of a more orthogonal protection to perform symbol detection does not bring any particular advantage for MIMO hard detection with moment-matching EP techniques. Similar results have been obtained in different scenarios, and this is the main reason why EPD-LLL is no longer included in following results. Since ML detection is prohibitively complex in this scenario, we could only run SpD detection as benchmark. Observe that the EPD gain with respect to GTA-SIC and GTA are about 1dB and 5dB respectively. Also, compared with Figure III.3, the crossing between EPD and GTA-SIC is produced at a lower SER point. Additionally, TS initialized with the EPD solution is included. Observed that TS-EPD improves EPD's

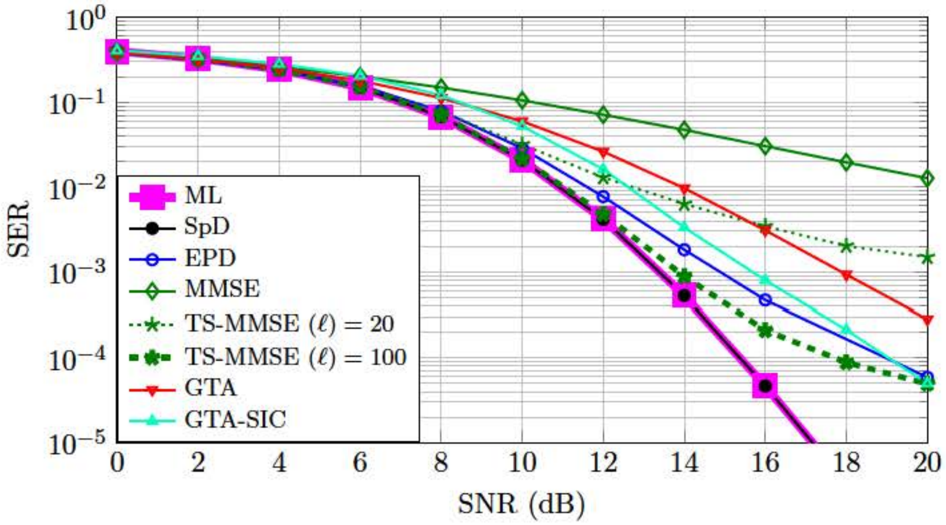


Figure III.3: SER performance in a $m = r = 6$ system with 4-QAM.

performance but the gain is not as big as it is in Figure III.3 for MMSE, because EPD provides a better initial point than MMSE and TS can not improves the final decision in many cases.

In Figure III.5 the number of antennas and the constellation are raised up to $m = r = 32$ and 16-QAM respectively. In this scenario, SpD or ML are no longer viable solutions. In this scenario, we test the influence in the EPD performance of the number I of iterations. In this case $I = 100$, $I = 10$ and $I = 2$ iterations are selected. Observe there is no difference in performance between the first two, meaning that it is pointless to increase the EPD complexity beyond $I = 10$ matrix inversions (III.11). Further, even the least complex EPD ($I = 2$) is already performing GTA-SIC. More specifically, the gain with respect to GTA-SIC at 10^{-3} are 0.5dB and 1.2dB for $I = 2$ and $I = 10 = 100$ respectively. In Figures Figures III.6 to III.9, we reproduce similar experiments with different scenarios with increasing number of antennas and constellation orders. All these experiments prove that the EPD solution with $I = 10$ iterations provides excellent performance with respect to state-of-the-art methods with similar complexity. Further, in Figure III.10, we illustrate how the EPD performance is robust and stable for increasing m and r (keeping $m = r$) if we maintain the constellation order. By comparing the high-SNR performance of EPD and GTA-SIC for $m = r = 32$, $m = r = 64$, and $m = r = 100$ we can see that the EPD degrades much slower than GTA-SIC. In particular, compare the SNR required by each method to achieve a SER of 10^{-4}. In particular, observe that EPD, regardless the dimension ($m = r$), typically needs around 25dB, while this

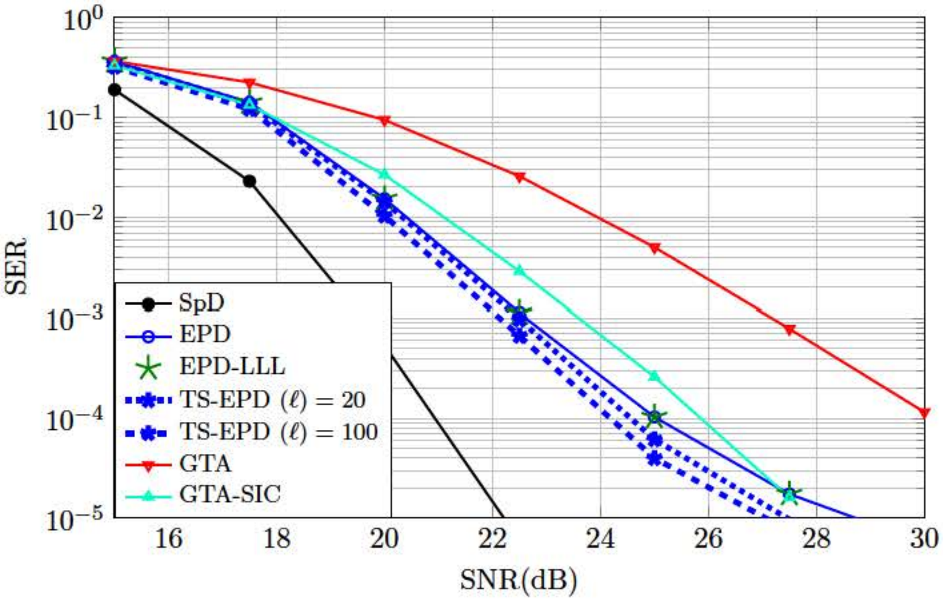


Figure III.4: SER performance in a $m = r = 12$ system with 16-QAM.

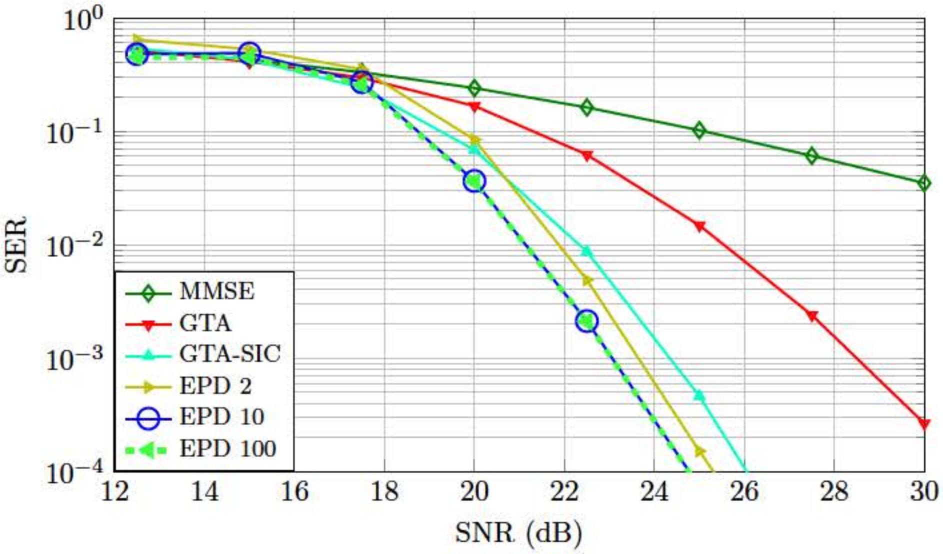


Figure III.5: SER performance in a $m = r = 32$ system with 16-QAM.

value grows with the dimension for GTA-SIC. Up to our knowledge, this behavior has not been reported for any other MIMO detection method of $\mathcal{O}(m^3)$ complexity.

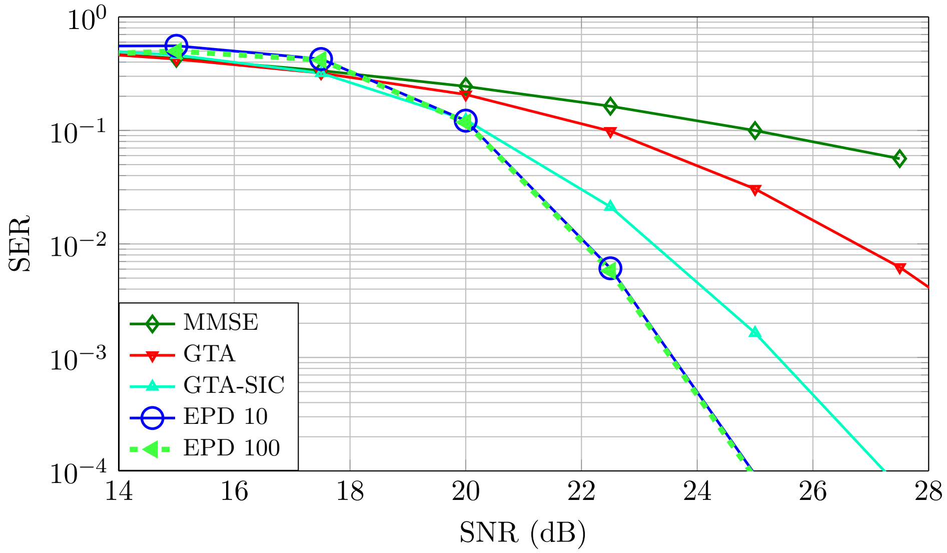


Figure III.6: SER performance in a $m = r = 64$ system with 16-QAM.

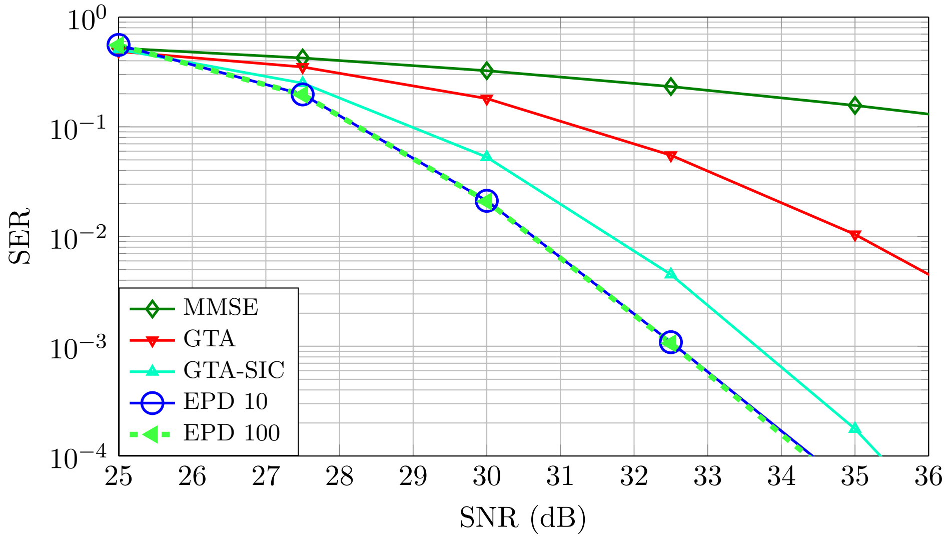


Figure III.7: SER performance in a $m = r = 64$ system with 64-QAM.

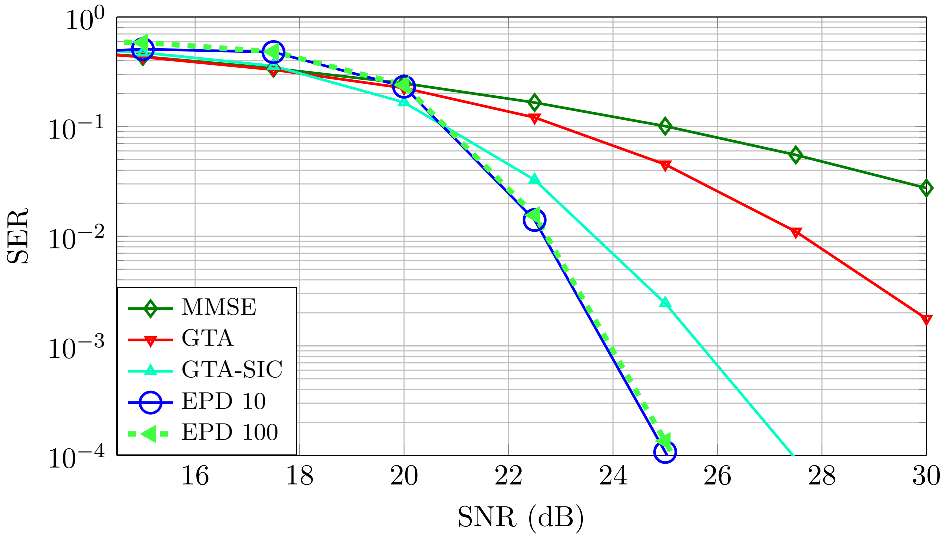


Figure III.8: SER performance in a $m = r = 100$ system with 16-QAM.

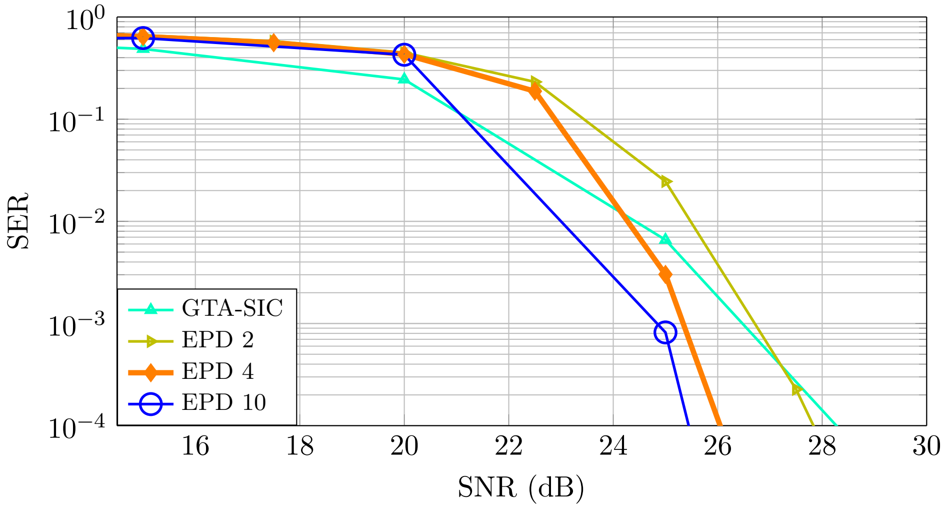


Figure III.9: SER performance in a $m = r = 250$ system with 16-QAM.

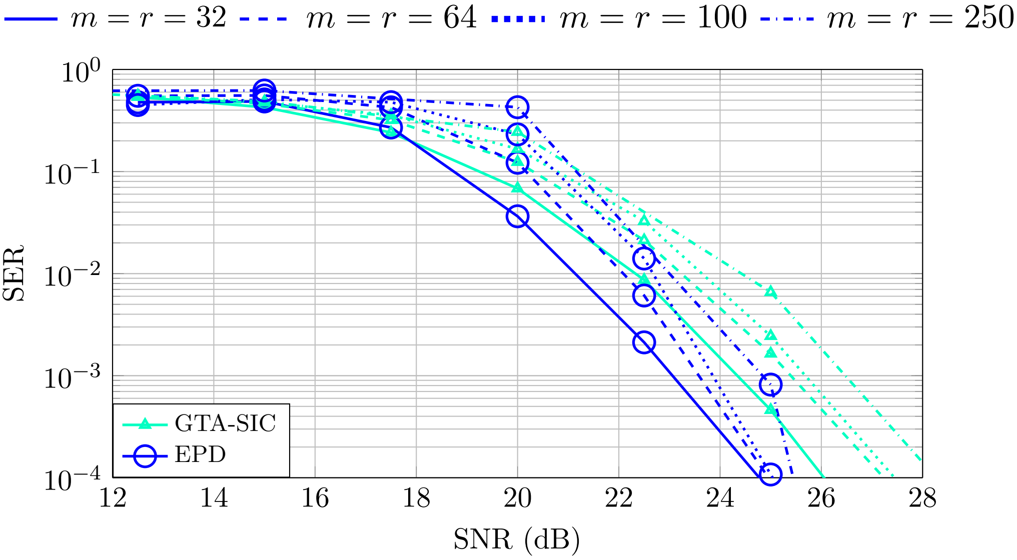


Figure III.10: SER performance in several systems (increasing $m = r$) and 16-QAM.

Chapter IV

Soft-Detection Methods

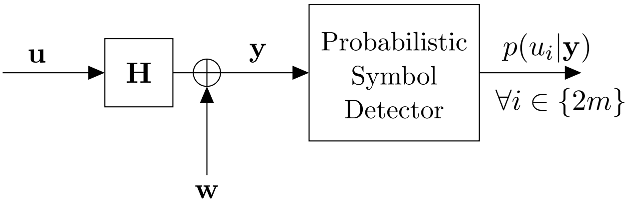
Modern channel-coding techniques, such as Turbo codes [75] or LDPC codes [76], are needed to achieve transmission rates close to the fundamental theoretical limits of the MIMO channel. Efficient decoding is possible using the BP algorithm [76, 77], which is a low-complexity message-passing approximate inference method that estimates marginal probabilities in a joint probability distribution. BP decoding needs as input an estimate to the posterior probability of each coded bit given the vector of channel observations. This information is provided by the so-called probabilistic symbol detector, see System Model in I.2, which has to marginalize the joint posterior pdf of the transmitted vector of QAM symbols, given the channel observation.

Multiple algorithms have been proposed to perform hard-output symbol detection in MIMO systems, some of them are introduced in Chapter II. On the contrary, the list of probabilistic symbol detection algorithms is comparatively much shorter. The focus on this chapter is on MIMO probabilistic symbol detection methods that can scale up to hundreds of antennas and high-order modulations. In particular, we focus on methods with polynomial complexity with the number of transmit antennas m. In addition, we will introduce some other works that use Markov-Chain Monte-Carlo (MCMC) algorithms to approximate marginal posterior probabilities. MCMC algorithms guarantee exact marginal computation in the limit of the number of samples taken, hence they can be used as a useful benchmark. For completeness, we include again the System Model IV.1 as the reference for these algorithms.

Recall that, as discussed in Chapter I, computing

$$p(u_i|\mathbf{y}) = \sum_{u \in \mathcal{A}} p(\mathbf{u}|\mathbf{y}) \tag{IV.1}$$

has $\mathcal{O}(M^m)$ complexity. Approximate Inference is the required to estimate (IV.1) at a feasible cost.



System Model IV.1: Probabilistic Symbol Detection Scenario

IV.1 MMSE and GTA as Soft Detectors

As it described in Chapter II, MMSE [20, 48] approximates the posterior distribution $p(\mathbf{u}|\mathbf{y})$ as a multivariate Gaussian distribution, that allows for tractable marginalization. The MMSE Gaussian approximation is determined by the following moments:

$$\mathbf{\Sigma}_{\text{MMSE}} = \left(\mathbf{H}^\top\mathbf{H} + \frac{\sigma_w^2}{E_s}\mathbf{I}\right)^{-1} \tag{IV.2}$$

$$\boldsymbol{\mu}_{\text{MMSE}} = \mathbf{\Sigma}_{\text{MMSE}}\mathbf{H}^\top\mathbf{y}, \tag{IV.3}$$

where for hard detection only the mode of the distribution was taken into account. We can obtain marginal posterior probabilities as follows:

$$p_{\text{MMSE}}(u_i|\mathbf{y}) = \mathcal{N}(u_i : \mu_{\text{MMSE},i}, \sigma^2_{\text{MMSE},i}) \ \forall i \in \{2m\}. \tag{IV.4}$$

where $\boldsymbol{\sigma}^2_{\text{MMSE}} = \text{diag}\left(\mathbf{\Sigma}_{\text{MMSE}}\right)$. Then, coded bit marginal probabilities are simply computed as follows

$$p_{\text{MMSE}}(c_{ji} = c|\mathbf{y}) = \sum_{u_i \in \mathcal{B}_j(c)} p_{\text{MMSE}}(u_i|\mathbf{y}), \tag{IV.5}$$

where recall that c_{ji} is the bit assigned to j-th position of the Gray code at the i-th antenna. Similarly, recall that GTA also constructs a tree-factorized approximation to $p(\mathbf{u}|\mathbf{y})$, see (II.13). Given $p_{\text{GTA}}(\mathbf{u}|\mathbf{y})$, marginalization to compute $p_{\text{GTA}}(u_i|\mathbf{y}) \approx p(u_i|\mathbf{y})$ is straightforward using BP. Then, bit posterior marginal probabilities are computed as in (IV.5).

IV.2 CHEMP algorithm

The Channel Hardening-Exploiting Message Passing (CHEMP) [78] algorithm is based on a message passing schedule, and it is inspired on the success of Approximate Message Passing (AMP) techniques in sparse signal reconstruction (compressed sensing) [79]. AMP algorithms essentially implement the standard rules of BP message passing [65] and all messages are approximated with univariate Gaussian distributions. Given the observation vector $\mathbf{y}$, we define:

$$\mathbf{H}^\top \mathbf{y} = \mathbf{H}^\top \mathbf{H}\mathbf{u} + \mathbf{H}^\top \mathbf{w}, \tag{IV.6}$$

and equation (IV.6) is rewritten as follows:

$$\mathbf{z} = \mathbf{J}\mathbf{u} + \mathbf{v}, \tag{IV.7}$$

where:

$$\mathbf{z} \triangleq \frac{\mathbf{H}^\top \mathbf{y}}{2m}, \quad \mathbf{J} \triangleq \frac{\mathbf{H}^\top \mathbf{H}}{2m}, \quad \mathbf{v} \triangleq \frac{\mathbf{H}^\top \mathbf{w}}{2m}. \tag{IV.8}$$

Note that the i-th element of $\mathbf{z}$ is finally given by

$$z_i = J_{ii} u_i + \underbrace{\sum_{j=1, j\neq i}^{2m} J_{ij} u_j + v_i}_{\triangleq\, g_i}, \tag{IV.9}$$

where g_i denotes the interference-plus-noise term. Given $\mathbf{z}$, CHEMP iteratively estimates $p(u_i|\mathbf{z})$, $\forall i \in \{2m\}$, by assuming that g_i is Gaussian distributed with mean $\mu_{g,i}$ and variance $\sigma^2_{g,i}$. These two moments are computed as follows. If $q_j^{(\ell-1)}(u_j)$ is the CHEMP estimate to $p(u_j|\mathbf{z})$ after iteration $\ell - 1$, $u_j \in \mathcal{A}$ $\forall j \in \{2m\}$, then:

$$\mu_{g,i} = \sum_{\substack{j=1\\ j\neq i}}^{2m} J_{ij} \mathbb{E}_{p(\mathbf{u})}[u_j] \approx \sum_{\substack{j=1\\ j\neq i}}^{2m} J_{ij} \left(\sum_{\forall s \in \mathcal{A}} s \; q_j^{(\ell-1)}(s) \right), \tag{IV.10}$$

and

$$\begin{aligned} \sigma^2_{g,i} &\approx \sum_{\substack{j=1\\ j\neq i}}^{2m} J_{ij}^2 \mathbb{E}_{p(\mathbf{u})}[u_j^2] + \sigma_v^2 \\ &= \sum_{\substack{j=1\\ j\neq i}}^{2m} J_{ij}^2 \left(\sum_{\forall s \in \mathcal{A}} s^2 q_j^{(\ell-1)}(s) - \mathbb{E}_{p(\mathbf{u})}[u_j]^2 \right) + \sigma_v^2. \end{aligned} \tag{IV.11}$$

where $\sigma_v^2 = \sigma_w^2/2r$. Note that to compute the variance $\sigma^2_{g,i}$ in (IV.11) it is assumed that all symbols are statistically independent, which is generally not true. Based on (IV.9) we can estimate $q_i^{(\ell)}(u_i)$, $u_i \in \mathcal{A}$, as follows:

$$q_i^{(\ell)}(u_i) = \frac{1}{Z} \exp\left(\frac{-1}{2\sigma_i^2} (z_i - \mu_i - J_{ii} u_i)^2 \right), \tag{IV.12}$$

where Z is a normalization constant that is adjusted once we evaluate (IV.12) for all u_i in $\mathcal{A}$. The algorithm is initialized using a uniform distribution for $q_i^{(0)}(u_i)$, $\forall i \in \{2m\}$. The complexity per iteration is $\mathcal{O}(rm^2)$. Finally, it is possible from $p(u_i|\mathbf{y})$ to compute $p(c_{ji}|\mathbf{y})$, as done in (IV.5).

The CHEMP algorithm relies on the fact that, in the limit $m \to \infty$, the $\mathbf{J}$ matrix in (IV.8) is approximately diagonal and, consequently, the noise variance $\sigma_{g,i}^2$, in (IV.11) tends to σ_v^2 [78]. For large m values and 4-QAM modulation, the CHEMP method provides excellent performance results. However, as we show in the experimental result section in Chapter V, for fixed m and increasing constellation order M the CHEMP performance is significantly degraded. Several factors might explain such degradation, in particular the fact that the noise variance $\sigma_{g,i}^2$ in (IV.11) grows with the constellation order. Following [78], for large constellation orders (16-QAM and above) CHEMP requires to decrease the number of transmitting antennas, $m < r$, in order to maintain the performance observed for the 4-QAM case.

IV.3 Markov Chain Monte Carlo Methods

MCMC algorithms have their roots in the Metropolis-Hastings algorithm [35], which attempts to compute complex integrals by expressing them as expectations for some distribution and then estimating this expectation by drawing samples. MCMC methods have been proposed to approximate the marginal posterior probabilities in MIMO detection in [80–84], among others. On one hand, MCMC methods guarantee that eventually they converge to the exact solution, the exact marginal distribution in our case. However, this is rather an asymptotic result as the number of samples required must be extremely large for high-dimensional MIMO system with high-order QAM constellations, specially if we want to perform probabilistic detection. The reason is that in this case we require a sufficiently large number of samples per constellation point at each transmitter. For large m and high-order constellations, MCMC methods become excessively burdensome. Another drawback of MCMC methods for MIMO detection is that it is quite hard to predict the scaling of the complexity (in terms of number of samples or different initializations required) as a function of m, r or the constellation order.

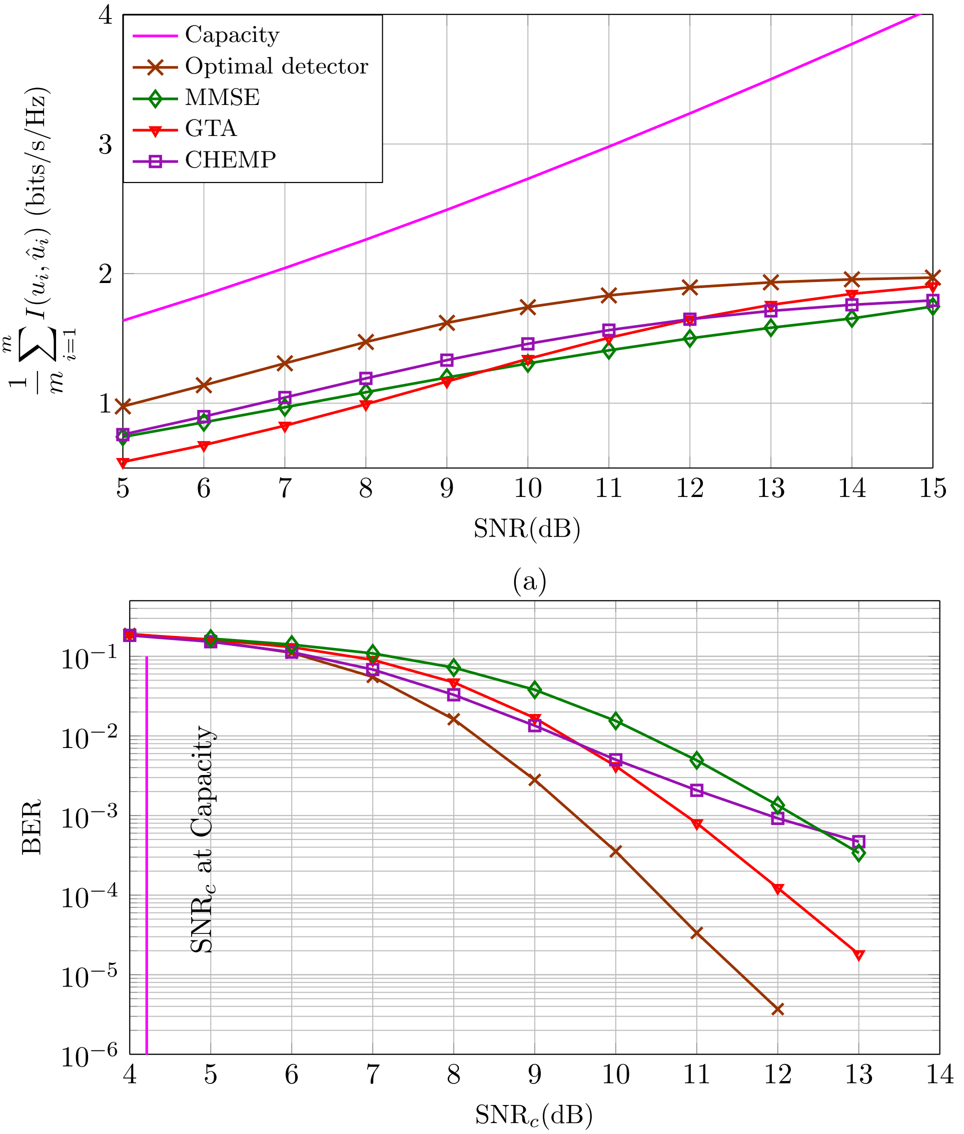
The easiest MCMC method for discrete distributions is the GIBBs sampler [35, 85]. This technique requires an initial estimation of $\mathbf{u}^{(0)}$ and afterwards each component is iteratively sampled using a conditional distribution using the last value sampled for the rest of variables. The initial estimation $\mathbf{u}^{(0)}$ can be randomly set or based on another estimation method, similarly as TS was initialized. Iteratively, each component $\forall i \in \{2m\}$ is updated sampling from the corresponding conditional distribution:

$$\begin{aligned}
u_1^{(\ell+1)} &\sim p(u_1|u_2^{(\ell)}, u_3^{(\ell)}, \cdots, u_{2m}^{(\ell)}, \mathbf{y}, \mathbf{H})\\
u_2^{(\ell+1)} &\sim p(u_2|u_1^{(\ell+1)}, u_3^{(\ell)}, \cdots, u_{2m}^{(\ell)}, \mathbf{y}, \mathbf{H})\\
&\vdots\\
u_{2m}^{(\ell+1)} &\sim p(u_{2m}|u_1^{(\ell+1)}, u_2^{(\ell+1)}, \cdots, u_{2m-1}^{(\ell+1)}, \mathbf{y}, \mathbf{H}).
\end{aligned} \tag{IV.13}$$

As discussed, the number of samples required to maintain a robust performance (accurate estimate of the symbol marginals using the obtain samples) grows very fast with either M and m. Other problem of these techniques is that the algorithm can be stucked in a local solution. In order to avoid this problem, usually these techniques are initialized with different $\mathbf{u}^{(0)}$ several times and the detected symbol vector is chosen as the least ML cost in several iterations.

IV.4 Comparison of State-of-the-art Methods

Consider again the $m = r = 5$ scenario with 4-QAM modulation which was the example to show the achievable rate in Figure I.4. Recall that the dimension is small enough so we are able to solve the marginalization in (IV.1) exactly. In Figure IV.1 (a) we include a comparison of the achievable rate for several soft detectors. We show how in the high-SNR regime most of the methods saturate to the constellation limit $\log_2(M) = 2$ bits. However in the low-SNR regime, the optimal detector is the closest to the capacity limit. Our aim is to get as close performance as possible in the low-SNR regime to the optimal one. Furthermore, in Figure IV.1(b) we include a $(3, 6)$-regular LDPC coded system with block length $k = 5120$ bits. In this case, System Model I.3 is used. Note that, to simulate the BER coded performance, the coding rate $R = 0.5$ is taken into account in the definition of $\mathrm{SNR_c}$. The not-so-good CHEMP performance for high-SNR regime, even worst than MMSE, m is small and assume independence is more difficult than for a higher value of m. However, in the low-SNR regime, when the differences between the optimal detector and the capacity limit are lower has the second performance. The aim of the propose soft detector based on matching moment is outperform CHEMP in low-SNR regime and GTA in high-SNR regime.



(a)

(b)

Figure IV.1: Performance in a $m = r = 5$ scenario with 4-QAM modulatio. (a), achievable transmission rates. (b), BER with a $(3, 6)$-regular LDPC code and block length $k = 5120$ bits.

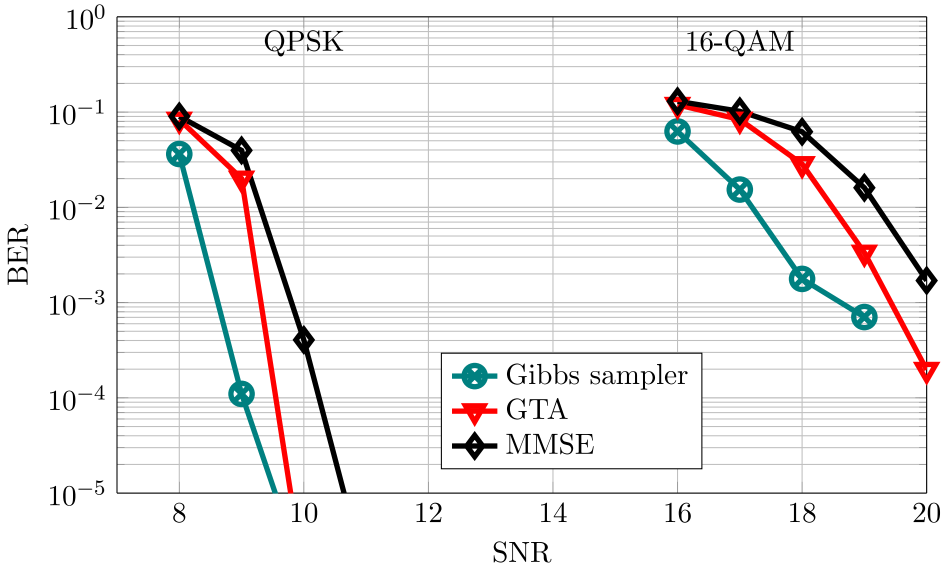


Figure IV.2: $m = r = 32$ MIMO system with 4-QAM and 16-QAM constellation. A $(3, 6)$-regular LDPC with code block $k = 10240$ bits has been used.

To provide some insight into how MCMC detection methods would perform in Figure IV.2 we show simulated performance for a $m = r = 32$ system with 4-QAM and 16-QAM using a $(3, 6)$-regular LDPC with code block $k = 10240$ bits. The MCMC method corresponds to a Gibbs sampling scheme (see [84]) where, for every channel observation $\mathbf{y}$ we run 10 parallel Gibbs samplers, each with a different random initialization and up to 5.10^3 samples (we consider a burn-in period of 5.10^3). Thus, marginals $p(u_i|\mathbf{y})$, for all $i \in \{2m\}$ are estimated using 5.10^4 samples. MCMC provides a better performance than GTA and MMSE, but in the 16-QAM case the we show how MCMC suffers a lack of performance because we are using the same number of samples than for 4-QAM case and in this case may need more to provide better approximations.

Chapter V

Soft-Detection via EC Approximations

In Chapter III, EP is used to find the mode of a posterior probability distribution that has been projected into a Gaussian approximation. The method cannot be easily tuned to perform probabilistic detection, as its description is essentially a message passing type of algorithm that does not provide the complete picture of the fundamental underlying inference problem. Indeed, in [86] we show that, while the MIMO EPD in [39] is able to significantly improve GTA in hard detection problem, both methods perform similarly when combined with an LDPC channel decoder that requires a probabilistic input.

In this chapter, we show how probabilistic MIMO symbol detection can be implemented using a more general approximate inference framework called Expectation Consistency, which was first described by Opper & Winther in [63]. In EC, we describe the inference problem as the search of an stationary point of an approximation to the free energy associated to the true posterior probability distribution of the transmitted symbols. Any stationary point satisfies a moment matching condition between the involved distributions. We discuss feasible methods to find such stationary points and show the fundamental tradeoffs between accuracy and speed of convergence. Based on this analysis, we find solutions that are robust and accurate across different modulation orders and system dimensions. The resulting EC probabilistic MIMO detector achieve excellent performance results compared to state-of-the-art methods with the same complexity order.

By computing the mutual information between the transmitted MIMO symbol vector and the corresponding output of the probabilistic symbol detection stage, we show that the transmission rate of a single-user MIMO system heavily depends on the probabilistic detector implemented. The proposed ECD MIMO detector achieves the closest gap to channel capacity at moderate SNR for all tested scenarios. Further, the gain at moderate SNRs obtained by ECD in comparison with EPD, is corroborated by performance simulation using optimized irregular LDPC block codes [87] and terminated convolutional-LDPC block codes [88, 89], in a similar scenario as was tested EPD in [86]. In all cases, we obtain remarkable SNR gains, proving that the accuracy of the MIMO probabilistic symbol detection stage is crucial in order to achieve close-to-capacity performance.

Finally, we show that the probabilistic output given by the BP algorithm after LDPC channel decoding can be fed back to the ECD symbol detection stage by a simple modification in the initialization point. Simulation results indicate that the performance of a MIMO receiver based on EC detection and LDPC channel coding with a feedback loop does not significantly improve the open-loop architecture if the LDPC code length is long enough. Therefore ECD probabilistic output is accurate enough and the use of a feedback loop in a more complex receiver would not be necessary. In any case, for moderate block lengths, as those typically used in mobile wireless communications, the closed-loop architecture can bring non-negligible performance gains.

V.1 Expectation Consistency Approximate Inference

A brief introduction to EC approximate inference [63] is presented next. The formulation given in this section is actually general and allows a straightforward description of ECD for MIMO. Let $\mathbf{u}$ be a random variable with a probability density function that factors in the following way

$$p(\mathbf{u}) = \frac{1}{Z} f_q(\mathbf{u}) f_r(\mathbf{u}). \tag{V.1}$$

We are in an scenario where the computation of $Z = \int f(\mathbf{u}) d\mathbf{u}$ is unfeasible, and so it is the computation of any moment over $p(\mathbf{u})$. Nevertheless, separately, $f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ are tractable with regard to a measure of the form $\exp(\boldsymbol{\lambda}^T \boldsymbol{\phi}(\mathbf{u}))$ for some real value natural parameter vector $\boldsymbol{\lambda}$ and some function vector

$$\boldsymbol{\phi}(\mathbf{u}) = [\phi_1(\mathbf{u}), \ldots, \phi_J(\mathbf{u})].$$

Namely, it is possible to perform inference over the following two distributions:

$$q(\mathbf{u}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{u}) \exp(\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})), \tag{V.2}$$

$$r(\mathbf{u}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{u}) \exp(\boldsymbol{\lambda}_r^\top \boldsymbol{\phi}(\mathbf{u})), \tag{V.3}$$

where the J-th dimensional parameter vectors $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ belong to a certain convex set Φ, and

$$Z_q(\boldsymbol{\lambda}_q) = \int f_q(\mathbf{u}) \exp(\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u}, \tag{V.4}$$

$$Z_r(\boldsymbol{\lambda}_r) = \int f_r(\mathbf{u}) \exp(\boldsymbol{\lambda}_r^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u}. \tag{V.5}$$

Note that both $q(\mathbf{u})$ and $r(\mathbf{u})$ define an exponential family of distributions[1], where $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ are respectively is the natural parameter vector, $\boldsymbol{\phi}(\mathbf{u})$ is the vector of sufficient statistics, and $\log Z_q(\boldsymbol{\lambda}_q)$ and $\log Z_r(\boldsymbol{\lambda}_r)$ are convex functions of $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ respectively that satisfy

$$\nabla_{\boldsymbol{\lambda}_q} \log Z_q(\boldsymbol{\lambda}_q) = \mathbb{E}_{q(\mathbf{u})} \left[\boldsymbol{\phi}(\mathbf{u})\right], \tag{V.6}$$

$$\nabla_{\boldsymbol{\lambda}_r} \log Z_r(\boldsymbol{\lambda}_r) = \mathbb{E}_{r(\mathbf{u})} \left[\boldsymbol{\phi}(\mathbf{u})\right]. \tag{V.7}$$

Since both $q(\mathbf{u})$ and $r(\mathbf{u})$ contain “partial information” of the true distribution $p(\mathbf{u})$ ($f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ respectively), the main idea behind EC approximate inference is to optimize $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ so that $q(\mathbf{u})$ and $r(\mathbf{u})$ have the same moments, i.e., (V.6) is consistent with (V.7). The first step to derive EC approximation is to note that the partition function Z in (V.1)

[1] See [65] for an introduction to exponential families and their properties.

can be expressed in the following way

$$Z = Z_q(\boldsymbol{\lambda}_q)\frac{Z}{Z_q(\boldsymbol{\lambda}_q)} = Z_q(\boldsymbol{\lambda}_q)\frac{\int f(\mathbf{u})d\mathbf{u}}{Z_q(\boldsymbol{\lambda}_q)} \tag{V.8}$$

$$\begin{aligned}
&= Z_q(\boldsymbol{\lambda}_q)\frac{\displaystyle\int f_q(\mathbf{u})f_r(\mathbf{u})d\mathbf{u}}{\displaystyle\int f_q(\mathbf{u})\exp(\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))d\mathbf{u}} \\
&= Z_q(\boldsymbol{\lambda}_q)\frac{\displaystyle\int f_q(\mathbf{u})f_r(\mathbf{u})\exp((\boldsymbol{\lambda}_q-\boldsymbol{\lambda}_q)^\top\boldsymbol{\phi}(\mathbf{u}))d\mathbf{u}}{\displaystyle\int f_q(\mathbf{u})\exp(\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))d\mathbf{u}} \\
&= Z_q(\boldsymbol{\lambda}_q)\frac{\displaystyle\int f_q(\mathbf{u})\exp(\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))d\mathbf{u}}{\displaystyle\int f_q(\mathbf{u})\exp(\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))d\mathbf{u}} \\
&= Z_q(\boldsymbol{\lambda}_q)\mathbb{E}_{q(\mathbf{u})}[f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))].
\end{aligned} \tag{V.9}$$

And thus,

$$\log Z = \log Z_q(\boldsymbol{\lambda}_q) + \log\left(\mathbb{E}_{q(\mathbf{u})}[f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))]\right). \tag{V.10}$$

Again we assume that we are in an scenario where the expectation

$$E_{q(\mathbf{u})}[f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top\boldsymbol{\phi}(\mathbf{u}))] \tag{V.11}$$

is not evaluable. In [63], the authors propose to approximate this expectation by replacing $q(\mathbf{u})$ by a simpler distribution $s(\mathbf{u})$ that belongs to the same exponential family than $q(\mathbf{u})$ and $r(\mathbf{u})$, i.e.,

$$s(\mathbf{u}) = \frac{1}{Z_s(\boldsymbol{\lambda}_s)}\exp(\boldsymbol{\lambda}_s^\top\boldsymbol{\phi}(\mathbf{u})), \tag{V.12}$$

where $\log Z_s(\boldsymbol{\lambda}_s)$ is a convex function of $\boldsymbol{\lambda}_s$ that satisfies $\nabla_{\boldsymbol{\lambda}_s}\log Z_s(\boldsymbol{\lambda}_s) = \mathbb{E}_{s(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})]$. While replacing $q(\mathbf{u})$ by $s(\mathbf{u})$ yields in general a poor approximation of (V.10), it can be a fairly reasonable solution if both $q(\mathbf{u})$ and $s(\mathbf{u})$ have the same moments, namely if $\mathbb{E}_{q(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})] = \mathbb{E}_{s(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})]$. In a beautiful way, this condition is naturally achieved as a stationary point of the $\log Z$ approximation, as follows. By replacing $q(\mathbf{u})$ by $s(\mathbf{u})$ in (V.10),

$\log Z$ is approximated by

$$
\begin{aligned}
&\log Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) \\
&= \log Z_q(\boldsymbol{\lambda}_q) + \log \left(\mathbb{E}_{s(\mathbf{u})}[f_r(\mathbf{u}) \exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u}))] \right) \\
&= \log Z_q(\boldsymbol{\lambda}_q) + \log \int \frac{\exp(\boldsymbol{\lambda}_s^\top \boldsymbol{\phi}(\mathbf{u}))}{Z_s(\boldsymbol{\lambda}_s)} f_r(\mathbf{u}) \exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u} \\
&= \log Z_q(\boldsymbol{\lambda}_q) + \log \int \exp(\boldsymbol{\lambda}_s^\top \boldsymbol{\phi}(\mathbf{u})) f_r(\mathbf{u}) \exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u} - \log Z_s(\boldsymbol{\lambda}_s) \\
&= \log Z_q(\boldsymbol{\lambda}_q) + \log \int f_r(\mathbf{u}) \exp \left((\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)^\top \boldsymbol{\phi}(\mathbf{u}) \right) d\mathbf{u} - \log Z_s(\boldsymbol{\lambda}_s) \\
&= \log Z_q(\boldsymbol{\lambda}_q) + \log Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q) - \log Z_s(\boldsymbol{\lambda}_s). \qquad \text{(V.13)}
\end{aligned}
$$

Note that $\log Z_{\mathrm{EC}}$ depends only on $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$, while it depends on three probability distributions: $q(\mathbf{u})$ with parameter vector $\boldsymbol{\lambda}_q$, $r(\mathbf{u})$ with parameter vector $(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)$ and $s(\mathbf{u})$ with parameter vector $\boldsymbol{\lambda}_s$ and that by assumption $Z_q(\boldsymbol{\lambda}_q)$, $Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)$ and $Z_s(\boldsymbol{\lambda}_s)$ can be computed efficiently. Recall that moment matching it is necessary between $q(\mathbf{u})$ and $r(\mathbf{u})$ and also between $q(\mathbf{u})$ and $s(\mathbf{u})$. While the first condition ensures that the two approximations that are constructing $p(\mathbf{u})$ are consistent, the second is required so that the measure replacement in the expectation in (V.10) is not too coarse. By using (V.6) and (V.7) it is easy to prove that:

$$\nabla_{\boldsymbol{\lambda}_q} \log Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) = \mathbb{E}_q[\boldsymbol{\phi}(\mathbf{u})] - \mathbb{E}_r[\boldsymbol{\phi}(\mathbf{u})], \qquad \text{(V.14)}$$

$$\nabla_{\boldsymbol{\lambda}_s} \log Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) = \mathbb{E}_r[\boldsymbol{\phi}(\mathbf{u})] - \mathbb{E}_s[\boldsymbol{\phi}(\mathbf{u})]. \qquad \text{(V.15)}$$

Therefore, any pair $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ that satisfies that the gradient of $\log Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ with respect to $\boldsymbol{\lambda}_q$ is zero yields moment consistency between $q(\mathbf{u})$ and $r(\mathbf{u})$, while any pair $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$ that satisfies that the gradient of $\log Z_{\mathrm{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ with respect to $\boldsymbol{\lambda}_s$ is zero yields moment consistency between $r(\mathbf{u})$ and $s(\mathbf{u})$.

While the search of stationary points can be challenging ($\log Z_{\mathrm{EC}}$ is convex in $\boldsymbol{\lambda}_q$ but is the sum of a convex term and a concave term with respect to $\boldsymbol{\lambda}_s$), we have found that a very simple iterative procedure works well in general. In this algorithm, called in [63] the EC single loop, "messages" are sent back and forth between the two distributions $q(\mathbf{u})$ and $r(\mathbf{u})$. $s(\mathbf{u})$ is updated to be consistent with either $q(\mathbf{u})$ or $r(\mathbf{u})$ depending in which way are propagated. The algorithm is resumed in Algorithm 5.

Convergence is achieved when $||\boldsymbol{\mu}_q^{(\ell-1)} - \boldsymbol{\mu}_r^{(\ell)}||$ is below a certain threshold. However, convergence is not guaranteed [63], so it is also necessary to set an upper limit in the number of iterations. Note that the definition of convergence is quite subtle, in the sense that the above iterative algorithm

Algorithm 5 The EC Single Loop

```
Initialize $\ell = 1$, $\boldsymbol{\lambda}_q^{(0)}$
repeat
    1) Given $\boldsymbol{\lambda}_q^{(\ell-1)}$, compute $\boldsymbol{\mu}_q^{(\ell-1)} = \mathbb{E}_q[\boldsymbol{\phi}(\mathbf{u})]$.
    2) Compute $\boldsymbol{\lambda}_s^{(\ell-1)}$ such that $\mathbb{E}_s[\boldsymbol{\phi}(\mathbf{u})] = \boldsymbol{\mu}_q^{(\ell-1)}$.
    3) Update $\boldsymbol{\lambda}_r^{(\ell)} = \boldsymbol{\lambda}_s^{(\ell-1)} - \boldsymbol{\lambda}_q^{(\ell-1)}$.
    4) Given $\boldsymbol{\lambda}_r^{(\ell)}$, compute $\boldsymbol{\mu}_r^{(\ell)} = \mathbb{E}_r[\boldsymbol{\phi}(\mathbf{u})]$.
    5) Compute $\boldsymbol{\lambda}_s^{(\ell)}$ such that $\mathbb{E}_s[\boldsymbol{\phi}(\mathbf{u})] = \boldsymbol{\mu}_r^{(\ell)}$.
    6) Update $\boldsymbol{\lambda}_q^{(\ell)} = \boldsymbol{\lambda}_s^{(\ell)} - \boldsymbol{\lambda}_r^{(\ell)}$.
    7) $\ell = \ell + 1$
until convergence (or stop criterion)
```

may get stuck in a $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$ point such that these parameters do not change anymore, but at the same time the moment matching condition is not fully met. Further, in order to avoid numerical issues, a *damping* (low-pass filter) is implemented in the update of $\boldsymbol{\lambda}_q$ at step 6) of the algorithm, in which $\boldsymbol{\lambda}_q$ is updated using a convex combination between the old value and the new one. Namely, updating $\boldsymbol{\lambda}_q$ as follows: $\boldsymbol{\lambda}_q^{(\ell)} = \beta(\boldsymbol{\lambda}_s^{(\ell)} - \boldsymbol{\lambda}_r^{(\ell)}) + (1-\beta)\boldsymbol{\lambda}_q^{(\ell-1)}$ for some damping factor $\beta \in [0, 1]$. Smoothing the parameter via damping is a fairly common technique to stabilize approximate inference iterative algorithms. See for instance [90–92] for discussions on stabilization in message passing algorithms.

V.2 EC MIMO detection

In this section, we adapt the EC inference methodology described above to construct an approximation to the MIMO posterior probability distribution in (I.8), in this case explicitly including the normalization parameter Z:

$$p(\mathbf{u}|\mathbf{y}) = \frac{1}{Z}\mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})\frac{1}{\sqrt{M}}\prod_{i=1}^{2m}\mathbb{I}_{u_i\in\mathcal{A}}. \tag{V.16}$$

V.2.1 EC distributions

We propose the following equivalences

$$f_q(\mathbf{u}) = \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I}), \tag{V.17}$$

$$f_r(\mathbf{u}) = \frac{1}{\sqrt{M}}\prod_{i=1}^{2m}\mathbb{I}_{u_i\in\mathcal{A}}. \tag{V.18}$$

Also we build $q(\mathbf{u}), r(\mathbf{u})$ and $s(\mathbf{u})$ with the following vector of sufficient statistics and vector of natural parameters

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{u}) &= \left[u_1, u_2, \ldots, u_{2m}, \frac{-u_1^2}{2}, \frac{-u_2^2}{2}, \ldots, \frac{-u_{2m}^2}{2}\right]^\top, && \text{(V.19)}\\ \boldsymbol{\lambda}_q &= [\gamma_{q,1}, \gamma_{q,2}, \ldots, \gamma_{q,2m}, \Lambda_{q,1}, \Lambda_{q,2}, \ldots, \Lambda_{q,2m}]^\top = [\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q]^\top \\ \boldsymbol{\lambda}_r &= [\gamma_{r,1}, \gamma_{r,2}, \ldots, \gamma_{r,2m}, \Lambda_{r,1}, \Lambda_{r,2}, \ldots, \Lambda_{r,2m}]^\top = [\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r]^\top \\ \boldsymbol{\lambda}_s &= [\gamma_{s,1}, \gamma_{s,2}, \ldots, \gamma_{s,2m}, \Lambda_{s,1}, \Lambda_{s,2}, \ldots, \Lambda_{s,2m}]^\top = [\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s]^\top, && \text{(V.20)}\end{aligned}$$

where $\boldsymbol{\gamma}_q, \boldsymbol{\gamma}_r, \boldsymbol{\gamma}_s \in \mathbb{R}^{2m}$ and $\boldsymbol{\Lambda}_q, \boldsymbol{\Lambda}_r, \boldsymbol{\Lambda}_s \in \mathbb{R}_+^{2m}$. According to (V.2) and (V.17), is possible to show

$$\begin{aligned}q(\mathbf{u}) &\propto f_q(\mathbf{u}) \exp(\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})) = f_q(\mathbf{u}) \exp\left(\boldsymbol{\gamma}_q^\top \mathbf{u} - \frac{\mathbf{u}^\top \operatorname{diag}(\boldsymbol{\Lambda}_q)\mathbf{u}}{2}\right) \\ &\propto \exp\left(\underbrace{\left(\sigma_w^{-2}\mathbf{H}^\top \mathbf{y} + \boldsymbol{\gamma}_q\right)^\top}_{\mathbf{g}^\top} \mathbf{u} - \frac{1}{2}\mathbf{u}^\top \underbrace{\left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \operatorname{diag}(\boldsymbol{\Lambda}_q)\right)}_{\mathbf{S}} \mathbf{u}\right). && \text{(V.21)}\end{aligned}$$

Therefore $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} : \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbb{E}_{q(\mathbf{u})}[(\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})] = \boldsymbol{\Sigma} = \mathbf{S}^{-1}, \qquad \text{(V.22)}$$

$$\mathbb{E}_{q(\mathbf{u})}[\mathbf{u}] = \boldsymbol{\mu} = \mathbf{S}^{-1}\mathbf{g}. \qquad \text{(V.23)}$$

On the other hand, from the definition of $f_r(\mathbf{u})$ in (V.18) we get

$$\begin{aligned}r(\mathbf{u}) &\propto \exp\left(\boldsymbol{\gamma}_r^T \mathbf{u} - \frac{\mathbf{u}^T \operatorname{diag}(\boldsymbol{\Lambda}_r)\mathbf{u}}{2}\right) \prod_{i=1}^{2m} \mathbb{I}_{u_i \in \mathcal{A}} \\ &\propto \prod_{i=1}^{2m} \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right) \mathbb{I}_{u_i \in \mathcal{A}}, && \text{(V.24)}\end{aligned}$$

and thus $r(\mathbf{u})$ is an independent discrete distribution over $\mathcal{A}^{2m}$ such that

$$\mathbb{E}_{r(\mathbf{u})}[u_i] = \frac{\sum_{u_i \in \mathcal{A}} u_i \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right)}{\sum_{q \in \mathcal{A}} \exp\left(\gamma_{ri} q - \frac{\Lambda_{ri} q^2}{2}\right)}, \qquad \text{(V.25)}$$

$$\mathbb{E}_{r(\mathbf{u})}[u_i^2] = \frac{\sum_{u_i \in \mathcal{A}} u_i^2 \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right)}{\sum_{q \in \mathcal{A}} \exp\left(\gamma_{ri} q - \frac{\Lambda_{ri} q^2}{2}\right)}. \qquad \text{(V.26)}$$

Finally, the distribution $s(\mathbf{u})$ is given by

$$s(\mathbf{u}) \propto \exp(\boldsymbol{\lambda}_s^\top \boldsymbol{\phi}(\mathbf{u})) = \exp\left(\boldsymbol{\gamma}_s^\top \mathbf{u} - \frac{\mathbf{u}^\top \operatorname{diag}(\boldsymbol{\Lambda}_s)\,\mathbf{u}}{2}\right), \tag{V.27}$$

and therefore $s(\mathbf{u})$ is an independent Gaussian distribution, i.e. $s(\mathbf{u}) = \mathcal{N}(\mathbf{u} : \boldsymbol{\Lambda}_s^{-1}\boldsymbol{\gamma}_s, \operatorname{diag}\left(\boldsymbol{\Lambda}_s^{-1}\right))$ where

$$\mathbb{E}_{s(\mathbf{u})}[(\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})] = \boldsymbol{\Lambda}_s^{-1} \tag{V.28}$$

$$\mathbb{E}_{s(\mathbf{u})}[\mathbf{u}] = \boldsymbol{\Lambda}_s^{-1}\boldsymbol{\gamma}_s. \tag{V.29}$$

V.2.2 Single Loop ECD detection

The aim in ECD inference is to find some $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ and $(\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s)$ pairs such that $q(\mathbf{u})$ in (V.21), $r(\mathbf{u})$ in (V.24) (evaluated at $\boldsymbol{\gamma}_r = \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q$ and $\boldsymbol{\Lambda}_r = \boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q$) and $s(\mathbf{u})$ in (V.27) that satisfy

$$\mathbb{E}_{q(\mathbf{u})}[u_i] = \mathbb{E}_{r(\mathbf{u})}[u_i] = \mathbb{E}_{s(\mathbf{u})}[u_i] \tag{V.30}$$

$$\mathbb{E}_{q(\mathbf{u})}[u_i^2] = \mathbb{E}_{r(\mathbf{u})}[u_i^2] = \mathbb{E}_{s(\mathbf{u})}[u_i^2], \tag{V.31}$$

$\forall i \in \{2m\}$.

To find such a point, we use Algorithm 5 described before, which is particularized to the MIMO scenario in Algorithm 6. The steps in this algorithm are straightforward to implement given the expressions of the moments of $q(\mathbf{u}), r(\mathbf{u})$ and $s(\mathbf{u})$. Furthermore, we inicialize $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ such that $q(\mathbf{u})$ in (V.21) coincides with the MMSE Gaussian approximation, i.e., $\boldsymbol{\gamma}_q^{(0)} = \mathbf{0}$ and $\boldsymbol{\Lambda}_q^{(0)} = E_s^{-1}$ [20, 27]. The complexity of the Single Loop per iteration is dominated by the computation of the covariance matrix of the $q(\mathbf{u})$ distribution in (V.21). This complexity is $\mathcal{O}(m^3)$, thus it is independent on the constellation size $|\mathcal{A}|$. Computing the $r(\mathbf{u})$ mean and variance in (V.25) and (V.26) requires $\mathcal{O}(m|\mathcal{A}|)$ operations. The complexity of the rest of steps do not depend on the constellation and thus the complexity is $\mathcal{O}(m)$. Therefore, if the algorithm is run for I iterations, the final complexity is $\mathcal{O}(m^3 I + m|\mathcal{A}|I + mI)$, exactly as the EP complexity described in III.

Algorithm 6 The Single loop ECD

```
Fix a damping factor $\beta$.
Initialize $\boldsymbol{\gamma}_q^{(0)} = \mathbf{0}$ and $\boldsymbol{\Lambda}_q^{(0)} = E_s^{-1}\mathbf{I}$.
repeat
    1) Given $\boldsymbol{\gamma}_q^{(\ell-1)}, \boldsymbol{\Lambda}_q^{(\ell-1)}$, compute $\mathbb{E}_{q(\mathbf{u})}[u_i]$ and $\mathbb{E}_{q(\mathbf{u})}[u_i^2]$, $\forall i \in \{2m\}$.
    2) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_{s(\mathbf{u})}(\mathbf{u})[u_i] = \mathbb{E}_{q(\mathbf{u})}[u_i]$ and $\mathbb{E}_{s(\mathbf{u})}[u_i^2] = \mathbb{E}_{q(\mathbf{u})}[u_i^2]$, $\forall i \in \{2m\}$.
    3) Update $\boldsymbol{\gamma}_r^{(\ell)} = \boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_q^{(\ell)}$, $\boldsymbol{\Lambda}_r^{(\ell)} = \boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_q^{(\ell)}$.
    4) Given $\boldsymbol{\gamma}_r^{(\ell)}, \boldsymbol{\Lambda}_r^{(\ell)}$, compute $\mathbb{E}_{r(\mathbf{u})}[u_i]$ and $\mathbb{E}_{r(\mathbf{u})}[u_i^2]$, $\forall i \in \{2m\}$.
    5) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_{s(\mathbf{u})}[u_i] = \mathbb{E}_{r(\mathbf{u})}[u_i]$ and $\mathbb{E}_{s(\mathbf{u})}[u_i^2] = \mathbb{E}_{r(\mathbf{u})}[u_i^2]$, $\forall i \in \{2m\}$.
    6) Update
        $$\boldsymbol{\gamma}_q^{(\ell)} = \beta\left(\boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\gamma}_q^{(\ell-1)}$$
        $$\boldsymbol{\Lambda}_q^{(\ell)} = \beta\left(\boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\Lambda}_q^{(\ell-1)}$$
    7) $\ell = \ell + 1$
until convergence (or stop criterion)
```

V.3 EPD vs ECD

In this section, we include a comparison between the proposed receivers EPD and ECD in order to show that both algorithms are actually equivalent, and that the EC approach is a more fundamental point of view for approximate inference that can be exploited to derive more robust algorithms and better understanding of their convergence.

The first thing that can be observed is the way in which both algorithms approach the posterior distribution (I.8):

$$p(\mathbf{u}|\mathbf{y}) = \frac{1}{Z}\mathcal{N}(\mathbf{y}:\mathbf{Hu},\sigma_w^2\mathbf{I})\frac{1}{\sqrt{M}}\prod_{i=1}^{2m}\mathbb{I}_{u_i\in\mathcal{A}} \tag{V.32}$$

$$p_{\text{EP}}(\mathbf{u}|\mathbf{y}) \propto \mathcal{N}(\mathbf{y}:\mathbf{Hu},\sigma_w^2\mathbf{I})\prod_{i=1}^{2m}\mathrm{e}^{\gamma_i u_i-\frac{1}{2}\Lambda_i u_i^2} \tag{V.33}$$

$$\text{EC}\begin{cases} q(\mathbf{u}) = \frac{1}{Z_q(\boldsymbol{\gamma}_q,\boldsymbol{\Lambda}_q)}\mathcal{N}(\mathbf{y}:\mathbf{Hu},\sigma_w^2\mathbf{I})\prod_{i=1}^{2m}\mathrm{e}^{\gamma_{q,i}u_i-\frac{1}{2}\Lambda_{q,i}u_i^2} \\ \\ r(\mathbf{u}) = \frac{1}{Z_r(\boldsymbol{\gamma}_r,\boldsymbol{\Lambda}_r)}\prod_{i=1}^{2m}\mathbb{I}_{u_i\in\mathcal{A}}\mathrm{e}^{\gamma_{r,i}u_i-\frac{1}{2}\Lambda_{r,i}u_i^2} \end{cases} \tag{V.34}$$

Thus, in the EC case, we dispose of two distributions to approximate $p(\mathbf{u}|\mathbf{y})$ that can provide further information about the posterior distribution beyond the moments that are set to enforce consistency. For instance, note that $q(\mathbf{u})$ may captures correlations from $p(\mathbf{u}|\mathbf{y})$, while $r(\mathbf{u})$ captures the non-Gaussianity of the true distribution. Further, not that EC also provides an approximation to the partition function of $p(\mathbf{u}|\mathbf{y})$, and this is not directly given by the EP algorithm described in Algorithm 4.

However, we can establish a per-step equivalence between the single loop ECD in Algorithm 6 and the EPD in Algorithm 4. The most subtle point to understand this equivalence is to reveal how the overlapping distribution $s(\mathbf{u})$ appears in the derivation of EPD. As we can infer form Tables V.1 and V.2, $s(\mathbf{u})$ is related to the cavity marginals in step 2) of EPD. While the MM condition between $q(\mathbf{u})$, $s(\mathbf{u})$ and $r(\mathbf{u})$ emerged naturally in the ECD formulation, concluding that EPD should be stopped when there is MM between $p_{\text{EP}}(\mathbf{u}|\mathbf{y})$, the cavity marginals $p_{\text{EP}}^{(\ell)\backslash i}(u_i|\mathbf{y})$ in 2) and the distributions $\hat{p}^{(\ell)}(u_i|\mathbf{y})$ is not easily justified. This is why assessing convergence for the EPD algorithm was a much harder task than for ECD, as we show in the next section.

EPD as in Algorithm 4
Fix a damping factor β
Initialize $\ell = 1$, $\boldsymbol{\gamma}^{(0)} = \mathbf{0}$ and $\boldsymbol{\Lambda}^{(0)} = E_s^{-1}$
1) Given $\boldsymbol{\gamma}^{(\ell-1)}$ and $\boldsymbol{\Lambda}^{(\ell-1)}$ compute $\boldsymbol{\Sigma}_{\text{EP}}^{(\ell)}$, $\boldsymbol{\sigma}_{\text{EP}}^{(\ell)} = \text{diag}(\boldsymbol{\Sigma}_{\text{EP}}^{(\ell)})$ and $\boldsymbol{\mu}_{\text{EP}}^{(\ell)}$ $\boldsymbol{\Sigma}_{\text{EP}} = \left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \text{diag}\left(\boldsymbol{\Lambda}^{(\ell-1)}\right)\right)^{-1}$ $\boldsymbol{\mu}_{\text{EP}} = \boldsymbol{\Sigma}_{\text{EP}}\left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{y} + \boldsymbol{\gamma}^{(\ell-1)}\right)$
2) Compute the cavity marginals $p_{\text{EP}}^{(\ell)\backslash i}(u_i \mid \mathbf{y}) \propto \frac{p_{\text{EP}}^{(\ell)}(u_i \mid \mathbf{y})}{\exp(\gamma_i^{(\ell)}u_i - \frac{1}{2}\Lambda_i^{(\ell)}u_i^2)} = \mathcal{N}(u_i : t_i^{(\ell)}, h_i^{2(\ell)})$, $\forall i \in \{2m\}$: $h_i^{2(\ell)} = \frac{\sigma_{i\text{EP}}^{(\ell)}}{1 - \sigma_{i\text{EP}}^{(\ell)}\Lambda_i^{(\ell)}}$ $t_i^{(\ell)} = h_i^{2((\ell))}\left(\frac{\mu_i^{(\ell)}}{\sigma_{i\text{EP}}^{(\ell)}} - \gamma_i^{(\ell)}\right)$
3) Introduce the true constellation and compute $\mathbb{E}_{\hat{p}^{(\ell)}}[u_i^2]$ and $\mathbb{E}_{\hat{p}^{(\ell)}}[u_i]$ of $\hat{p}^{(\ell)}(u_i \mid \mathbf{y}) \propto p_{\text{EP}}^{(\ell)\backslash i}(u_i \mid \mathbf{y})\mathbb{I}_{u_i \in \mathcal{A}_i}$, $\forall i \in \{2m\}$.
4) Update $\forall i \in \{2m\}$ $\Lambda_i^{(\ell+1)} = \beta\left(\frac{1}{\sigma_{\hat{p}_i}^{2(\ell)}} - \frac{1}{h_i^{2(\ell)}}\right) + (1-\beta)\Lambda_i^{(\ell)}$ $\gamma_i^{(\ell+1)} = \beta\left(\frac{\mu_{\hat{p}_i}^{(\ell)}}{\sigma_{\hat{p}_i}^{2(\ell)}} - \frac{t_i^{(\ell)}}{h_i^{2(\ell)}}\right) + (1-\beta)\gamma_i^{(\ell)}$ 5) $\ell = \ell + 1$

Table V.1: EPD steps

ECD as in Algorithm 6
Fix a damping factor β
Initialize $\ell = 1$, $\boldsymbol{\gamma}_q^{(0)} = \mathbf{0}$ and $\boldsymbol{\Lambda}_q^{(0)} = E_s^{-1}$
1) Given $\boldsymbol{\gamma}_q^{(\ell-1)}, \boldsymbol{\Lambda}_q^{(\ell-1)}$, compute $\mathbb{E}_q[u_i]$ and $\mathbb{E}_q[u_i^2]$, $\forall i \in \{2m\}$ $\mathbb{E}_{q(\mathbf{u})}[\mathbf{u}^2] = \boldsymbol{\Sigma} = \left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{H} + \mathrm{diag}\left(\boldsymbol{\Lambda}_q^{(\ell-1)}\right)\right)^{-1}$ $\mathbb{E}_{q(\mathbf{u})}[\mathbf{u}] = \boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\sigma_w^{-2}\mathbf{H}^\top\mathbf{y} + \boldsymbol{\gamma}_q^{(\ell-1)}\right)$ 2) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that : $\mathbb{E}_{s(\mathbf{u})}[u_i] = \mathbb{E}_{q(\mathbf{u})}[u_i] = \mu_i$ $\mathbb{E}_{s(\mathbf{u})}[u_i^2] = \mathbb{E}_{q(\mathbf{u})}[u_i^2] = \mathrm{diag}(\boldsymbol{\Lambda})_i$
3) Update $\boldsymbol{\gamma}_r^{(\ell)} = \boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_q^{(\ell)}$ and $\boldsymbol{\Lambda}_r^{(\ell)} = \boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_q^{(\ell)}$
4) Given $\boldsymbol{\gamma}_r^{(\ell)}, \boldsymbol{\Lambda}_r^{(\ell)}$, compute $\mathbb{E}_{r(\mathbf{u})}[u_i]$ and $\mathbb{E}_{r(\mathbf{u})}[u_i^2]$, $\forall i \in \{2m\}$. 5) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_{s(\mathbf{u})}[u_i] = \mathbb{E}_{r(\mathbf{u})}[u_i]$ and $\mathbb{E}_{s(\mathbf{u})}[u_i^2] = \mathbb{E}r(\mathbf{u})[u_i^2]$, $\forall i \in \{2m\}$.
6) Update $\boldsymbol{\gamma}_q^{(\ell)} = \beta\left(\boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\gamma}_q^{(\ell-1)}$ $\boldsymbol{\Lambda}_q^{(\ell)} = \beta\left(\boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\Lambda}_q^{(\ell-1)}$ 7) $\ell = \ell + 1$

Table V.2: ECD steps

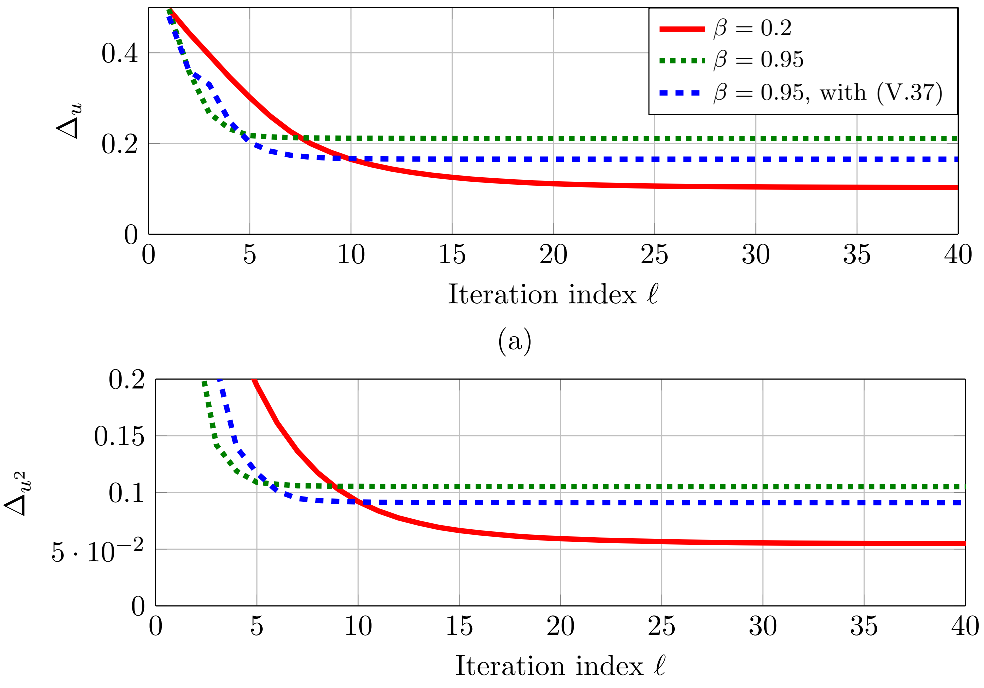
V.3.1 Assessing convergence

The moment matching condition in (V.30) and (V.31) represents the optimal operational point of the EC approximation. In practice, however, we will show next that the stationary point reached by the Single Loop in Algorithm 6 shows an irreducible gap between the moments of $q(\mathbf{u})$ and $r(\mathbf{u})$. Our goal in this section is to illustrate how the quality of the ECD approximation depends on the update rules implemented during the Single Loop and it introduces a modification that achieves a better tradeoff between accuracy and complexity. We emphasize that the quality of the approximation is measured in terms of moment matching between tractable approximations to $p(\mathbf{u}|\mathbf{y})$ ($q(\mathbf{u})$ and $r(\mathbf{u})$ respectively), and not with regards to the distribution $p(\mathbf{u}|\mathbf{y})$ itself. We analyze the evolution of the following two quantities along iterations

$$\Delta_u = \frac{1}{2m}\sum_{i=1}^{2m}\Big|\mathbb{E}_q[u_i] - \mathbb{E}_r[u_i]\Big|, \tag{V.35}$$

$$\Delta_{u^2} = \frac{1}{2m}\sum_{i=1}^{2m}\Big|\mathbb{E}_q[u_i^2] - \mathbb{E}_r[u_i^2]\Big|. \tag{V.36}$$

In Figure V.1 we represent Δ_u and Δ_{u^2} for a $m = r = 5$ scenario with 4-QAM modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. According to Figure IV.1(a), this SNR value is far from the saturation regime (largest gap to channel capacity), and it is in this range where we want the ECD to work and improve state-of-the-art methods. Three implementations of the Single Loop are compared in Figure V.1. For the red solid line we have used $\beta = 0.2$, i.e., a very slow parameter update in Step 6) of Algorithm 6. This is exactly the EPD algorithm implemented in Chapter III. The opposite case is represented by the green dotted line, which has been computed with $\beta = 0.95$. While the $\beta = 0.2$ case achieves a better approximation, i.e. smaller Δ_u and Δ_{u^2} values, it requires in average 25 iterations to converge to such a stable point. Recall that each Single Loop is as complex as computing the MMSE estimate, due to the matrix inversion in (V.21). On the other hand, the $\beta = 0.95$ case quickly saturates (around 5 iterations), but its solution is actually worse. In order to achieve a better tradeoff between accuracy and complexity, we maintain the fast updates using $\beta = 0.95$, but modify the parameter update in Algorithm 6 and introduce a gradual decrease in the variance per component allowed at each iteration. More precisely, we set an iteration-dependent minimum value of the variance $\mathbb{E}_{s(\mathbf{u})}[u_i^2]$ at Step 6)



(a)

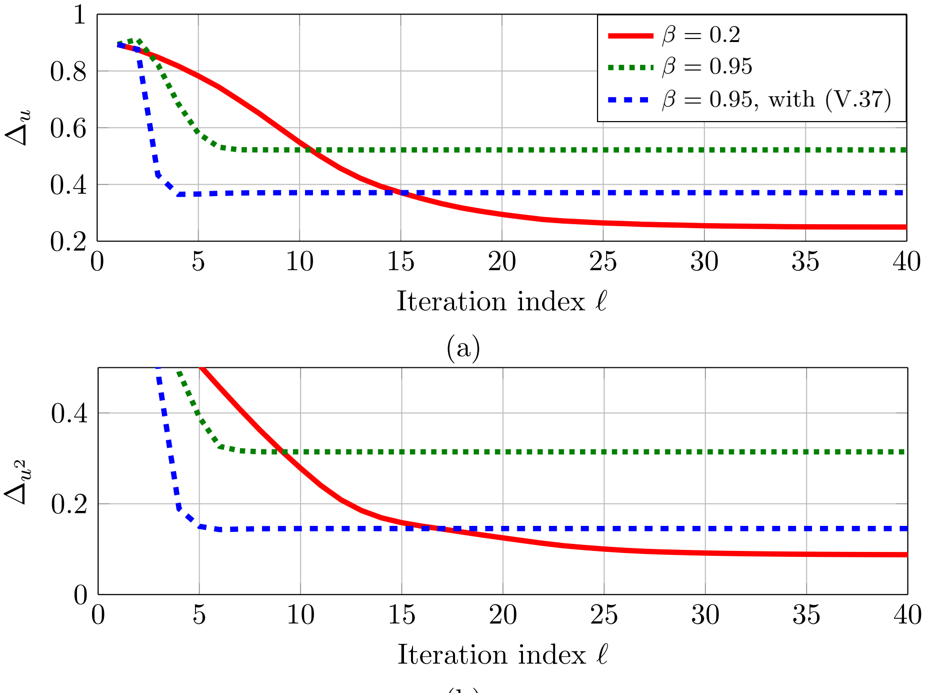
(b)

Figure V.1: Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 5$ scenario with 4-QAM modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$.

of Algorithm 6) with the following form:

$$\mathbb{E}_{s(\mathbf{u})}[u_i^2] = \max\left(2^{-\max(\ell-3,0)}, \mathbb{E}_{r(\mathbf{u})}[u_i^2]\right), \tag{V.37}$$

namely during the first 3 iterations we set a reasonably minimum high variance per component (0.5) and, from iteration 4, we let this minimum value to decrease exponentially fast with ℓ. Setting a high-variance parameter during the first iterations of the algorithm is crucial in the low-SNR regime in order to avoid over-fitting. Namely, provided that the MMSE solution (initialization of our algorithm) is a poor estimate at high-noise levels, setting a large value for ϵ during the first iterations prevents the algorithm to provide unreasonable confident estimates, which would also restrain the EC algorithm to move far away from the MMSE estimate. The convergence of this implementation of the EC algorithm is represented in In Figure V.1 with dashed lines. Observe that an improvement is achieved with respect to the $\beta = 0.95$ case, reducing the gap to the stationary point achieved



(a)

(b)

Figure V.2: Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 32$ scenario with 4-QAM modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$.

by $\beta = 0.2$, without a significant penalty in speed of convergence, as it typically converges in less than 10 iterations. These effects are even more evident when we move to higher-dimensional scenarios. In Figure V.2 we consider a $m = r = 32$ scenario with 4-QAM modulation at an SNR of 6dB, and in Figure V.3 we consider a $m = r = 32$ scenario with 16-QAM modulation at an SNR of 13dB. Convergence speed is actually maintained and the gap with respect to $\beta = 0.2$ case is clearly reduced. Using the ECD moment matching criterion many other variants of the single loop update methods can be tested and compared with our proposal. However, no significant differences have been appreciated when we measure the system performance in terms of the mutual information in (I.14) or BER. In the following results, regardless of the dimension of the system or constellation order, we implement the ECD detector using $\beta = 0.95$, the progressive variance limit in (V.37) and a maximum number of iterations of $I = 10$. Thus, the complexity order the ECD detector implemented is roughly 10 times the MMSE complexity.

V.4 Experimental Results

In the following, we include simulation performance results that show the accuracy of the ECD approximation. In our experiments, we compare our proposal with the soft output MMSE solution in [20, 27], the GTA algorithm in [50] and the CHEMP method in [78].

V.4.1 Calibration Curves

In Figures V.4 and V.5 we show calibration curves, where we compare the true marginal symbol posterior probability $p(u_i|\mathbf{y})$ in (I.11) with the one estimated by each method. The obtained curves are shown for SNR = 7 dB and SNR = 13.5 dB respectively. These figures have been generated using 2500 random channel matrices and transmitting five MIMO symbol vectors per channel matrix $\mathbf{H}$. For each MIMO symbol vector, we include in the plot

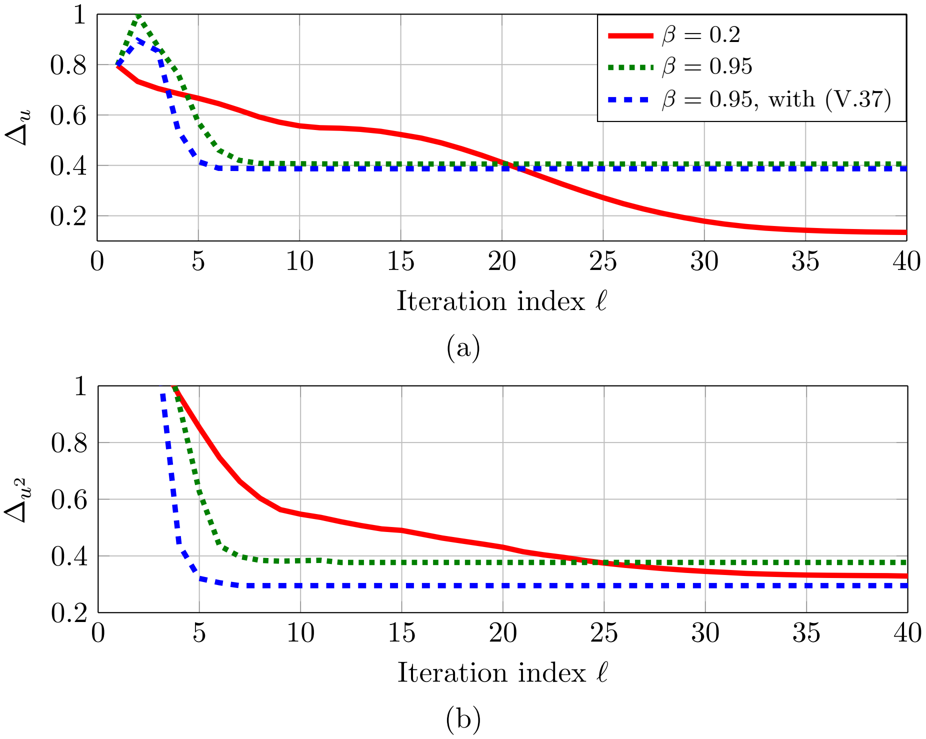
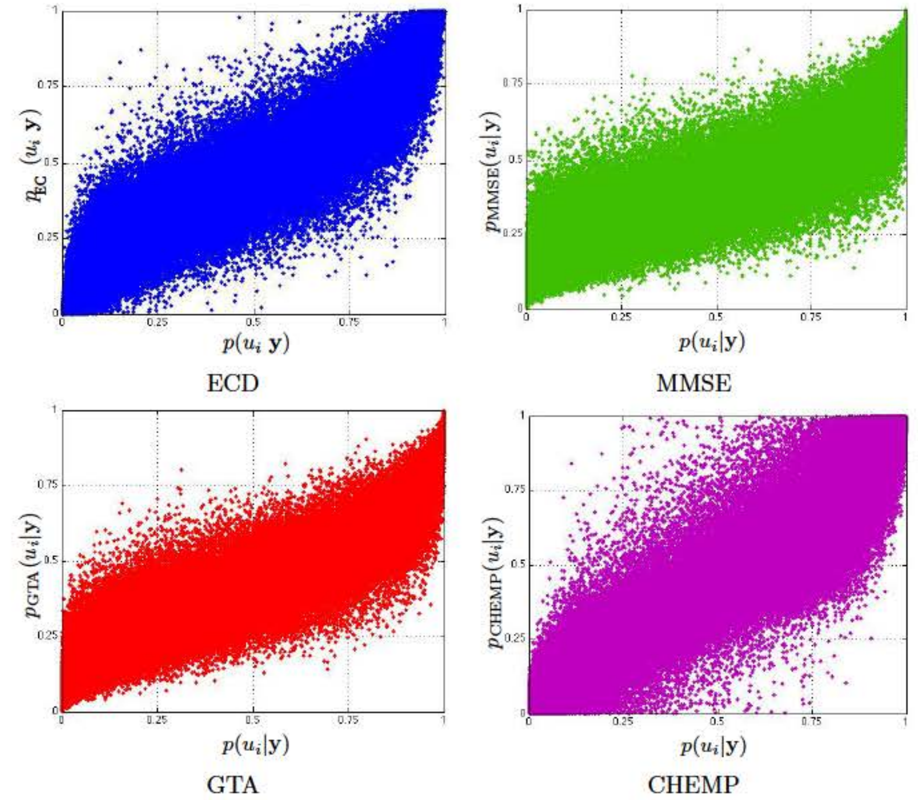


Figure V.3: Δ_u and Δ_{u^2} in (V.35) and (V.36) for a $m = r = 32$ scenario with 16-QAM modulation at a SNR of 13dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$.

$p(u_i|\mathbf{y})$ and its approximation for every transmitted antenna, i.e., $\forall i \in \{2m\}$ and for every symbol in $\mathcal{A}$. Thus, in total every plot contains $M \times m \times 2500 \times 5 = 2.5 \cdot 10^5$ points. Observe that when the noise variance is large,

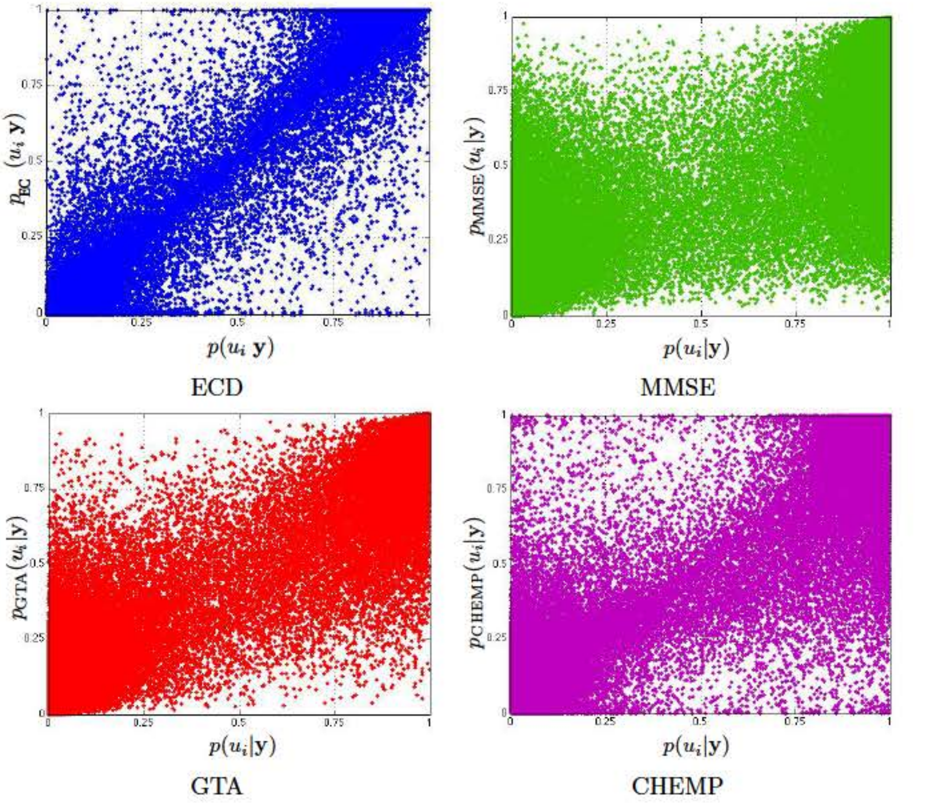


ECD MMSE

GTA CHEMP

Figure V.4: Calibration curves in $m = r = 5$ system with a 4-QAM at $\text{SNR} = 7\text{dB}$.

the calibration curves for ECD, GTA and CHEMP are roughly diagonal. At low-SNR regime, the joint posterior probability $p(\mathbf{u}|\mathbf{y})$ is dominated by the Gaussian noise and hence it presents an unimodal Gaussian-like shape, where approximate inference methods typically perform well. As we increase the SNR, note that only ECD is able to maintain and approximate diagonal calibration curve while the rest of methods present a large dispersion in the corners ($p(u_i|\mathbf{y})$ close to either zero or one). Compared to ECD (and also to GTA), the CHEMP solution for high SNR presents a significant density of points in the upper left and bottom right quadrants of the calibration curve, where either $p(u_i|\mathbf{y})$ tends to zero and $p_{\text{CHEMP}}(u_i|\mathbf{y})$ to one or the other way around. These errors are particularly harmful for the LDPC decoding stage.



ECD

MMSE

GTA

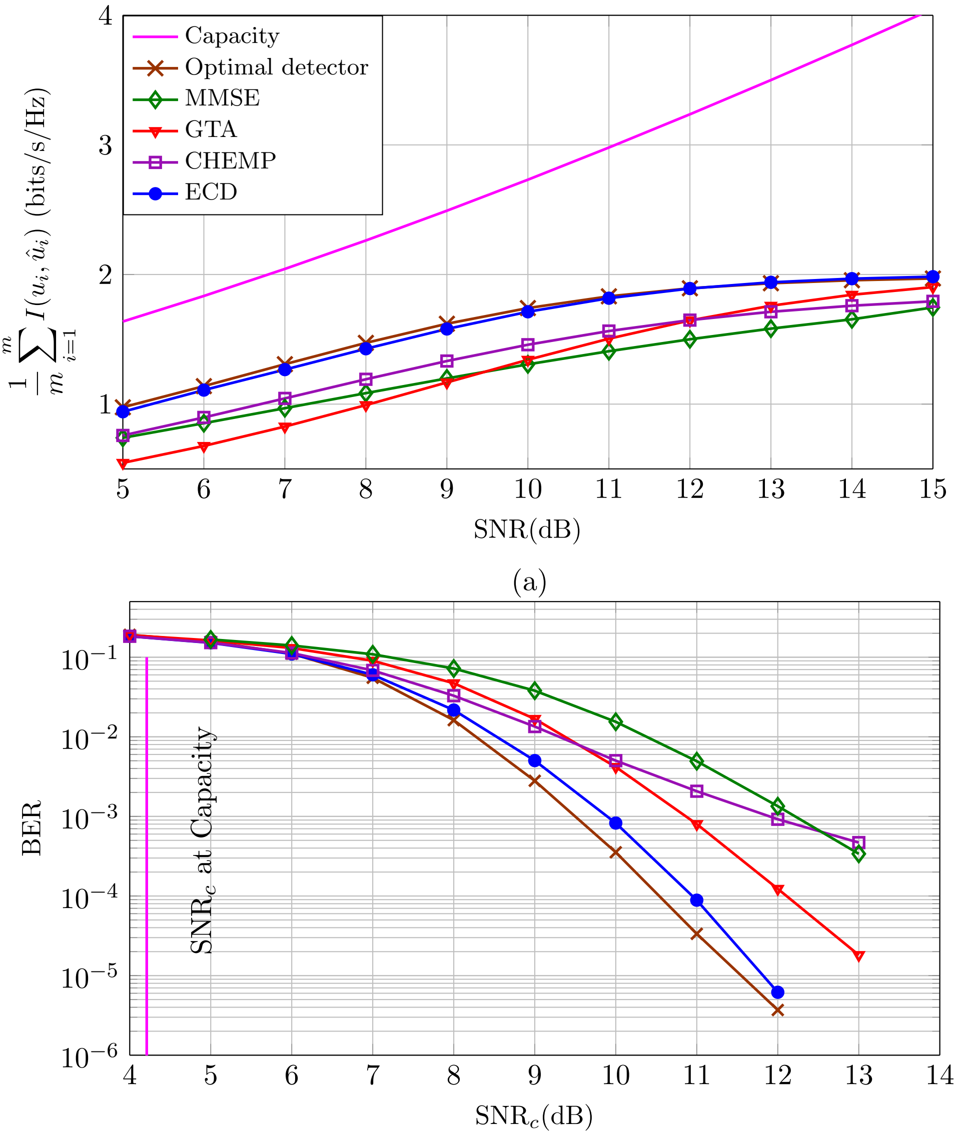
CHEMP

Figure V.5: Calibration curves in $m = r = 5$ system with a 4-QAM at SNR = 13.5dB.

V.4.2 A Low Dimensional MIMO System

We consider again the $m = r = 5$ scenario with 4-QAM modulation already considered in Figure IV.1 and extended in Figure V.6(a) with the achievable rate of the ECD MIMO detector. Remarkably, it essentially overlaps the optimal detector performance, achieving a large gain with respect to GTA, MMSE and CHEMP. When the number of antennas is small (5 in our case), the columns of the channel matrix $\mathbf{H}$ are typically non-orthogonal and this limits the MMSE performance [20, 27]. Also, as discussed in Section IV.2, the CHEMP method relies on the matrix $\mathbf{J} = m^{-1}\mathbf{H}^{\top}\mathbf{H}$ being diagonal and for a small m, this assumption is unrealistic.

Results in Figure V.6(a) indicate that the MIMO system performance can highly benefit from the more accurate estimates to the symbol posterior marginals $p(u_i|\mathbf{y})$ provided by ECD. To corroborate this fact, we include an LDPC channel encoding stage at the transmitter and an LDPC channel decoder at the receiver (System Model I.3). The LDPC channel decoder is fed by soft coded bit probabilities computed using the symbol posterior marginals $p(u_i|\mathbf{y})$ (or their approximations), according to the bit-modulation mapping. It is well known that the more accurate the probabilistic detector is, the better performance is obtained after the LDPC decoding stage using BP [27–29]. In Figure V.6(b), we show for this scenario the simulated BER measured after the LDPC decoding stage (solid lines). A $(3, 6)$-regular LDPC code with block length $k = 5120$ bits has been used. In terms of coded performance, the gap between optimal detection and ECD is only about 0.4 dB measured at a BER of 10^{-4} while the gap to GTA is over 1.5 dB. Note that both MMSE and CHEMP provide poor performance, which slowly decreases with $\mathrm{SNR_c}$.



(a)

(b)

Figure V.6: $m = r = 5$ system with 4-QAM modulation, in (a) achievable transmission rates. In (b), with a $(3, 6)$-regular LDPC code with block length $k = 5120$ bits.

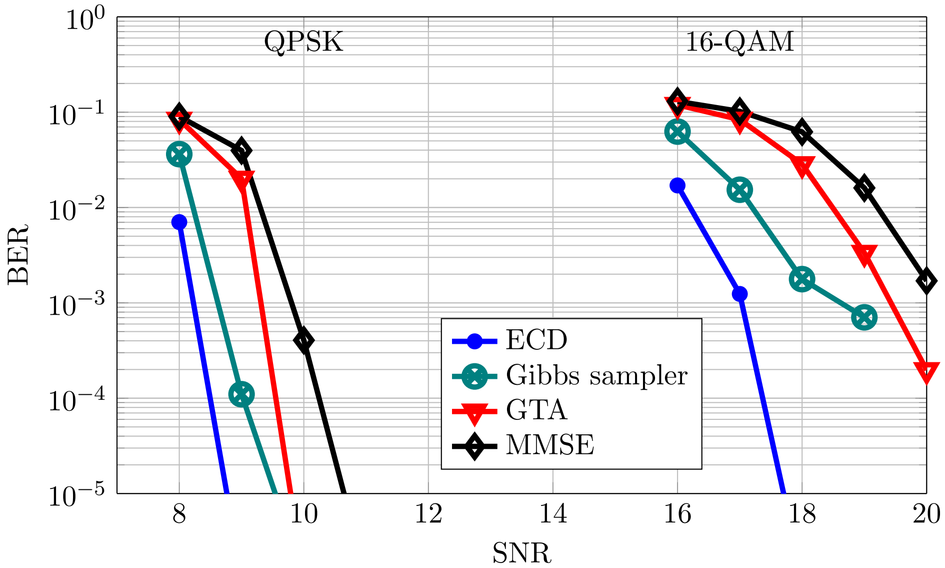


Figure V.7: $m = r = 32$ MIMO system with 4-QAM and 16-QAM constellation. A $(3, 6)$-regular LDPC with code block $k = 10240$ bits has been used.

V.4.3 A $m = r = 32$ MIMO system

In a larger scenario, exact marginalization is not viable anymore and we rely on approximate methods. Recovering the larger scenario from Chapter IV, we include the ECD performance. We show in Figure V.7 that even for the simplest case, 4-QAM, the MCMC performance seems to degrade at high-SNR regime. This effect is more severe in 16-QAM case, as MCMC typically gets trapped in a local mode. Works on MCMC MIMO detection propose different heuristic methods to compensate for this behavior, but ultimately the only way is to increase both the number of samplers that are run in parallel and the number of samples generated by each one of them. Additionally, the result remarks the performance provided by ECD.

In Figures V.8 to V.10, we represent the obtained achievable rates for a $m = r = 32$ MIMO system using 4-QAM modulation, 16-QAM modulation, and 64-QAM modulation respectively. While CHEMP and ECD are competitive for the 4-QAM case, CHEMP is no longer a viable option in the 16-QAM or 64-QAM cases. As discussed in Section IV.2, the variance of the interference noise that CHEMP aims to iteratively cancel grows with the constellation order. For $m = r$ and high order constellations the interference noise becomes excessively large.

Following [78], it can be checked that CHEMP becomes effective again

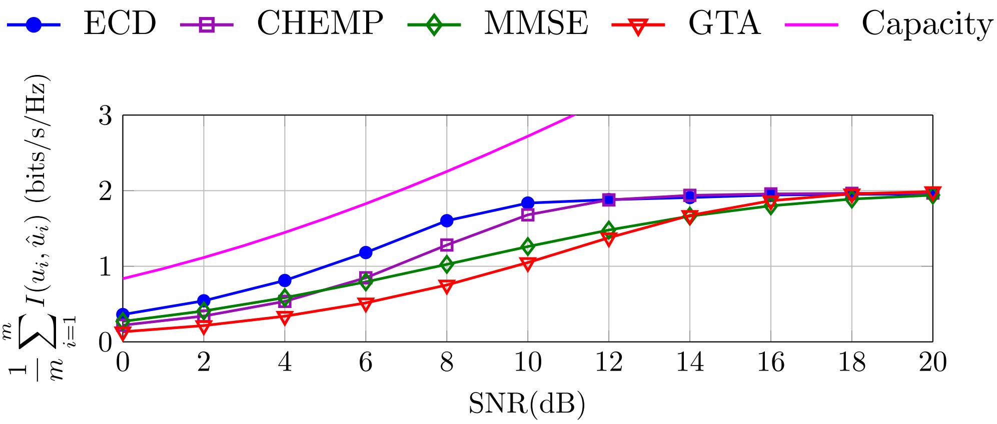


Figure V.8: Acheivable rate computed for a $m = r = 32$ MIMO system with 4-QAM.

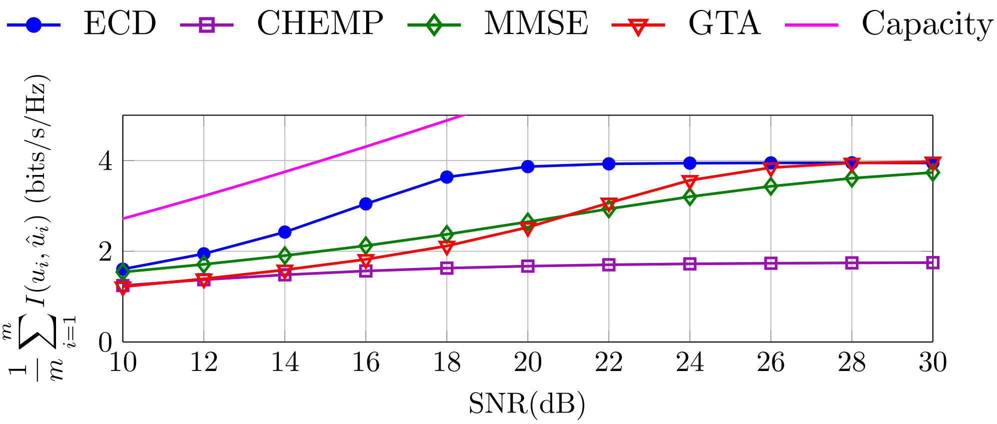


Figure V.9: Acheivable rate computed for a $m = r = 32$ MIMO system with 16-QAM.

as we reduce the number of transmitting antennas, i.e., if $m < r$. In Figure V.11 (a), we compare the ECD and CHEMP transmission rates for a 16-QAM modulation with $r = 32$ and $m = 16$, 20, 25 and 32. In (b), we include BER simulation results using the $(3, 6)$-regular LDPC with code block length $k = 5120$ bits. For small m values, CHEMP is comparative to the EC based solution. However, its performance severely degrades as m approaches r. CHEMP can be regarded as a Gaussian message-passing distributed implementation of the EC algorithm for those cases where interference is "locally" tractable. Unlike CHEMP, the ECD algorithm performs the update of all parameters at the same time in a centralized manner.

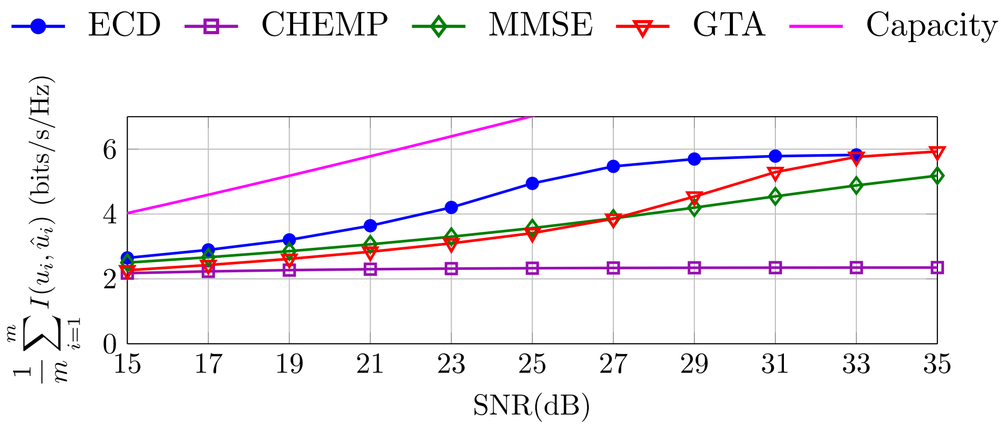


Figure V.10: Acheivable rate computed for a $m = r = 32$ MIMO system with 64-QAM.

These results show that ECD is robust against the increase in the constellation order. In the following we solely consider $m = r$ scenarios with high order constellations and hence we omit CHEMP from the results.

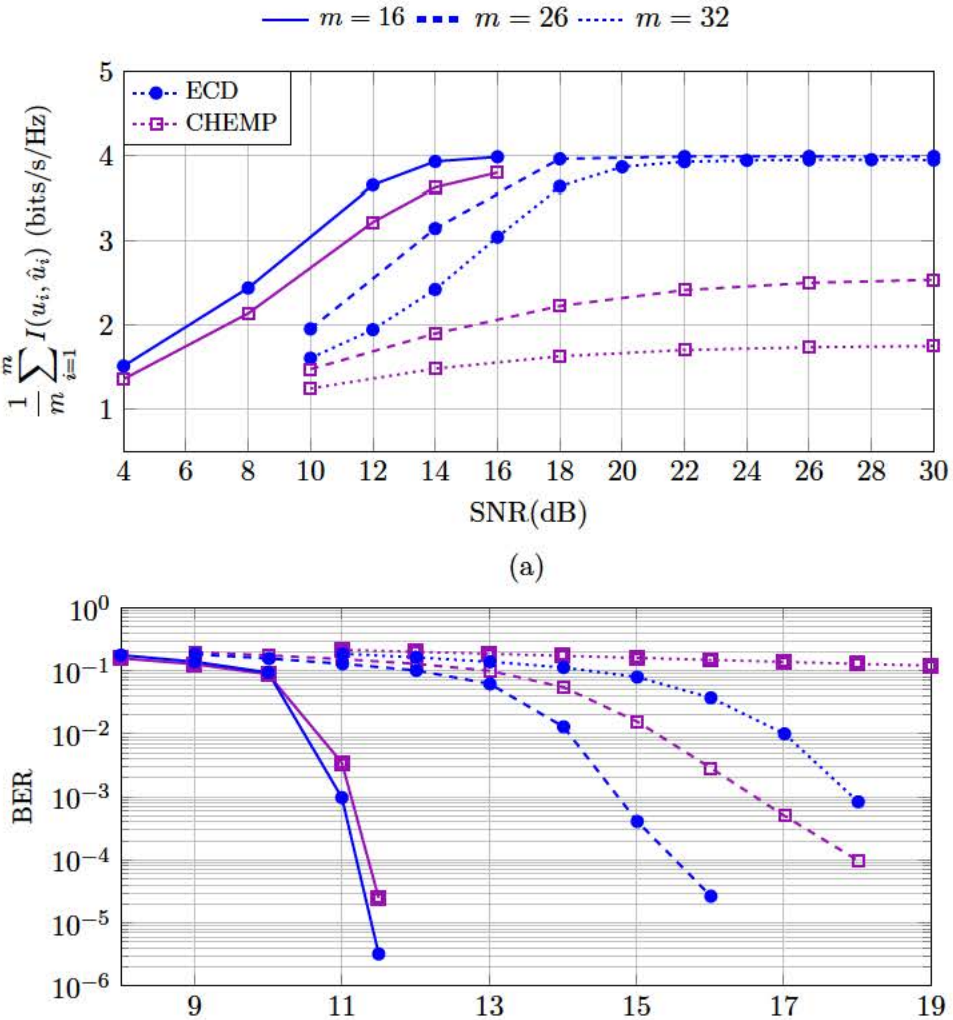
We complete the analysis of this scenario in Figure V.12, by including BER performance results using LDPC constructions that are designed to improve the performance of the $(3, 6)$-regular LDPC code used in previous experiments. In dashed lines we show the performance of the rate-1/2 irregular LDPC code[2] with block length $k = 30720$ bits. We also include simulation results (solid lines) for a Convolutional Low-Density Parity-Check (LDPCC) code constructed by spatially-coupling 48 independent copies of a $(3, 6)$-regular LDPC code, each having block length $k = 640$ bits, with low-rate terminations [93]. The resulting coding rate is 0.479 and a total block length $k = 30720$ bits[3]. For the irregular LDPC code, at moderate SNRs, ECD is able to provide a significant gain, which vanishes at high-SNR regime because of the LDPC error floor. In contrast, because the LDPCC code has large minimum distance, no error floor has been observed in the range of SNRs considered and ECD achieves a stable gain of 2.5 dB with respect to GTA.

V.4.4 Feedback helps for intermediate LDPC block lengths

The BP algorithm for LDPC decoding recomputes a probability for each coded bit that takes into account the correlations imposed by the LDPC

[2] The code is generated using ([76], Example 3.99).

[3] The code is generated using protographs [94] in order to optimize its minimum distance, as described in [95].



(a)

(b)

Figure V.11: $r = 32$ MIMO system with 16-QAM modulation, in (a) achievable transmission rates for different m values. In (b), BER performance when a $(3, 6)$-regular LDPC with code block $k = 5120$ bits.

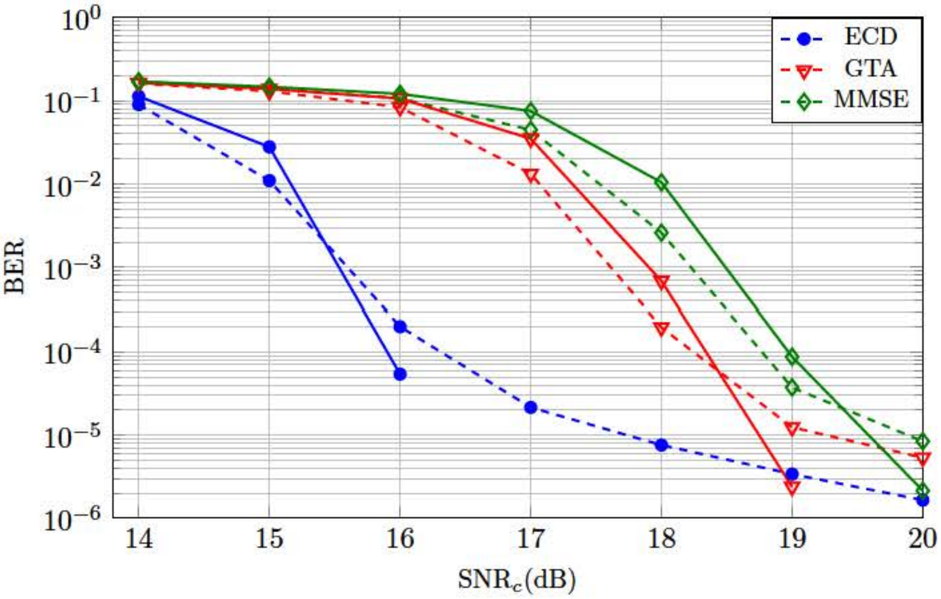


Figure V.12: BER performance of a $m = r = 32$ system with 16-QAM, using the irregular (dashed lines), $R = 1/2$ LDPC with code block $k30720$ bits and a $(3, 6)$-regular (solid lines) LDPCC with the same block and coding rate 0.479.

code. We can use such an estimate to recompute a prior probability for each transmitted symbol, in order to provide to the EC MIMO detector in Algorithm 6 with a refined initialization (different from the default MMSE initialization). Denote the estimation computed this way by $p_{\mathrm{BP}}(u_i)$ for any $u_i \in \mathcal{A}$, $i \in \{2m\}$. Note that the transmission of an LDPC codeword require in general the transmission of several symbol vectors through the (known) MIMO channel, each of them independently equalized by the ECD. The re-initialization that we propose for ECD is performed similarly for all cases. Given $p_{\mathrm{BP}}(u_i)$ for any $u_i \in \mathcal{A}$, we run Algorithm 6 with the following values

$$\Lambda_{q,i}^{(0)} = \left(\mathbb{E}_{p_{\mathrm{BP}}(u_i)}[u_i^2]\right)^{-1}, \tag{V.38}$$

$$\gamma_{q,i}^{(0)} = \Lambda_{q,i}^{(0)} \mathbb{E}_{p_{\mathrm{BP}}(u_i)}[u_i]. \tag{V.39}$$

for $i \in \{2m\}$. Each time we call the ECD stage using the BP LDPC soft output is referred to as a "loop". In Figure V.13 we include the BER measured after the LDPC decoding stage after two loops for $m = r = 32$ system with 16-QAM modulation. A $(3, 6)$-regular LDPC with code block $k = 1280$ bits (solid lines) and $k = 10240$ bits (dashed lines) have been used. The ECD has been run with 10 iterations in all cases, despite we have observed that after the first loop the number of iterations can be reduced as the initialization is better. In all cases, observe that feedback indeed improves the performance and that the highest gain is achieved after the first loop. However, such a gain quickly decreases with the block length, as it is around 1.2 dB for $k = 1280$ bits (measured at 10^{-4}) and less than 0.1 dB already for $k = 10240$ bits. As we consider large block lengths, the BP LDPC decoding probabilistic outputs tend to be extreme, i.e., estimated probabilities are close to either zero or one [76], preventing the ECD stage to substantially diverge from this point.

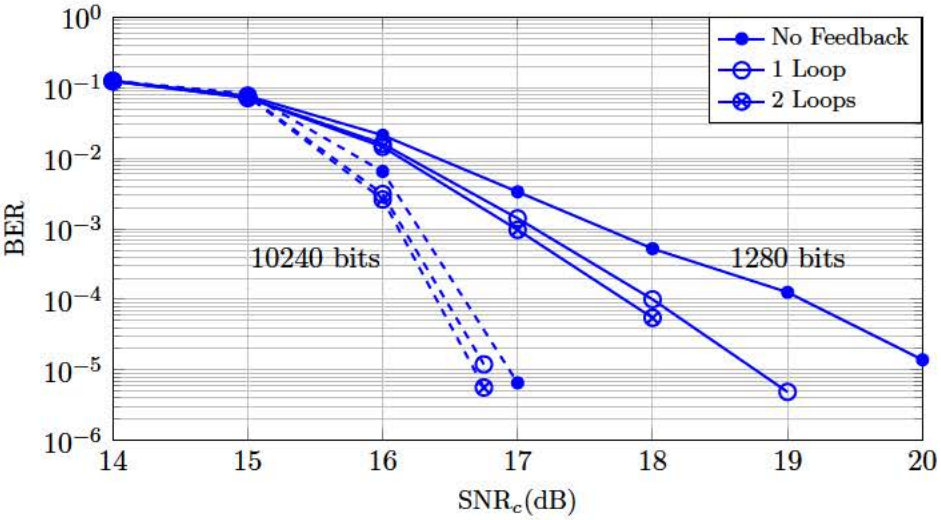


Figure V.13: ECD performance in a $m = r = 32$ system with 16-QAM using the feedback loop. We consider a $(3, 6)$-regular LDPC code with different k lengths.

The use of a more-complex receiver scheme that implements the feedback loop can be justified for those applications where power and delay constraints limit the block length that users at the transmitter side can afford. A common scenario is the uplink in mobile wireless communications, where the m users, each one using a single antenna, independently encode its information stream using a channel code with a moderate block length, around a few hundred bits. We emphasize here that the design of a receiver with a feedback loop between channel equalizer and channel decoder is not trivial [96] and that the proposed scheme should be regarded as a proof-of-concept, rather than a final solution. The main conclusion that we draw here is that the ECD is accurate enough when combined with a sufficiently long channel code so the use of a feedback loop may not be necessary.

Chapter VI

Conclusions

When the number of antennas at the transmit side is increased, the design of efficient MIMO receivers is a challenging problem due to the high-dimensional discrete input space. Classical solutions either reduce the search space (Sphere decoding or Tabu search methods) or proyect the problem into a continuous space to solve a quadratic problem (MMSE) or to construct a Gaussian-like approximations (GTA or CHEMP). However, on one hand, SpD or TS badly scale with the number of antennas. On the other hand, MMSE presents poorer performance but keeps the complexity cubic with respect to the number antennas at the transmit side. Recent methods such as GTA or CHEMP are the state-of-the-art detectors and are able to perform both hard symbol detection and to provide probabilistic information. While GTA constructs a tree-factorized Gaussian approximation to the posterior pdf, CHEMP is a message-passing inference algorithm. As we have shown, by callibration curves or BER results, the accuracy of the probabilistic MIMO detector is crucial to the overall performance.

In this Thesis, Machine Learning techniques based on moment matching are proposed to solve the inference process. EP is a powerful approximate inference method to construct tractable approximations to a given pdf. In experimental results, we show that it outperforms state-of-the-art methods. We consider several scenarios from small to a very large number of antennas. The results show that our proposal is robust for all constellation orders. Furthermore, by keeping fix the constellation order and increasing the number of antennas of the system, we demonstrate that EPD suffers less degradation than its competitors and thus its scales much better for massive MIMO scenarios.

While the presented results for MIMO hard detection with EPD significantly improve state-of-the-art, our second focus is to develop effective methods for probabilistic MIMO symbol detection, as it can be combined with modern capacity-achieving channel codes. The EPD implementation proposed is shown to be effective only for mean estimation in the high-SNR regime. However, we found difficult to properly tune the algorithm to obtain accurate covariance estimates in the low-to-moderate SNR regimen. This task turned to be cumbersome, due to its heuristic nature. To overcome this problem, we have proposed EC approximation inference methodology, which generalizes EP and provides a framework to evaluate convergence. Using the EC point of view, we propose algorithms that are able to accurately estimate the marginal distribution for each transmitted symbol, i.e. in a Gaussian approximation through its mean and variance. With this information, soft-bit inference can be performed to feed the channel decoder. Furthermore, we show that the system implementing ECD is able to achieve a transmission rate closer to capacity than state-of-the-art methods for probabilistic detection.

It should be noted that our work deals with MIMO scenario where we have a high-dimensionality of the transmitted constellation, and the detection process intends to recover all transmitted symbols. There are several MIMO scenarios matching our premises. One example could be uplink MU-MIMO scenario in which several users try to transmit to a Base Station (BS) with at least the same number of receive antennas as the sum of the transmit antennas, otherwise the performance is highly degraded. The BS needs to detect all those symbols transmitted at the same time sharing frequency. Another possibility is a downlink SU-MIMO scenario in which a BS with several antennas can use some of them to transmit to a user, the number of transmit antennas should be as much as the number of receiver antennas. The user needs to detect the symbols transmitted by those antennas assigned to it.

Future Lines

Several future research lines can be proposed with the actual-research state.

- During this Thesis perfect CSI was assumed for the receiver. However, this assumption can be unrealistic. Analyzing the behavior of the proposed methods when there is uncertainty on the CSI is an important problem. A further step on this direction can be done by assuming a pdf over the channel matrix $\mathbf{H}$.

- The design of a code to reach channel capacity using EC estimates is one important open problem if we would like to achieve the gains in transmission rate predicted by mutual information plots.

- A deep study of the Turbo-like architecture should be performed, as many different parameters are involved in the proposed architecture. In particular, the achievable rate should be analyzed with regard to the ECD parameters and the number of feedback iterations. Increasing the number of feedbacks may reduce the number of ECD iterations I, and eventually this behavior may not be constant for all length codes.

- The application of the EC algorithm to perform fast precoding design and power allocation in MIMO system replacing current implementations based on MCMC approximations [97].

- Finally, an important step forward to the communication industry in order to reach throughput promises [98], can be to test the already mentioned application in a simulation framework, within a cellular architecture. This will provide a more general perspective of the key strengths of the proposed receiver.

Author Bibliography

[39] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation Propagation Detection for High-Order High-Dimensional MIMO Systems", *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014. [Online]. Available: `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6841617`.

[40] ——, "Probabilistic MIMO Symbol Detection with Expectation Consistency Approximate Inference ", *Submitted to IEEE Transactions on Vehicular Technologies*, Oct. 2016.

[86] ——, "Improved performance of LDPC-coded MIMO systems with EP-based soft-decisions", in *IEEE International Symposium on Information Theory*, (Honolulu, USA), Jun. 2014, pp. 1997–2001. [Online]. Available: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6875183`.

Bibliography

[1] P. Robertson, E. Villebrun, and P. Hoeher, "A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain", in *IEEE International Conference on Communications*, (Seattle, USA), vol. 2, Jun. 1995, pp. 1009–1013.

[2] L. Zheng, P. Viswanath, and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels", *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1095, May 2003.

[3] A. Goldsmith, *Wireless Communications.* Cambridge University Press, 2005.

[4] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays.", *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[5] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband.* Academic press, 2010.

[6] "IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC)and Physical Layer (PHY) Specifications Amendment 5: Enhancements for Higher Throughput", *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pp. 1–565, Oct. 2009.

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

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

[9] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, “An overview of MIMO communications-a key to gigabit wireless”, *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198–218, Feb. 2004.

[10] D. Gesbert, M. Shafi, D. S. Shiu, P. J. Smith, and A. Naguib, “From theory to practice: an overview of MIMO space-time coded wireless systems”, *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3, pp. 281–302, Apr. 2003.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[12] J. R. Hampton, *Introduction to MIMO communications.* Cambridge University Press, 2013.

[13] A. Lozano and N. Jindal, “Transmit diversity vs. spatial multiplexing in modern MIMO systems”, *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 186–197, Jan. 2010.

[14] G. D. Golden, C. J. Foschini, R Valenzuela, and P. W. Wolniansky, “Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture”, *Electronics Letters*, vol. 35, no. 1, pp. 14–16, Jan. 1999.

[15] J. Hoydis, S. ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?”, *IEEE Journal on Selected Areas in Communications,*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[16] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An Overview of Massive MIMO: Benefits and Challenges”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[17] E. G. Larsson, “MIMO detection methods: How they work [lecture notes]”, *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 91–95, May 2009.

[18] M. Lu, T. Lo, and J. Litva, “A physical spatio-temporal model of multipath propagation channels”, in *IEEE Vehicular Technology Conference*, (Phoenix, USA), vol. 2, May 1997, pp. 810–814.

[19] M. Bayes and M. Price, “An essay towards solving a problem in the doctrine of chances.”, *Philosophical Transactions (1683-1775)*, pp. 370–418, 1763.

[20] G. Caire, R. R. Muller, and T. Tanaka, “Iterative multiuser joint decoding: optimal power allocation and low-complexity implementation”, *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 1950–1973, Sep. 2004.

[21] K. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A Low-Complexity Detector for Large MIMO Systems and Multicarrier CDMA Systems", *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 3, pp. 473–485, Apr. 2008.

[22] P. Li and R. D. Murch, "Multiple output selection-LAS algorithm in large MIMO systems", *IEEE Communications Letters*, vol. 14, no. 5, pp. 399–401, May 2010.

[23] L. Dai, X. Gao, X. Su, S. Han, C. L. I, and Z. Wang, "Low-Complexity Soft-Output Signal Detection Based on Gauss Seidel Method for Uplink Multiuser Large-Scale MIMO Systems", *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4839–4845, Oct. 2015.

[24] P. D. Bai Lin and J. Choi, *Low complexity MIMO detection.* Springer Verlag, 2012.

[25] R. Gallager, "Low-density parity-check codes", *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1962.

[26] C. Berrou and A. Glavieux, "Turbo codes", *Encyclopedia of Telecommunications*, 2003.

[27] A. Sanderovich, M. Peleg, and S. Shamai, " LDPC coded MIMO multiple access with iterative joint decoding", *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1437–1450, Apr. 2005.

[28] P. M. Olmos, J. J. Murillo-Fuentes, and F. Pérez-Cruz, "Joint nonlinear channel equalization and soft LDPC decoding with Gaussian processes", *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1183–1192, Mar. 2010.

[29] A. G. D. Uchoa, R. C. D. Lamare, and C. Healy, "Iterative Detection and Decoding Algorithms For Block-Fading Channels Using LDPC Codes", in *IEEE Wireless Communications and Networking Conference*, (Istambul, Turkey), Apr. 2014, pp. 815–820.

[30] D. Silva, F. R. Kschischang, and R. Kotter, "Communication Over Finite-Field Matrix Channels", *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1296–1305, Mar. 2010.

[31] Y. Katayama and S. Morioka, "One-shot Reed-Solomon decoding for high-performance dependable systems", in *International Conference on Dependable Systems and Networks*, (New York, USA), May 2000, pp. 390–399.

[32] Y. Li, L. Wang, and Z. Ding, "An Integrated Linear Programming Receiver for LDPC Coded MIMO-OFDM Signals", *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2816–2827, Jul. 2013.

[33] E. Telatar, “Capacity of multi-antenna Gaussian channels”, *European Transactions on Telecommunication*, vol. 10, no. 6, pp. 585–596, Nov. 1999.

[34] R. Motwani and P. Raghavan, *Randomized algorithms.* Chapman & Hall/CRC, 2010.

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

[36] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[37] T. P. Minka, “Expectation propagation for approximate Bayesian inference”, in *Conference on Uncertainty in Artificial Intelligence*, (Seattle, USA), Aug. 2001, pp. 362–369.

[38] M. W. Seeger, “Expectation propagation for exponential families”, Tech. Rep., 2005.

[41] U. Fincke and M. Pohst, “Improved methods for calculating vectors of short length in a lattice, including a complexity analysis”, *Mathematics of computation*, vol. 44, no. 170, pp. 463–471, Apr. 1985.

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

[43] A. Burg, M. Borgmann, C. Simon, M. Wenk, M. Zellweger, and W. Fichtner, “Performance tradeoffs in the VLSI implementation of the sphere decoding algorithm”, in *IEE International Conference on 3G Mobile Communication Technologies*, (London, England), Oct. 2004, pp. 93–97.

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

[45] U. Madhow and M. L. Honig, “MMSE interference suppression for direct-sequence spread-spectrum CDMA”, *IEEE Transactions on Communications*, vol. 42, no. 12, pp. 3178–3188, Dec. 1994.

[46] M. Honig, U. Madhow, and S. Verdu, “Blind adaptive multiuser detection”, *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 944–960, Jul. 1995.

[47] Y. Li and R. Liu, “Adaptive Blind Source Separation and Equalization for Multiple-Input/Multiple-Output Systems”, *IEEE Transactions on Information Theory*, vol. 44, pp. 2864–2876, Nov. 1998.

[48] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, "MMSE decision-feedback equalizers and coding. I. Equalization results", *IEEE Transactions on Communications*, vol. 43, no. 10, pp. 2582–2594, Oct. 1995.

[49] T. h. Liu and Y. L. Liu, "Modified fast recursive algorithm for efficient MMSE-SIC detection of the V-BLAST system", *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3713–3717, Oct. 2008.

[50] J. Goldberger and A. Leshem, " MIMO Detection for High-Order QAM Based on a Gaussian Tree Approximation", *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4973–4982, Aug. 2011.

[51] J. Goldberger, "Improved MIMO Detection based on Successive Tree Approximations", in *IEEE International Symposium on Information Theory*, (Istanbul, Turkey), Jul. 2013, pp. 2004–2008.

[52] ——, *GTA C code*, `http://www.eng.biu.ac.il/~goldbej/papers/MIMO_source_files.rar`, 2013.

[53] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization", *IEEE Transactions on Communications*, vol. 56, no. 2, pp. 309–318, Feb. 2008.

[54] Y. H. Gan and W. H. Mow, "Complex lattice reduction algorithms for low-complexity MIMO detection", in *IEEE Global Communications Conference*, (St. Louis, USA), vol. 5, Dec. 2005, pp. 2953–2957.

[55] K. A. Singhal, T. Datta, and A. Chockalingam, "Lattice reduction aided detection in large-MIMO systems", in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, (Darmstadt, Germany), Jun. 2013, pp. 594–598.

[56] J. Boutros, N. Gresset, L. Brunel, and M. Fossorier, "Soft-input soft-output lattice sphere decoder for linear channels", in *IEEE Global Communications Conference*, (San Francisco, USA), vol. 3, Dec. 2003, pp. 1583–1587.

[57] D. Wübben, D. Seethaler, J. Jaldén, and G. Matz, "Lattice Reduction", *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 70–91, May 2011.

[58] Q. Zhou and X. Ma, "Element-Based Lattice Reduction Algorithms for Large MIMO Detection", *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 274–286, Feb. 2013.

[59] H. Zhao, H. Long, and W. Wang, "Tabu Search Detection for MIMO Systems", in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, (Athens, Greece), Sep. 2007, pp. 1–5.

[60] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Random-Restart Reactive Tabu Search Algorithm for Detection in Large-MIMO Systems", *IEEE Communications Letters*, vol. 14, no. 12, pp. 1107–1109, Dec. 2010.

[61] N. Srinidhi, T. Datta, A. Chockalingam, and B. S. Rajan, "Layered Tabu Search Algorithm for Large- MIMO Detection and a Lower Bound on ML Performance", *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 2955–2963, Nov. 2011.

[62] M. W. Seeger, "Bayesian Inference and Optimal Design for the Sparse Linear Model", *Journal of Machine Learning Research*, vol. 9, pp. 759–813, Jun. 2008.

[63] M. Opper and O. Winther, "Expectation Consistent Approximate Inference", *Journal of Machine Learning Research*, vol. 6, pp. 2177–2204, Dec. 2005.

[64] S. L. Lauritzen, "Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models", *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1098–1108, Dec. 1992.

[65] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference.* Foundations and Trends in Machine Learning, 2008.

[66] P. M. Olmos, J. J. Murillo-Fuentes, and F. Pérez-Cruz, "Tree-Structure Expectation Propagation for LDPC Decoding Over the BEC", *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3354–3377, Jun. 2013.

[67] L. Salamanca, P. M. Olmos, F. Perez-Cruz, and J. J. Murillo-Fuentes, "Tree-structured expectation propagation for LDPC decoding over BMS channels", *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4086–4095, Oct. 2013.

[68] I. Santos, J. J. Murillo-Fuentes, R. Boloix-Tortosa, E. A. de Reyna, and P. M. Olmos, "Expectation Propagation as Turbo Equalizer in ISI Channels", *IEEE Transactions on Communications*, vol. -, no. -, pp. –, 2016.

[69] M. Senst and G. Ascheid, "How the Framework of Expectation Propagation Yields an Iterative IC-LMMSE MIMO Receiver", in *IEEE Global Communications Conference*, (Houston, USA), Dec. 2011, pp. 1–6.

[70] T. P. Minka, "A family of algorithms for approximate Bayesian Inference", PhD thesis, Massachusetts Institute of Technology, 2001.

[71] D. P. Bertsekas, *Nonlinear programming.* Athena scientific Belmont, 1999.

[72] T. P. Minka, “The EP energy function and minimization schemes”, Tech. Rep., 2001.

[73] K. T. Truong and R. W. Heath, “Effects of Channel Aging in Massive MIMO Systems”, *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338–351, Dec. 2013.

[74] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, “Multiple-antenna techniques for wireless communications-a comprehensive literature survey”, *IEEE Communications Surveys & Tutorials*, vol. 11, no. 2, pp. 87–105, Feb. 2009.

[75] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error-correcting coding and decoding: turbo-codes”, in *IEEE International Conference on Communications*, (Geneva, Switzerland), May 1993, pp. 1064 –1070.

[76] T. J. Richardson and R. Urbanke, *Modern coding theory.* Cambridge University Press, 2008.

[77] 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, Feb. 2001.

[78] T. L. Narasimhan and A. Chockalingam, “Channel Hardening-Exploiting Message Passing (CHEMP) Receiver in Large-Scale MIMO Systems”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 847–860, Oct. 2014.

[79] D. L. Donoho, A. Javanmard, and A. Montanari, “Information-Theoretically Optimal Compressed Sensing via Spatial Coupling and Approximate Message Passing”, *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7434–7464, Nov. 2013.

[80] Y. Jia, C. Andrieu, R. J. Piechocki, and M. Sandell, “Improving soft output quality of MIMO demodulation algorithm via importance sampling”, in *IEE International Conference on 3G Mobile Communication Technologies*, (London, England), Oct. 2004, pp. 388–391.

[81] B. Farhang-Boroujeny, H. Zhu, and Z. Shi, “Markov chain Monte Carlo algorithms for CDMA and MIMO communication systems”, *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1896–1909, May 2006.

[82] M. Hansen, B. Hassibi, A. G. Dimakis, and W. Xu, “Near-Optimal Detection in MIMO Systems Using Gibb, year = 2009, venue=Honolulu,USA”, in *IEEE Global Telecommunications Conference*, pp. 1–6.

[83] R. R. Chen, R. Peng, A. Ashikhmin, and B. Farhang-Boroujeny, “Approaching MIMO capacity using bitwise Markov Chain Monte Carlo detection”, *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 423–428, Feb. 2010.

[84] T. Datta, N. A. Kumar, A. Chockalingam, and B. S. Rajan, “A Novel Monte-Carlo-Sampling-Based Receiver for Large-Scale Uplink Multiuser MIMO Systems”, *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3019–3038, Sep. 2013.

[85] K. P. Murphy, *Machine learning: A probabilistic perspective.* MIT press, 2012.

[87] T. J. Richardson, M. A. Shokrollahi, and R. Urbanke, “Design of capacity approaching irregular low-density parity-check codes”, *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, Feb. 2001.

[88] D. J. Costello, Jr., L. Dolecek, T. Fuja, J. Kliewer, D. G. M. Mitchell, and R. Smarandache, “Spatially coupled sparse codes on graphs: theory and practice”, *IEEE Communications Magazine*, vol. 52, no. 7, pp. 168–176, Jul. 2014.

[89] S. Kudekar, T. Richardson, and R. Urbanke, “Spatially Coupled Ensembles Universally Achieve Capacity Under Belief Propagation”, *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 7761–7813, Dec. 2013.

[90] T. Heskes, “Stable fixed points of loopy belief propagation are minima of the Bethe free energy”, in *Advances in Neural Information Processing Systems*, (Vancouver, Canada), vol. 14, Dec. 2003, pp. 343–350.

[91] J. M. Mooij and H. J. Kappen, “Sufficient Conditions for Convergence of the Sum-Product Algorithm”, *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4422–4437, Dec. 2007.

[92] G. Elidan, I. McGraw, and D. Koller, “Residual belief propagation: informed scheduling for asynchronous message passing”, in *Conference on Uncertainty in Artificial Intelligence*, (Cambridge, USA), Jul. 2006, pp. 165–173.

[93] D. G. M. Mitchell, A. E. Pusane, M. Lentmaier, and D. J. Costello, Jr., “Exact Free Distance and Trapping Set Growth Rates for LDPC Convolutional Codes”, in *IEEE International Symposium on Information Theory*, (St. Petersburg, Russia), Aug. 2011.

[94] J. Thorpe, “Low-density parity-check (LDPC) codes constructed from protographs”, *IPN progress report*, vol. 42, no. 154, pp. 42–154, 2003.

[95] D. G. M. Mitchell, M. Lentmaier, and D. J. Costello, Jr., “Spatially Coupled LDPC Codes Constructed From Protographs”, *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4866–4889, Sep. 2015.

[96] M. Tuchler, R. Koetter, and A. C. Singer, “Turbo equalization: principles and new results”, *IEEE Transactions on Communications*, vol. 50, no. 5, pp. 754–767, May 2002.

[97] F. Perez-Cruz, M. R. D. Rodrigues, and S. Verdu, “MIMO Gaussian Channels With Arbitrary Inputs: Optimal Precoding and Power Allocation”, *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1070–1084, Mar. 2010.

[98] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, “Towards 1 Gbps/UE in Cellular Systems: Understanding Ultra-Dense Small Cell Deployments”, *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2078–2101, Apr. 2015.