Interference management in large-scale MIMO systems for 5G

Zahran Hajji

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Télécommunications*

Par

Zahran HAJJI

Gestion des interférences des systèmes large-scale MIMO pour la 5G

Thèse présentée et soutenue à IMT Atlantique BREST, le 17/12/2018

Unité de recherche : Lab-STICC

Thèse N° : 2018IMTA0109

Rapporteurs avant soutenance :

Marie-Laure BOUCHERET Professeur INP ENSEEIHT
Jean-Marc BROSSIER Professeur GIPSA-Lab

Composition du Jury :

Président : Ali MANSOUR Professeur ENSTA Bretagne Professeur IMT Atlantique

Examinateur : Raphaël VISOZ Ingénieur de recherche ORANGE LABS

Rapporteurs : Marie-Laure BOUCHERET Professeur INP ENSEEIHT
Jean-Marc BROSSIER Professeur GIPSA-Lab

Dir. de thèse : Karine AMIS Professeur IMT Atlantique
Abdeljalil AISSA EL BEY Professeur IMT Atlantique

Invité(s)
Carlos Fouazi BADER Maître de conférences Centrale Supélec
Abdel-Ouahab BOUDRAA Maître de conférences École Navale

Acknowledgments

I WOULD like to express my sincerest thanks and appreciation to all the members of the Signal and Communications department for welcoming me in your laboratory. I have met many people inside and outside the work sphere that made the Ph.D an enjoyable adventure.

I am deeply grateful to my supervisors, Prof. Karine AMIS and Prof. Abdeldjalil AISSA-EL-BEY, for their invaluable advice, kindness, encouragements, patience and support during these three years. Their profound scientific knowledge, invaluable insight and experience have had a great impact on the success of the thesis. There is no doubt in my mind that without their comments, criticisms and guidance, my Phd will not be accomplished. I am also indebted to them for giving me the opportunity to improve my research background and experience. I am very lucky to have had the opportunity to work with them. It was always a pleasure to share an unforgettable moments rich with new results.

Furthermore, I also wish to thank the committee members for reviewing my thesis and enhancing me with their valuable comments.

Last but not least, the whole acknowledgment is dedicated to my family and my girlfriend Fiona for their unlimited support, guidance and help. I am unable to count their graces and without them and their unselfish love, and kindness, and tenderness, and affection, I have not been come thus far and achieved my thesis.

Résumé des Travaux de Thèse

Introduction

Au cours de la dernière décennie, les communications sans fil et les services Internet se sont infiltrés dans la société et ont radicalement changé notre vie, dépassant toutes les attentes. Ils sont devenus un réel besoin pour beaucoup d'entre nous. En outre, la demande en communications sans fil fiables continue de croître rapidement lorsque les mobiles sans fil prenant en charge les communications vocales vers des services multimédia à haut débit sont déployés avec succès. Pour répondre à cette demande, depuis les années 1980, diverses technologies et normes innovantes ont été proposées pour faire évoluer les systèmes de communications.

Afin de desservir les utilisateurs mobiles dans une zone géographique donnée, les concepteurs ont proposé de partitionner cette zone en petites surfaces appelées cellules. Les utilisateurs qui se trouvent dans chaque cellule sont desservis par une station de base située au centre de la cellule. Les communications cellulaires sont bidirectionnelles. Liaison montante permet à la station de base de détecter les signaux envoyés par les utilisateurs. La liaison descendante permet à la station de base transmettre des signaux à ses utilisateurs.

L'objectif est maintenant de concevoir un réseau cellulaire efficace offrant des communications sans fil fiables sur les ressources spectrales attribuées et limité par une puissance de transmission maximale. À cette fin, les chercheurs ont défini les paramètres de performance à optimiser. Un paramètre important à prendre en compte est l'efficacité spectrale justifiée par la rareté et le coût élevé du spectre radio-magnétique.

De la 2G à la 3G, des canaux SISO à entrée unique et à sortie unique ont été considérés en combinaison avec des techniques d'accés multiples telles que TDMA, FDMA et CDMA. À partir de la 4G et au-delà, avec l'augmentation importante du nombre d'utilisateurs et l'utilisation de techniques d'accés multiples, le spectre radio-magnétique risque de saturer et de nouvelles technologies doivent être proposées.

Grâce à l'expérience du V-BLAST implémenté en temps réel dans le laboratoire Bell Labs, les systèmes sans fil multi-antennes ont retenu l'attention des chercheurs. Ces derniers ont démontré que les systèmes MIMO pouvaient améliorer l'efficacité spectrale en augmentant le nombre d'antennes sans augmenter la puissance ou la bande passante par rapport aux systèmes SISO. En raison de ces avantages, la technologie MIMO est implémentée dans la 4G. Cependant, elle est utilisée avec un nombre d'antennes ne dépassant pas dix dans la plupart des cas.

Des chercheurs ont récemment découvert que le potentiel des systèmes MIMO à atteindre une efficacité spectrale très élevée n'a pas été complètement exploité. En conséquence, une nouvelle version de MIMO appelée MIMO à grande échelle

est devenue un domaine de recherche dynamique. L'idée est d'utiliser un grand nombre, par exemple des dizaines voire des centaines d'antennes au niveau de la station de base. Cela peut conduire à des efficacités spectrales très élevées avec l'augmentation du débit de données sans modification de la bande passante. Ces avantages par rapport aux systèmes MIMO font des systèmes MIMO à grande échelle une technologie prometteuse pour la 5G afin de répondre à la demande croissante de communications sans fil dans les années à venir.

L'objectif de la thèse était de concevoir des algorithmes de détection adaptés au contexte des systèmes MIMO à grande échelle afin d'exploiter leur potentiel.

Le modèle du système et les algorithmes de l'état de l'art

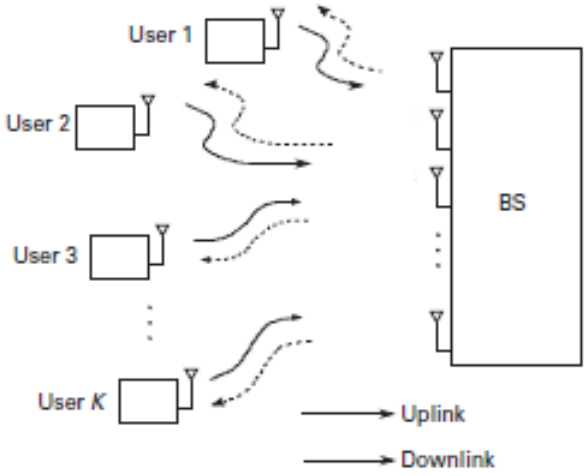

Figure 1: Communication MIMO multi-utilisateurs

Dans cette première partie, nous avons présenté le modèle mathématique des systèmes des communications MIMO à grande échelle dans le lien montant comme illustré sur la Figure. 1. Le modèle (1) repose sur l'expression d'un vecteur $\boldsymbol{y}$ dont les éléments sont les différentes observations au niveau des antennes à la station de base. Ce vecteur peut s'écrire sous la forme de la somme du vecteur représentant le bruit de l'environnement $\boldsymbol{\zeta}$ et d'une transformation du vecteur $\boldsymbol{x}$ contenant les différents symboles émis par les différents utilisateurs à l'aide de la matrice du canal $\boldsymbol{H}$ qui représentent l'état de canal entre les antennes à l'émission et à la réception.

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\zeta}, \tag{1}$$

Dans la première partie de la thèse, nous avons supposé que la matrice de canal est parfaitement connue et que les symboles émis appartiennent à une modulation d'alphabet fini et de constellation uniforme et carrée. Notre objectif est alors la détection des symboles émis tout en connaissant l'état de canal et le vecteur des observations au niveau de la station de base.

Dans notre état de l'art nous avons considéré deux groupes d'algorithmes de détection qui peuvent nous servir comme des références de la littérature et nous permettre de positionner nos algorithmes.

Une première classe est basée sur le critère de maximum de vraisemblance (MV) qui nécessite une recherche exhaustive. Cependant, dans les systèmes MIMO, en augmentant le nombre d'antennes, le critère MV devient un problème NP-hard qui ne peut pas être résolu. L'algorithme de décodage par sphère est une solution alternative. C'est un algorithme de décodage basé sur le critère MV qui peut avoir une complexité polynomiale pour certaines plages de valeurs de SNR. Lorsque les dimensions du système MIMO augmentent, la complexité de décodage de la sphère est nettement inférieure à la complexité exponentielle de la détection MV. L'idée est de limiter l'espace de recherche des points de réseau à une sphère centrée sur le point reçu. Cependant, pour les valeurs de SNR faibles et moyennes, sa complexité reste exponentielle, ce qui le rend inadapté lorsque les dimensions du système sont élevées.

La deuxième classe comprend les algorithmes de détection linéaires. Le filtrage adapté (MF), le forçage à zéro (ZF) et la minimisation de l'erreur quadratique moyenne (MMSE) sont des algorithmes de détection linéaire bien connus qui, contrairement au décodage par sphère, peuvent être appliqués lorsque les dimensions du système sont élevées en raison de leur complexité polynomiale. Mais ils obtiennent dans de tels cas une performance de détection médiocre par rapport aux algorithmes optimaux. Pour remédier à cet inconvénient, des schémas d'annulation d'interférence successive non linéaires basés sur des détecteurs linéaires ont été proposés, tels que ZF-SIC et MMSE-SIC. Leur complexité reste polynomiale et les performances de détection sont améliorées par rapport aux structures linéaires d'origine. La détection linéaire peut également être améliorée en utilisant les techniques de réduction de réseau de points.

Pour toute technique de détection proposée, deux critères d'intérêt sont la complexité et la performance de détection. L'objectif est donc d'obtenir un algorithme peu complexe et performant. Ceci est difficile à réaliser dans la plupart des cas. À titre d'exemple, le décodeur par sphère réalise les performances MV avec une complexité élevée contrairement aux détecteurs linéaires qui présentent une complexité faible et des performances faibles. Les chercheurs sont souvent obligés de gérer ce compromis qui doit encore être amélioré, en particulier dans le cas des systèmes MIMO à grande échelle. Un phénomène intéressant qui apparaît quand les dimensions du système deviennent suffisantes peut être exploité, il s'agit du durcissement de canal. Il est montré que lorsqu'un système MIMO surdéterminé est pris en compte (le nombre d'antennes de réception est beaucoup plus élevé que le nombre d'antennes d'émission), les détecteurs linéaires tels que MF, ZF et MMSE tendent vers des

performances optimales grâce au phénomène de durcissement de canal. Cependant, dans ce cas, l'efficacité spectrale est limitée par le nombre d'antennes d'émission qui doit être faible. Par conséquent, nous perdons l'un des avantages les plus importants des grands systèmes MIMO. Pour obtenir une efficacité spectrale élevée, les deux dimensions du système MIMO doivent être grandes, ce qui mène à une dégradation des performances à cause des interférences entre les différents utilisateurs. Pour résoudre ce problème, des algorithmes basés sur la recherche locale conviennent parfaitement aux systèmes MIMO à grande échelle, tels que LAS et RTS, qui atteignent des performances quasi-optimales tout en conservant la même complexité que la détection linéaire. L'idée ici est d'obtenir une première solution fournie par la détection linéaire et de l'améliorer en recherchant un meilleur minimum local choisi parmi les voisins de la sortie de la détection linéaire.

Détection basée sur l'acquisition comprimée

Dans la première partie de nos contributions, nous avons abordé le problème de la détection dans les systèmes MIMO à grande échelle. Nous avons proposé d'abord un critère de détection pour les systèmes de mélange sans bruit de signaux d'alphabet fini en exploitant leur simplicité. Un vecteur est simple lorsqu'il contient des élements égaux à ses bornes. Le terme simplicité a été introduit pour la première fois par Donoho. Il a démontré que des systèmes de mélange sous-déterminés où le nombre de sources dépasse le nombre d'observations peuvent être détectés avec succès. Nous avons montré que cette propriété peut être exploitée de la même façon que la propriété de parcimonie de tels signaux afin de proposer de nouveaux algorithmes basés sur la technique de l'acquisition comprimée. Nous avons ensuite étendu le critère proposé, basé sur la simplicité, au cas bruité, afin de concevoir un détecteur de faible complexité pour les systèmes MIMO à grande échelle. À notre connaissance, il s'agit du premier algorithme proposé capable de se comporter efficacement dans des configurations de systèmes MIMO à grande échelle, déterminés et sous-déterminés. On peut s'attendre à une telle configuration dans les communications montantes car le nombre d'utilisateurs multiplié par le nombre de leurs antennes pourrait être beaucoup plus élevé que le nombre d'antennes à la station de base.

Ensuite, nous avons montré l'efficacité du critère proposé appelé (FAS) par rapport aux algorithmes de détection les plus efficaces en analysant théoriquement les conditions de succés de détection, la probabilité dâerreur de détection et la complexité de calcul.

En premier lieu, nous avons examiné la condition nécessaire d'unicité et d'existence de la solution du critère proposé. Cette condition couvre le cas déterminé et partiellement le cas sous-déterminé. Par rapport aux techniques précédentes basées sur la parcimonie, nous avons obtenu une réduction suffisante des coûts de

calcul avec une préservation du taux d'erreur. Les résultats de la simulation ont corroboré avec l'analyse théorique. Nous avons ensuite comparé le critère proposé aux algorithmes de l'état de l'art. Sur la Figure. 2, on remarque que l'algorithme FAS est meilleur que les algorithmes de détection linéaire ainsi que leurs versions améliorées. Sur la Figure 3, on compare l'algorithme proposé avec les algorithmes de recherche locale (LAS et RTS). Le critère proposé (FAS) surpasse l'algorithme LAS dans tous les cas étudiés et dépasse le RTS au-dessous d'un seuil de BER qui augmente en augmentant l'ordre de modulation. La distribution théorique de la sortie du détecteur a aussi été demontrée et validée par des simulations. Dans la partie suivante, les résultats analytiques seront utilisés premièrement pour définir un récepteur itératif basé sur le concept de zone d'ombre afin d'améliorer encore les performances dans le cas non codé et deuxièmement pour définir un récepteur itératif de type turbo prenant en compte un décodeur FEC.

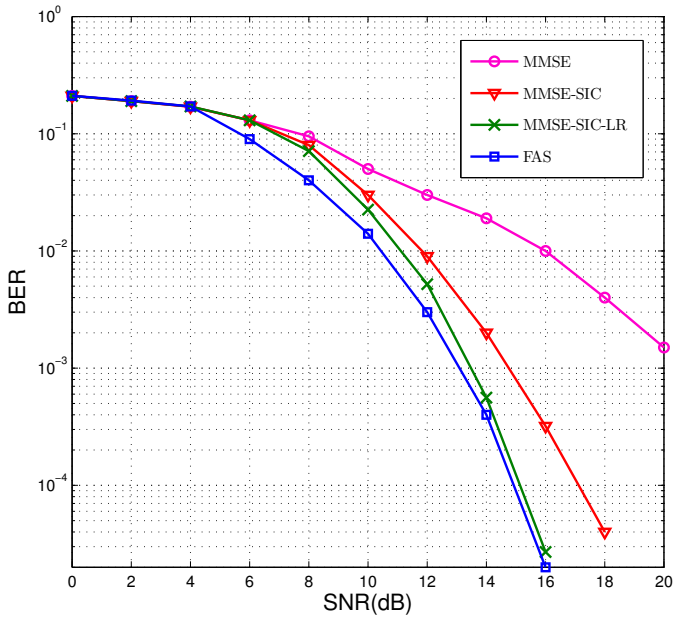

Figure 2: Comparaison de taux d'erreur binaire pour un système 64×64 avec une modulation 4-QAM.

Algorithmes itératifs basés sur le FAS

Dans la deuxième partie de nos contributions, nous avons de nouveau considéré le problème de la détection dans les systèmes MIMO à grande échelle et nous avons défini des récepteurs itératifs qui utilisent l'algorithme de détection basé sur la sim-

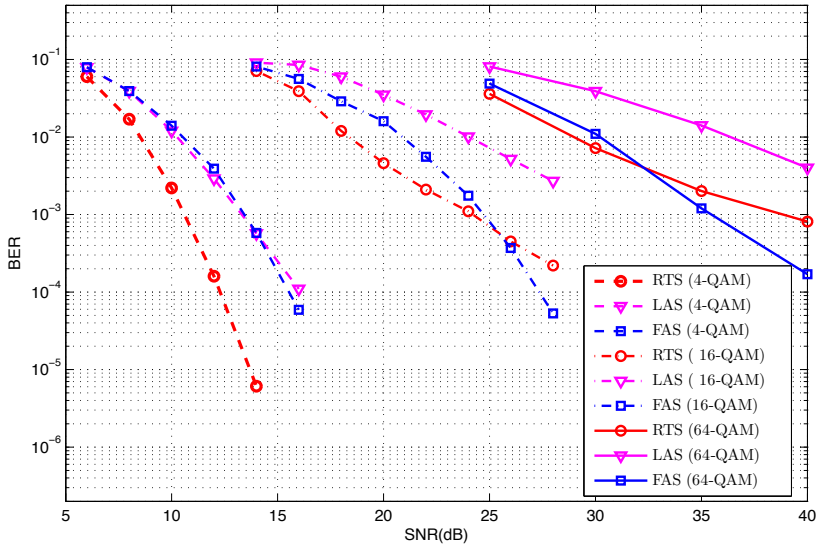

Figure 3: Comparaison de taux d'erreur binaire pour un système 32×32 avec 4-QAM,16-QAM and 64-QAM.

plicité (FAS). Dans un premièr lieu, nous avons travaillé sur les systèmes MIMO à grande échelle non codés. Nous avons proposé un nouvel algorithme d'annulation d'interférence successif appelé (FAS-SAC), basé sur le principe de la zone d'ombre et nous avons optimisé ses paramètres en exploitant l'analyse théorique de la sortie du détecteur. En deuxième lieu, nous avons considéré les systèmes MIMO à grande échelle codés et notre objectif était la définition des récepteurs itératifs de type turbo. Nous avons proposé un récepteur itératif, appelé (FAS-ML), basé sur une détection de type ML dont les sous-ensembles de candidats sont définis comme le voisinage de la sortie du critère FAS. Pour réduire davantage la complexité du récepteur, nous avons également introduit un autre récepteur basé sur un algorithme FAS dont le critère est pénalisé par une fonction de la valeur absolue de lâerreur moyenne pondérée par les sorties fiables du décodeur. Le paramètre de régularisation été défini analytiquement. Nous avons appelé ce récepteur (FAS-MAE). Les résultats de la simulation sur la Figure 4 ont montré que l'algorithme FAS-SAC proposé surpasse de manière significative les algorithmes standard FAS, LAS et RTS dans presque tous les cas, avec le même ordre de complexité de calcul. Dans le cas codé, sur la Figure 5, nous avons montré que les deux récepteurs itératifs proposés sont meilleurs que le Turbo-MMSE dans toutes les configurations déterminées et sous-déterminées et que FAS-MAE donne de meilleurs résultats que FAS-ML pour les modulations de taille élevée. Jusqu'à présent, tout le travail effectué repose sur une connaissance parfaite de la matrice de canal. Néanmoins, dans le cas de systèmes MIMO à grande échelle, ce scénario est difficile à obtenir. La partie suivante est consacrée à l'évaluation d'impact de l'estimation de canal sur les performances

du récepteur et à la conception d'algorithmes efficaces traitant de l'estimation de canal.

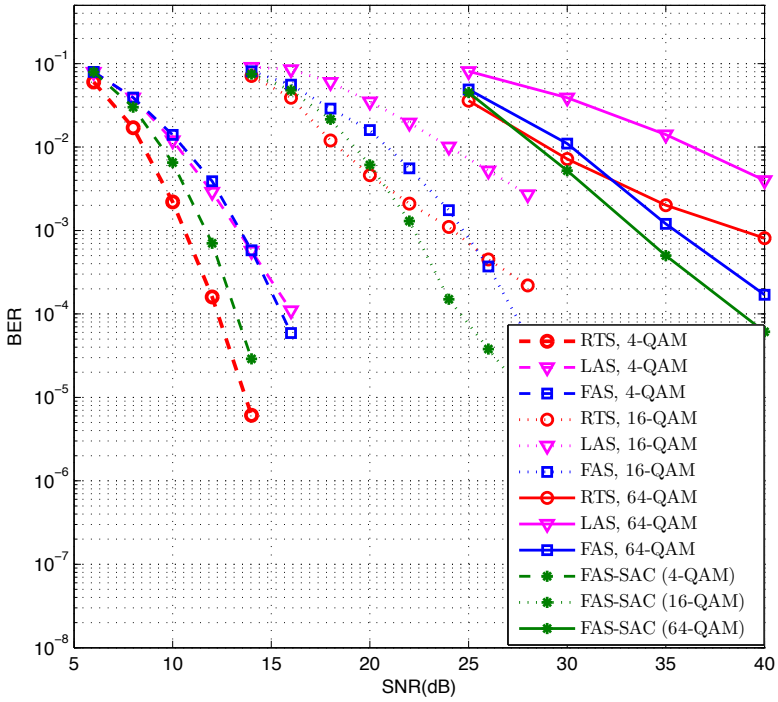

Figure 4: Comparaison de taux d'erreur binaire pour un système 32×32 avec 4-QAM,16-QAM and 64-QAM.

Estimation de canal dans les systèmes MIMO à grande échelle

Dans les premières parties, les algorithmes de détection étaient présentés sous l'hypothèse d'une connaissance parfaite du canal au niveau du récepteur. Toutefois, dans la pratique, les gains de canal sont estimés au niveau du récepteur, soit à l'aveugle, soit en semi-aveugle, soit en utilisant des séquences pilotes seulement. En raison du bruit et du nombre fini des séquences pilotes, les estimations de canal ne sont pas parfaites. Cela a une influence sur la capacité du canal MIMO et sur les performances des algorithmes de détection. Dans cette dernière partie, nous avons traité tout d'abord l'effet de l'estimation imparfaite de canal sur la performance du système MIMO. Ensuite, nous avons proposé des algorithmes d'estimation de canal

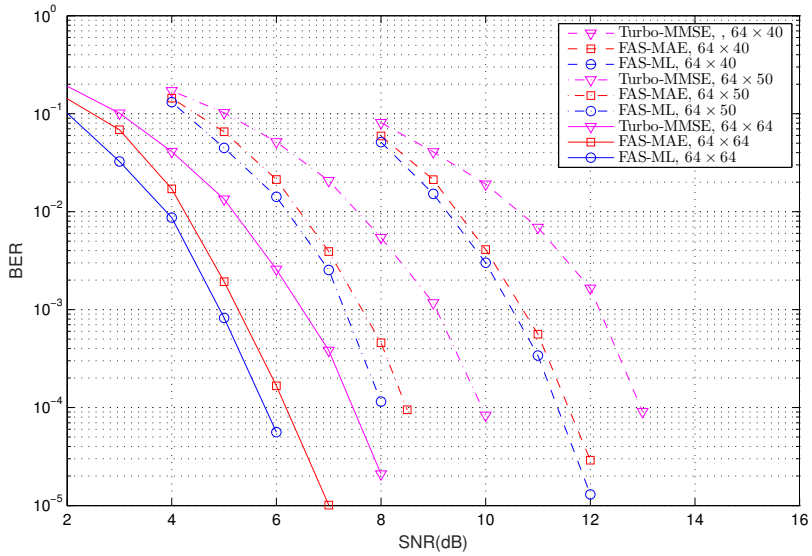

Figure 5: Comparaison de taux d'erreur binaire pour un système 32×32 des algorithmes FAS-MAE, FAS-ML and Turbo-MMSE avec une modulation 4-QAM codé.

semi-aveugles dans les systèmes MIMO à grande échelle non codés et codés avec des alphabets finis en supposant une longueur de séquence pilote limitée. Nous avons proposé des schémas d'estimation de canal fondés sur des décisions souples et dures des algorithmes FAS et FAS-SAC. Nous avons montré qu'il suffisait de prendre en compte un nombre de séquences pilotes égal au nombre d'utilisateurs. Des études théoriques pour les deux algorithmes ont été menées et nous avons déterminé les bornes de Crame Rao (CRB) lorsque des décisions souples sont prises en compte et les erreurs quadratiques moyennes (EQM) asymptotiques lorsque des décisions dures sont utilisées. Les résultats de simulation ont montré la validité de l'étude théorique. Ensuite, nous avons proposé un récepteur basé sur le turbo FAS-MAE qui combine l'estimation, la détection et le décodage FEC et nous avons défini deux manières de mettre à jour l'estimation de canal basée sur FAS à partir de la sortie du décodeur FEC. Les résultats de simulations sur les Figures 6, 7 et 8 ont montré l'efficacité du schéma proposé, qui fonctionne presque à la limite inférieure des performances avec estimation du canal MV basée sur une connaissance de toute la séquence de symboles émis, avec une supériorité des schémas basées sur celles basées sur les décisions dures dans le cas non codé, est des schémas basées sur celles basées sur les décisions dures dans le cas codé.

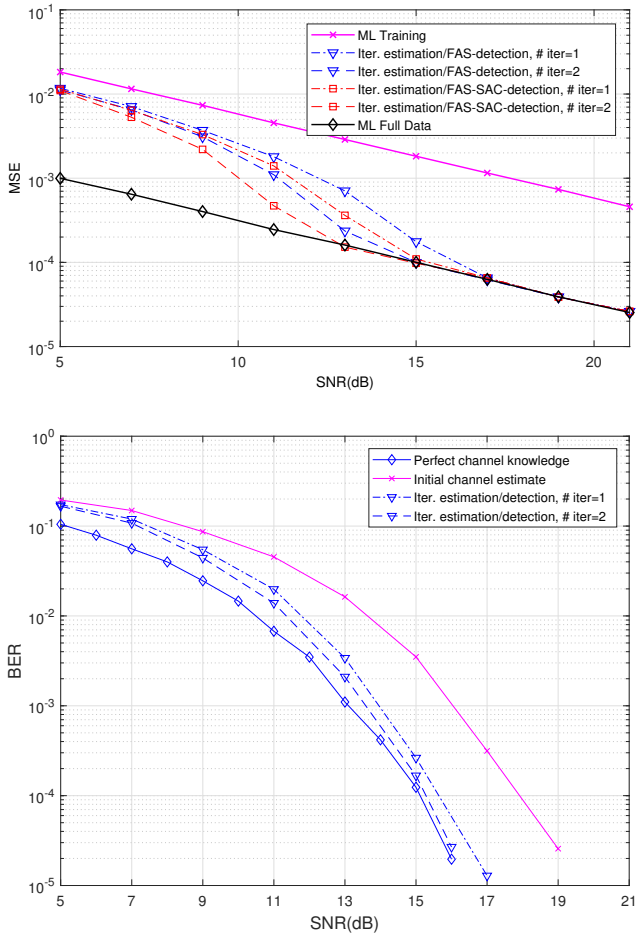

Figure 6: EQM vs RSB (Décision dure (FAS)), TEB vs RSB (Décision dure (FAS)) avec une modulation 4-QAM non codée, $n = N = 64$, $T_p = 64$ and $T = 1280$.

Conclusions

Cette thèse était motivée par les nouvelles opportunités offertes par les systèmes MIMO à grande échelle et par les différents défis à relever pour les rendre opérationnels. Par exemple, les algorithmes connus pour bien fonctionner avec un petit nombre d'antennes ne peuvent pas supporter le passage aux grandes dimensions. Par conséquent, de nouvelles approches et alternatives sont nécessaires.

Les travaux de la thèse peuvent être résumés comme suit :

Dans la première partie, nous avons d'abord présenté les systèmes MIMO et

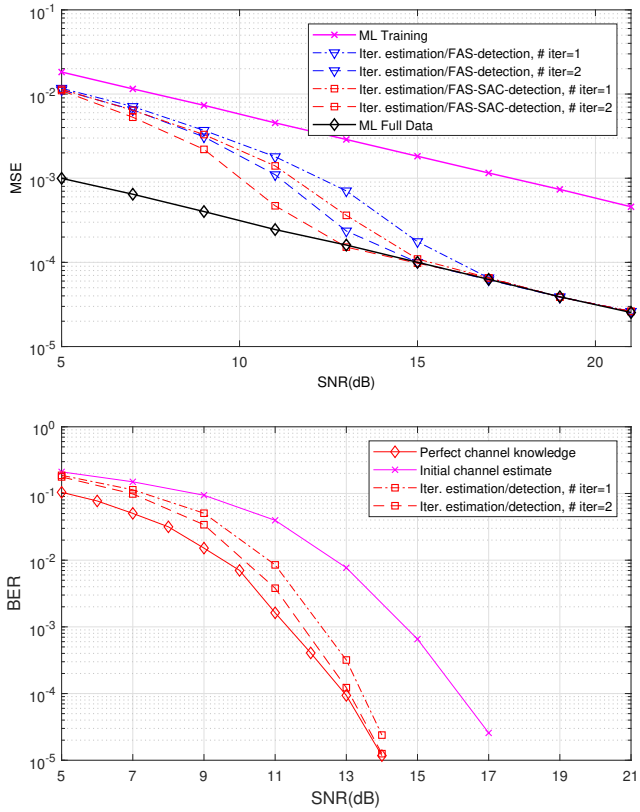

Figure 7: EQM vs RSB (Décision dure (FAS-SAC)), TEB vs RSB (Décision dure (FAS-SAC)) avec une modulation 4-QAM non codée, $n = N = 64$, $T_p = 64$ and $T = 1280$.

nous avons montré les avantages de tels systèmes par rapport aux systèmes SISO. Différents paramètres de performance tels que l'efficacité spectrale et la probabilité d'erreur ont été présentés. Nous avons ensuite présenté certains défis lorsque de tels systèmes sont déployés, tels que la nécessité d'algorithmes de faible complexité pour le traitement des systèmes d'antennes à grande dimension. L'estimation de canal a également été évoquée.

Ensuite, nous avons présenté l'état de l'art sur les techniques et les algorithmes de détection dans les systèmes MIMO et les systèmes MIMO à grande échelle. Nous avons détaillé les différents algorithmes et nous avons montré leurs points forts et leurs points faibles.

Dans la deuxième partie consacrée à la détection dans les systèmes MIMO à grande échelle, nous avons examiné les techniques de détection basées sur l'acquisition comprimée. Nous avons examiné les signaux à alphabet fini et nous avons exploité leur propriété de simplicité pour proposer des algorithmes efficaces dans les cas bruité et non bruité. Le cas sans bruit a été théoriquement étudié pour obtenir les conditions nécessaires de la bonne détection. Le schéma de récupération

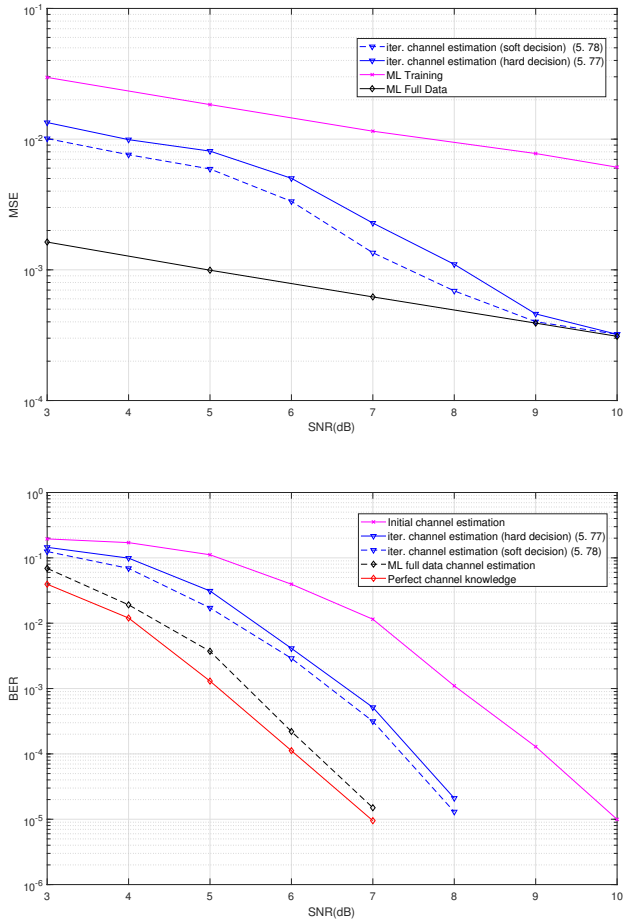

Figure 8: EQM vs RSB (FAS-MAE), TEB vs RSB (FAS-MAE) avec une modulation 4-QAM codée,, $n = N = 64$, $T_p = 64$ and $T = 1280$.

proposé a ensuite été étendu aux systèmes MIMO à grande échelle et une étude théorique des statistiques de sortie de détection a été réalisée avec succès. Nous avons montré que l'algorithme proposé présente le même ordre de complexité que les algorithmes peu complexes de l'état de lâart (MMSE, LAS, RTS) et atteint des meilleures performances (gains importants par rapport au LAS et au MMSE).

Ensuite, notre objectif était d'intégrer l'algorithme proposé dans une procédure itérative pour les cas non codés et codés afin d'améliorer ses performances. Au premier abord, nous avons examiné le cas non codé et nous avons proposé un schéma d'annulation successive d'interférences basé sur le principe de la zone d'ombre, avec des paramètres théoriquement fixés sur la base des statistiques de la sortie du critère proposé à l'origine. Nous avons ensuite proposé d'intégrer le schéma original dans un récepteur itératif de type turbo.

Dans la dernière partie, nous avons examiné le cas d'une estimation imparfaite de canal et nous avons introduit des algorithmes itératifs d'estimation semi-aveugles

basés sur le critère des moindres carrés alimentés par les sorties des algorithmes de détection proposés. Le cas codé a également été traité et des schémas d'estimation de canal et de détection itératifs de type turbo ont été proposés.

Abbreviations

MIMO multiple-input multiple-output

2G second generation

3G third generation

4G fourth generation

5G fifth generation

SISO single-input single-output

TDMA time division multiple access

FDMA frequency division multiple access

CDMA code division multiple access

OFDMA orthogonal frequency-division multiple access

OFDM orthogonal frequency-division multiplexing

V-BLAST vertical-Bell labs layered space-time

BS base station

CSI channel state information

CSIT channel state information at the transmitter

DPC dirty paper coding

SNR signal-to-noise ratio

QAM quadrature amplitude modulation

PAM pulse amplitude modulation

FEC forward error correction

ML maximum likelihood

MF matched filter

ZF zero forcing

MMSE minimum mean square error

SIC successive interference cancellation

IC interference cancellation

LAS likelihood ascent search

RTS reactive tabu search

AWGN additive white gaussian noise

SD sphere decoding

SINR Signal-to-interference-plus-noise ratio

LR lattice reduction

SA Seysen's algorithm

NP non-deterministic polynomial

CS compressive sensing

BP basis pursuit

FAS finite alphabet simplicity

SER symbol error rate

BER bit error rate

SAC shadow area constraints

LLR log likelihood ratio

MAE mean absolute error

SSE soft symbol error

TDD time-division duplexing

FDD frequency-division duplexing

CRB Cramer-Rao bound

EM expectation-maximization

MSE mean square error

PB pilot block

DB data block

SBC symbol-to-binary converter

BSC binary-to-symbol converter

ISI Inter-symbol interference

ZP zero padding

CP cyclic prefix

Notations

$(.)^H$	Hermitian transposition
$(.)^T$	Transposition
$(.)^*$	Complex conjugation
$\lvert.\rvert$	Absolute value of a number or a cardinality of a set
$\lVert.\rVert$	Euclidean norm of a vector
$\lVert.\rVert_p$	ℓ_p norm of a vector
$\lfloor a \rfloor$	Largest integer less than a
$\lfloor . \rceil$	Rounding operation to the nearest integer
$\mathbb{E}[.]$	Expectation operation
$\mathrm{Re}(.)$	Real part of the complex argument
$\mathrm{Im}(.)$	Imaginary part of the complex argument
$\mathrm{sgn}(.)$	Sign
$\Pr(.)$	Probability
$\boldsymbol{x}$	Complex vector $\boldsymbol{x}$
x_i	i-th element of the vector $\boldsymbol{x}$
$\underline{\boldsymbol{x}}$	Real transformation of the complex vector $\boldsymbol{x}$ defined by $\underline{\boldsymbol{x}} = \begin{pmatrix} \mathrm{Re}(\boldsymbol{x}) & \mathrm{Im}(\boldsymbol{x}) \end{pmatrix}$
$\boldsymbol{H}$	Complex matrix $\boldsymbol{H}$
$\underline{\boldsymbol{H}}$	Real transformation of the complex vector $\boldsymbol{H}$ defined by $\underline{\boldsymbol{H}} = \begin{pmatrix} \mathrm{Re}(\boldsymbol{H}) & -\mathrm{Im}(\boldsymbol{H}) \\ \mathrm{Im}(\boldsymbol{H}) & \mathrm{Re}(\boldsymbol{H}) \end{pmatrix}$
$\mathrm{tr}(\boldsymbol{H})$	Trace of the matrix $\boldsymbol{H}$
$\boldsymbol{h}_i$	The i-th column of the matrix $\boldsymbol{H}$
$\boldsymbol{H}_\Lambda$	The concatenation matrix of the columns of the matrix $\boldsymbol{H}$ with indices in the set Λ
$\otimes$	Kronecker product
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\mathbf{1}_n$	length-n ones vector
$\mathbb{C}$	Field of complex numbers
$\mathbb{R}$	Field of real numbers
n	Number of receive antennas
N	Number of transmit antennas

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Contents

1.1 Introduction

During last decade, wireless communications and internet services have infiltrated the society and changed our life seriously exceeding all expectations. They became a real need for many of us. In addition, the demand of reliable wireless communications is still growing rapidly when wireless mobiles that support voice communications to high data rate multimedia services are successfully deployed. To meet this demand, since 1980s, various and novel technologies and standards have been proposed as demonstrated in Fig. 1.1 from the first generation to the fifth generation communications systems.

In order to serve mobile users in a given geographic area, the designers proposed to partition this area into small surfaces called cells. Then, the users who are in each cell will be served by a base station located in the center of the cell as demonstrated in Fig. 1.2. The wireless communications are two-way. A first one is the uplink, when base station detects its assigned users signals. The other way is the downlink when the base station transmits signals to its users.

The objective now is to design an efficient cellular network that provides reliable wireless communications over the allocated spectrum resources and constrained by a maximum transmission power. To that purpose, researchers defined performance parameters to be optimized. An important parameter that should be taken into

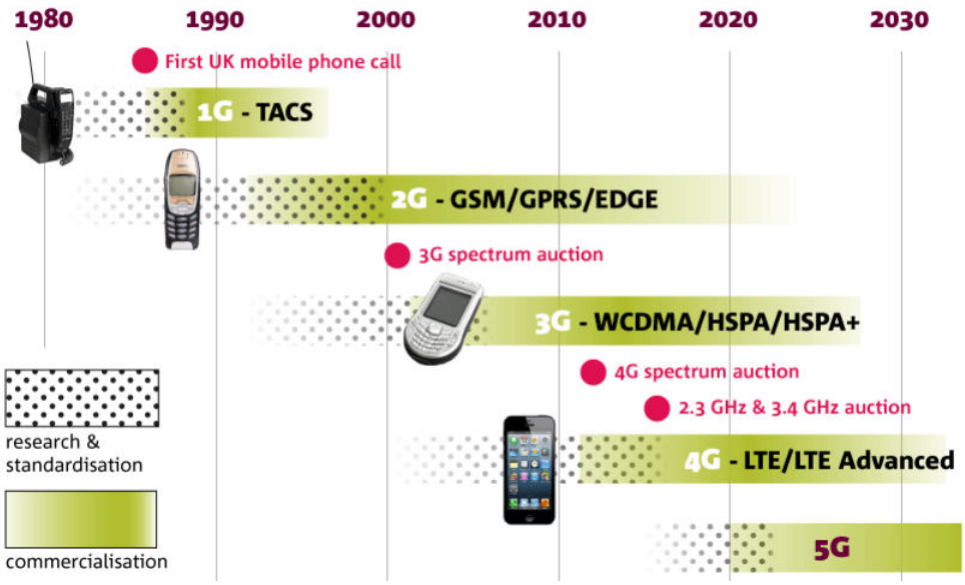

Figure 1.1: Wireless communications evolution

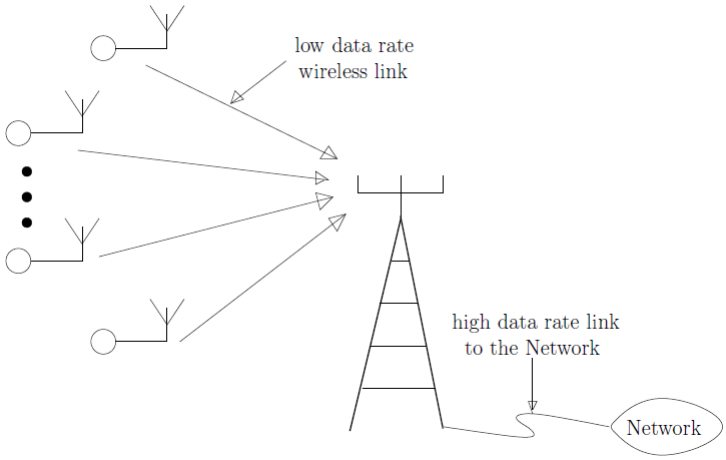

Figure 1.2: Cellular communication

account is the spectral efficiency justified by the scarcity and then the high cost of the radio spectrum.

From 2G to 3G, single-input single-output SISO channels were considered in combination with multiple access techniques such as TDMA, FDMA and CDMA. From 4G and beyond, as the number of users highly increases, using previous multiple access techniques, the radio spectrum risks to be saturated and new technologies must be proposed.

Motivated by the experience of the vertical BLAST (Bell Laboratories Layered Space-Time) or V-BLAST system implemented in real-time in the Bell Labs laboratory [1], multiantenna wireless systems have regained the attention of researchers.

One of the first theoretical study of these systems referred to as multiple-input multiple output (MIMO) systems was published in [2]. The authors demonstrated that the MIMO system can lead to better spectral efficiency increasing the number of antennas without extra power nor bandwidth compared to the SISO systems. Due to these advantages the MIMO technology is implemented in the 4G. However, it is used with a number of antennas that does not exceed ten in most cases.

Recently, researchers in MIMO field discover that the full potential of MIMO systems to achieve very high spectrum efficiency has not been yet exploited in practice. Consequently a new version of MIMO called large-scale MIMO has become an attractive field of research. The idea is to use large number such as tens to hundreds of antennas at the base station (BS). This can lead to achieve very high spectral efficiencies when increasing the data rate without changing the bandwidth. These advantages over MIMO systems make large MIMO systems a promising technology for 5G to meet the growing demand of wireless communications in the coming years.

In this chapter, we first present the encoded MIMO system communication and we detail the performance achieved by the MIMO communication compared to SISO communication. Then we show the motivation to migrate to large-scale systems.

1.2 MIMO communication

1.2.1 Point-to-point MIMO communication

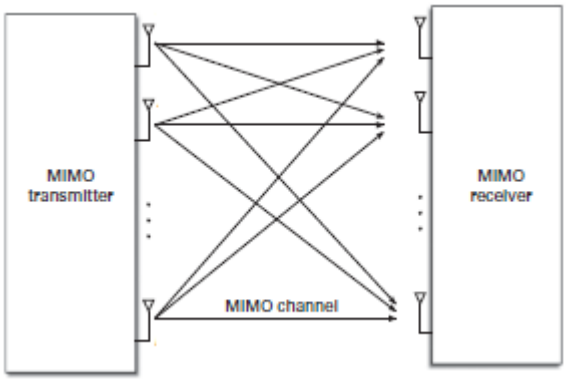

Figure 1.3: Point-to-point MIMO system

1.2.1.1 MIMO system model

In the point-to-point MIMO systems (Fig. 1.3), the encoded system model can be described by:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}_c + \boldsymbol{\zeta}, \tag{1.1}$$

where $\boldsymbol{y}$ is the received vector, $\boldsymbol{x}_c$ is the transmitted vector, $\boldsymbol{\zeta}$ is the additive white Gaussian noise and $\mathbb{E}[\boldsymbol{\zeta}\boldsymbol{\zeta}^H] = \sigma^2 I_n$ with σ^2 the noise variance at each receiver

antenna. $\boldsymbol{H}$ is the $(N \times n)$ channel matrix with entries assumed to stay constant over the signaling interval and modeled as independent and identically distributed (i.i.d). The matrix $\boldsymbol{H}$ can be estimated using pilot symbol vectors known at the receiver during a training phase. To approach the limit performance, the transmitted vector $\boldsymbol{x}_c$ is encoded. Then the information bits are grouped into messages that correspond to vectors belonging to a codebook $\mathcal{X}$. Each message contains $R = \log_2|\mathcal{X}|$ bits where R is defined as the transmission rate.

To study the utility of considering the MIMO system, the capacity of the transmission channel must be determined. The MIMO capacity for a codeword $\boldsymbol{x}_c$ is given by:

$$C_{MIMO}(SNR, \boldsymbol{H}) = \log_2 \det \left(\boldsymbol{I}_N + \frac{1}{\sigma^2} \boldsymbol{H} \mathbb{E}[\boldsymbol{x}_c \boldsymbol{x}_c^H] \boldsymbol{H}^H \right) \tag{1.2}$$

where $\mathbb{E}[\boldsymbol{x}_c \boldsymbol{x}_c^H]$ is the transmit covariance matrix with $tr(\mathbb{E}[\boldsymbol{x}_c \boldsymbol{x}_c^H]) = P_t$, P_t is the average power of the transmitted vector. The SNR is then defined as the ratio $\frac{P_t}{\sigma^2}$. Given the MIMO capacity, to get a low error probability detection of the codeword $\boldsymbol{x}_c$, the transmission rate R should be inferior to $C_{MIMO}(SNR, \boldsymbol{H})$. Without the knowledge of the channel state information (CSI) at the transmitter, for a given SNR value, the transmission rate is fixed arbitrary. The transmitter uses a fixed codebook that does not change when changing the channel gains and the transmitter codebook selection depends highly on whether the channel is slow or fast fading.

With slow fading, the channel may remain approximately constant long enough to allow reliable estimation of CSI at the receiver. The channel does not change during the length of the codeword. If the transmission rate R is chosen such that $C_{MIMO}(SNR, \boldsymbol{H}) < R$, it is impossible to successfully recover the transmitted vector.

In order to design a performing transmission-reception scheme, it is proposed to evaluate the probability of that scenario which is called the outage probability. Its theoretical limit for large-length codewords is defined as:

$$P_{outage}(SNR, R) = \min_{\boldsymbol{x} | tr(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H]) = P_t} \Pr(C_{MIMO}(SNR, \boldsymbol{H}) < R) \tag{1.3}$$

With a transmission rate R satisfying $C_{MIMO}(SNR, \boldsymbol{H}) < R$, any encoding-decoding scheme would have a codeword error rate higher than the channel outage probability given in 1.3. Therefore, researchers look for designing encoding-decoding schemes that can perform close to the channel outage probability for all values of SNR and R.

For the slow fading MIMO channels, the diversity gain which is a reliability measure and the multiplexing gain which measures the degrees of freedom are important to design efficient coding schemes. They are related by the so-called diversity-multiplexing gain tradeoff [3]. Their maximum values are about nN and $\min(n, N)$ respectively. For example, the maximum diversity order of V-BLAST algorithm [1] is only n. This is because the transmitted symbols are independent and they reach the receiver through n different paths. Another performing scheme, which achieves

the maximum diversity gain is the Alamouti space-time block code [4] thanks to the coding across both space and time.

In a fast fading channel, the channel changes many times over the duration of the used codeword. The codeword reception can be more reliable by spreading the codeword over different fades. In such scenario, an error-free communication can be achieved when the transmission rate R is below the ergodic MIMO capacity defined as follows:

$$C_{ergodic}(SNR) = \max_{\boldsymbol{x}|tr(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H])=P_t} \mathbb{E}_{\boldsymbol{H}}\left[C_{MIMO}(SNR, \boldsymbol{H})\right] \tag{1.4}$$

Note that the ergodic capacity is achieved when $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] = (P_t/N)\boldsymbol{I}_N$ and is therefore given by:

$$C_{ergodic}(SNR) = \mathbb{E}_{\boldsymbol{H}}\left[\log_2 \det\left(\boldsymbol{I}_n + \frac{SNR}{N}\boldsymbol{H}\boldsymbol{H}^H\right)\right] \tag{1.5}$$

It can be approximated as follows:

$$C_{ergodic}(SNR) \approx \min(n, N)\log_2(1 + SNR) \tag{1.6}$$

The ergodic MIMO capacity linearly increases when increasing $n = N$ [5]. Note that when the channel state information (CSI) is known at the transmitter side, the ergodic capacity of slow fading MIMO channel can be achieved with independent Gaussian beamformed signals where the linear precoding depends on singular vectors of the channel. In such a case, the MIMO channel can be transformed into $min(N, n)$ parallel non-interfering subchannels. Then, applying the waterfilling power allocation over these subchannels the ergodic capacity is achieved [5].

In addition, when the channel state information is available at the transmitter, the channel capacity can be increased by allocating more power to the most reliable channels with higher channel gain.

1.2.2 Multiuser MIMO communication

The multiuser MIMO communication refers to the communication between the base station and the user terminals in cellular communication. This is illustrated in Fig. 1.4. The uplink and the downlink are a multipoint-to-point communication and a point-to-multipoint communication respectively. In order to serve different user terminals, different multiple access techniques are used like TDMA, FDMA, CDMA and orthogonal frequency multiple access (OFDMA). In a broadcast channel (equivalent to the downlink of a cellular communication), the base station sends information to all users simultaneously and then each user extracts its private message. The dirty paper coding (DPC) [6] can achieve the maximum capacity of the broadcast channel. The sum capacity which is the sum of all data rates achieved by the users linearly increases with the number of transmit antennas at the base station and the

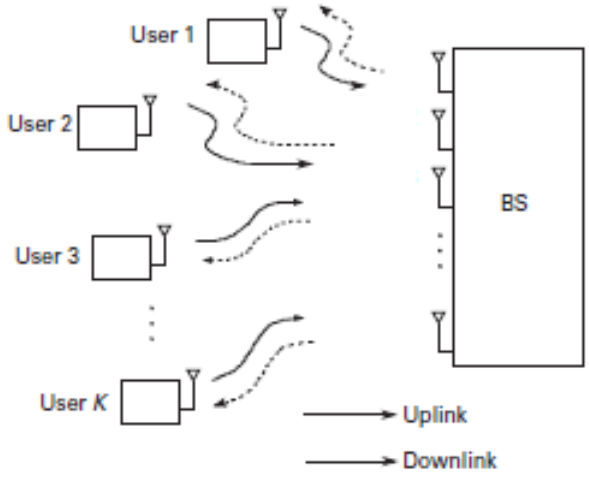

Figure 1.4: Multiuser MIMO system

	System dimension	Error probability	Channel capacity
SISO (non-fading)	$n = N = 1$	$\exp(-SNR)$	$\log_2(1 + SNR)$
MIMO (non-fading)	$n > 1, N > 1$	$(SNR)^{-1}$	$\log_2(1 + SNR)$
MIMO (fading)	$n > 1, N > 1$	$(SNR)^{-nN}$	$\min(n, N)\log_2(1 + SNR)$

Table 1.1: Performance indicators for SISO and MIMO systems.

number of single antenna user terminals. When a large number of transmit antennas is considered as significantly exceeding the number of user terminals, simple encoding-decoding techniques can achieve high performance thanks to additional spatial diversity.

1.2.3 Advantages of MIMO over SISO communication

In Table 1.1, a brief comparison of the link reliability in terms of error probability and channel capacity is given for both SISO and MIMO cases. In SISO non-fading channel, it is shown that the channel increases logarithmically contrary to the error probability which decreases exponentially with the SNR increases. However, when fading is considered the error probability decreases only linearly when increasing SNR. This performance degradation can be alleviated by considering MIMO systems exploiting the diversity of multiple channels and the error probability can be reduced exponentially when increasing the system dimensions. Compared to SISO systems, in addition to the gain in terms of link reliability, it is also shown that the channel capacity can be multiplied by the minimum of the system dimensions in MIMO systems. This makes MIMO systems more attractive in terms of both channel capacity and link reliability. The spectral efficiency can be increased either by increasing the size of the constellation alphabet (in SISO and MIMO systems) or

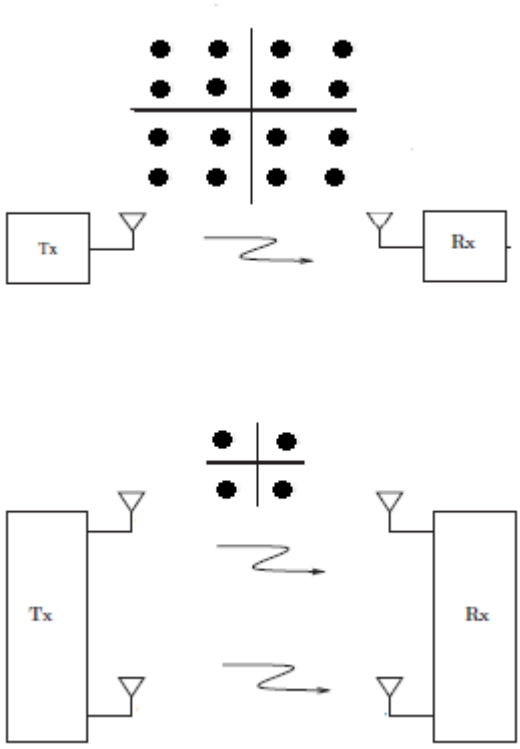

Figure 1.5: SISO and MIMO systems with the same spectral efficiency (4 bits per channel use)

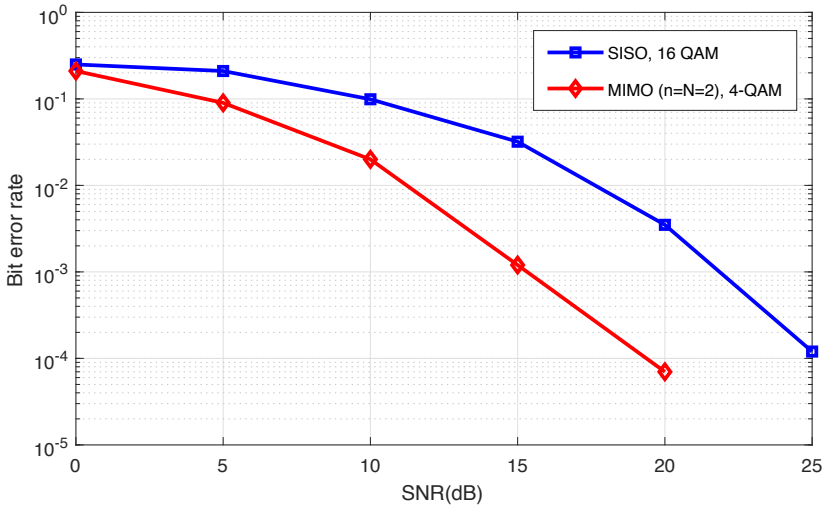

Figure 1.6: Bit error performance comparison of SISO and MIMO systems for the same spectral efficiency (4 bits per channel use)

the number of the transmitted antennas only in MIMO systems. For instance, given a frequency bandwidth to transmit 4 bits per channel use, a small modulation order (4-QAM) with two transmit antennas can be used in MIMO whereas a 16-QAM is required in SISO (see Fig.1.5). In Fig. 1.6 we show the bit error rate performance based on the zero forcing detection in both cases. The SISO system with 16-QAM achieves poor performance and the improvement when SNR increases is slow. This can be explained as mentioned in Table 1.1 by the linear dependency on the SNR value $P_{error} \propto SNR^{(-1)}$. However, when increasing the number of antennas, the MIMO system outperforms significantly the SISO systems thanks to spatial diversity. Therefore, increasing the number of antennas is a better choice to increase spectral efficiency.

To rise the data trafic explosion and to better exploit the detailed benefits of MIMO systems compared to SISO systems, researchers have proposed large-scale MIMO systems motivated by increasing the different communication performance indicators such as link reliability and spectral efficiency when increasing the MIMO system dimensions. In next Section, large-scale MIMO systems are introduced and their different benefits as well as raised challenges are presented.

1.3 Large-scale MIMO systems

1.3.1 Description

Large-scale MIMO systems involve a large number of antennas in both transmission and reception sides. This number goes from tens to hundreds. They include different MIMO configurations. The point-to-point MIMO configuration can be considered when the number of antennas is large. A typical real scenario for such configuration is the communication between two neighboring base stations possessing both a large number of antennas. The multiuser MIMO is also taken into account in large-scale MIMO systems. For the case of cellular communication, a multipoint-to-point uplink communication and point-to-multipoint downlink communication between the base station and the user terminals are established.

As user terminals are size-limited, large number antenna array is infeasible and the interference has to be managed at the base station either thanks to interference cancellation algorithms (uplink) or sophisticated beamforming (downlink). The higher the number of antennas at the base station the higher the degrees of freedom (space diversity).

1.3.2 Motivations

1.3.2.1 Full exploitation of MIMO advantages

Large-scale MIMO systems present numerous advantages. First, thanks to the additional degrees of freedom, they should support the expected increase of data traffic. Indeed according to Table 1.1 the capacity is proportional to the minimum between transmit and receive antenna number, and thus higher-data rate applications can

be afforded. Second as the maximum diversity gain equals the product of transmit and receive antennas, the transmission quality is ensured and services sensitive to errors should be better protected. Third, thanks to large number of antennas at the base station, the maximum number of users that can be served by the base station should increase and beamforming solutions with interference limited to the immediate neighboring of the user should be made possible which makes the resource allocation and scheduling less complex but requires accurate channel state information at the base station.

1.3.2.2 Channel hardening

Large-scale MIMO communications can offer new benefits which can not be exploited in smaller dimensions and which come from the large-scale MIMO channel properties. Channel matrix is all the better conditioned as its dimensions are high. More precisely, when the channel matrix becomes larger keeping the ratio of its dimensions fixed, the distribution of its singular values becomes less sensitive to the actual distribution of its entries. This was evoked as the Marchenko-Pastur law [7]. This phenomenon is also known as channel hardening [8]. As the size of $\boldsymbol{H}$ increases, the diagonal entries of the matrix $\boldsymbol{H}^H\boldsymbol{H}$ are relatively more prominent than the non-diagonal entries. This can be illustrated in Fig. 1.7. The channel hardening can be well exploited and offers a lot of advantages for large-scale MIMO communications. It can help simple detection scheme to achieve higher performance in large dimensions. Also as shown in Fig. 1.7 increasing the system dimensions the matrix $\boldsymbol{H}^H\boldsymbol{H}$ tends to be proportional to the identity matrix and thus enables algorithms based on matrix inversion less complex and faster. In brief, benefiting from this phenomenon, designers can exploit some approximations to propose new detection schemes with affordable complexity.

1.3.3 Challenges

Large-scale MIMO communications seem promising to establish efficient and reliable communications. With such systems one can expect multigigabit rate transmissions with hundreds of bits per channel. However, practical implementation raises many technical challenges. Low complexity signal processing algorithms are needed for channel estimation, synchronization, detection and encoding-decoding schemes. Channel estimation is one major bottleneck of the performance of such systems. Without accurate knowledge of the channel, large-scale MIMO won't keep their promises. In classical systems, CSI is acquired with the help of transmitted pilot sequences known at the receiver. The number of necessary pilots increases linearly with the system dimensions. The problem is twice. First the design of orthogonal pilot sequences is difficult when their number increases. Second it involves long sequences, which results in extended overhead and thus a loss in terms of spectral efficiency.

New encoding frame schemes are thus necessary for large-scale MIMO commu-

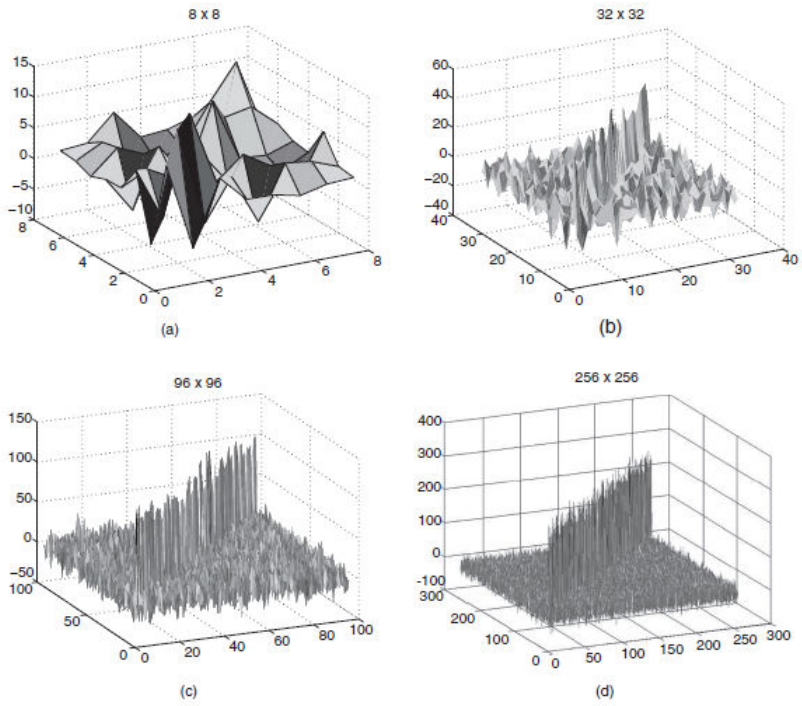

Figure 1.7: Intensity of $\boldsymbol{H}^H\boldsymbol{H}$

nications.

Another critical technological barrier is the detection of large dimensional systems. Near-optimal maximum-likelihood solutions suffer from a complexity which signifigantly increases with the system dimensions making them impractical.
Linear detection is of course compatible with practical implementation but its performance is too poor. New schemes of enhancement of existing ones are required to keep the promises of large-scale MIMO systems.

An utmost issue in large-scale MIMO communications is the multicell design. Its organization and functioning must be re-addressed like cell dimensions, neighboring base stations link, resource allocation and inter-cell interference management (interference between the users of different neighboring cells). In particular, as earlier mentioned, channel estimation requires new pilot sequence formats. The risk with existing pilot sequences is to have non-orthogonal pilot sequences between neighboring cells. This phenomenon called pilot contamination can disturb the multicell system and could be reduced by cells cooperation. Further investigation is still in-progress.

1.4 Scope of the thesis

This thesis is motivated by the new opportunities in large-scale MIMO systems and the different challenges to be addressed to make them operational. We have

undertaken two main studies.

The first challenge that the PhD aims to rise to is large-scale MIMO detection. After a deep state-of-the-art, we investigate solutions based on compressed-sensing which exploit the finite source alphabet simplicity property. The proposed algorithm is theoretically analysed and its insertion within an iterative procedure either with successive interference cancellation in the case of uncoded systems or with forward error correction code decoder is carefully designed. The algorithm and its extensions are compared to state-of-the-art schemes in terms of detection error and computation cost.

The second line of research is channel estimation. Our goal is to preserve the large-scale MIMO potential in terms of spectral efficiency by proposing efficient channel estimation algorithms that use the least possible number of pilot symbols. To that purpose, we propose to integrate the simplicity-based detection algorithms in an iterative semi-blind channel estimation in both uncoded and coded cases. An analytical study supported by simulations is carried out to evaluate the performance of the resulting CSI estimation algorithms compared to the state-of-the-art.

1.5 Outline of the thesis

The document is organized as follows:

Chapter 2 is dedicated to state-of-the–art detection algorithms for MIMO and large-scale MIMO systems. We stress weak and strong points of each algorithm.

In Chapter 3, we address the problem of large-scale MIMO detection. We exploit the simplicity property of finite alphabet signals to propose an efficient detection algorithm. We first focus on the noiseless case and we determine the necessary conditions on system dimensions for successful recovery. Then, we extend the principle to the noisy case and investigate the theoretical statistics of the algorithm output and we establish its theoretical error probability.

In Chapter 4, we insert the proposed simplicity-based detection algorithm within an iterative procedure. We first consider uncoded large-scale MIMO systems. We propose a successive interference cancellation scheme based on shadow area constraints applied to the simplicity-based algorithm. Then we focus on FEC-coded large-scale MIMO systems. We first design ML-like detection whose search area is fixed by the simplicity-based detection output. Then to further reduce the iterative receiver complexity, we propose a second scheme based on a regularization of the original detection criterion which uses the decoder output.

Chapter 5 is devoted to the channel state information estimation in both cases: uncoded and coded large-scale MIMO systems. First, we investigate semi-blind channel estimation algorithms based on the simplicity-based detection in the uncoded case. Different versions are studied. The proposed estimation schemes are theoretically analyzed and simulations support the conclusions. Second, we focus on the uplink of FEC-coded large-scale MIMO systems. We combine the iterative algorithm developed in Chapter 4 with the optimized estimation algorithm to propose

an iterative receiver which processes CSI estimation, large-scale MIMO detection and FEC decoding.

Finally Chapter 6 concludes the document and gives some perspectives.

Chapter 2

State-of-the-art of MIMO and large-scale MIMO detection algorithms

Contents

2.1 Introduction

Detection in MIMO systems is an important key to get an efficient communication. Migrating from SISO to MIMO communications, the detection task becomes more complicated because of the additional interference due to the simultaneous signals transmission from the other transmit antennas. Consequently, the design and the development of new detection algorithms are needed.

A first reflex is to extend the detection algorithms of multiuser CDMA detection to be candidates in the case of MIMO communications motivated by the structure similarities of their two system models. The algorithms which can implemented in MIMO systems are classified into two groups.

A first class is based on Maximum Likelihood (ML) criterion which needs an exhaustive search to find the solution [9]. However, in MIMO systems, by increasing the number of antennas, the ML criterion becomes an NP-hard problem which cannot be solved. An alternative solution is the sphere decoding algorithm. It is an ML-based decoding algorithm that can have a polynomial complexity for certain range of SNR values [10]. When the MIMO system dimensions increase, the sphere decoding complexity is significantly less than the exponential complexity of the ML detection. The idea is to limit the search space of lattice points to a sphere centered at the received point. However, for low and medium SNR values its complexity remains exponential which makes it not suitable when the system dimensions are high.

Another class consists of linear detection algorithms which are extended from CDMA detection. The matched filter (MF), zero forcing (ZF) and minimum mean square error (MMSE) are well-known linear detection algorithms which, contrary to the sphere decoding, can be applied when the system dimensions are high thanks to their polynomial complexity. But they achieve in such cases a poor detection performance compared to the optimum algorithms. To overcome this drawback, non-linear successive interference cancellation schemes based on linear detectors were proposed such as ZF-Successive Interference Cancellation and MMSE-SIC [11]. Their complexity keeps polynomial and the detection performance is improved compared to original linear structures. Linear detection can be also improved by using lattice reduction techniques [12].

The research goes on to find even more efficient algorithms and other domains such as machine learning and artificial intelligence emerge to look for promising solutions [13].

For any proposed detection technique, there are two criteria of interest which are the complexity and the detection performance. Hence, the objective is to get an algorithm with low complexity and good performance. This is difficult to achieve in most cases. As an example, the sphere decoder achieves the ML performance with a high complexity contrary to the linear detectors which have low complexity with low performance. The researchers are often forced to manage this trade-off which needs again to be improved especially in the case of large-scale MIMO systems. An interesting phenomenon in large-scale MIMO can be exploited, it is the channel hardening [8]. It is shown that when an overdetermined MIMO system is considered (the number of receive antennas is much higher than the number of transmit antennas), the linear detectors such as MF, ZF and MMSE perform close to the optimum thanks to the channel hardening phenomenon and become attractive from an implementation point of view. However, in this case the spectral efficiency is limited by the number of transmit antennas which must be low. Hence, we lose one of the most important benefits of large MIMO systems. To get high spectral efficiency, the two MIMO system dimensions should be large and the performance of linear detectors degrades in that case. To overcome this problem, local search-based algorithms well suited for large-scale MIMO systems like Likelihood Ascent Search (LAS) [14] and Reactive Tabu Search (RTS) represent near optimal performance

while keeping the same range of complexity as linear detectors. The task here is to get a first solution delivered by linear detection and to improve it looking for a better chosen local minimum from the solution neighbors of the linear detection output.

In this chapter, we first describe the MIMO system model and its extension to large-scale MIMO. Next, selected well-known detection algorithms are detailed. In addition to the structure description, we discuss their performance and complexity.

2.2 System model

Herein, we present the MIMO system model which is taken into account in the PhD report. Let N and n stand for the number of transmit and receive antennas respectively. We consider a noisy mixing model which can be described by the following linear equations:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\zeta}, \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{C}^N$ is the $N \times 1$ complex-valued source vector. The elements of $\boldsymbol{x}$ come from a known modulation alphabet $\boldsymbol{\beta}$ (for example QAM constellation) and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] = \boldsymbol{I}_N$, $\boldsymbol{y} \in \mathbb{C}^n$ is the $n \times 1$ complex-valued observation vector and $\boldsymbol{H} \in \mathbb{C}^{n \times N}$ is an $n \times N$ complex-valued random matrix that represents the channel matrix. We assume that the components of $\boldsymbol{H}$ are i.i.d and circularly symmetric complex Gaussian with zero mean and unit variance and $\boldsymbol{\zeta}$ is a complex AWGN noise vector with variance σ^2 such that $\mathbb{E}[\boldsymbol{\zeta}\boldsymbol{\zeta}^H] = \sigma^2 \boldsymbol{I}_n$. We assume that the channel matrix $\boldsymbol{H}$ is perfectly known at the receiver and the output statistical distribution given the input vector and the channel matrix is defined by:

$$\Pr(\boldsymbol{y}|\boldsymbol{H}, \boldsymbol{x}) = \frac{1}{(\pi\sigma^2)^n} \exp\left(-\frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2\right) \tag{2.2}$$

Most detection algorithms work directly with the complex-valued system model. For computational reasons researchers consider the equivalent real-valued system. As the vector $\boldsymbol{x}$ belongs to a complex-valued finite alphabet $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$, it can be decomposed from its real and imaginary parts as $\boldsymbol{x} = \boldsymbol{a} + j\boldsymbol{b}$ where $(\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{F}^N \times \mathcal{F}^N$ and $\mathcal{F} = \{\alpha_1, \alpha_2, .., \alpha_p\}$. We denote by $M = p^2$ the complex alphabet size. The equivalent real-valued system can be then written as:

$$\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\,\underline{\boldsymbol{x}} + \underline{\boldsymbol{\zeta}}, \qquad \underline{\boldsymbol{x}} \in \mathcal{F}^{2N}, \tag{2.3}$$

where $\underline{\boldsymbol{x}} = \begin{pmatrix}\mathrm{Re}(\boldsymbol{x}) & \mathrm{Im}(\boldsymbol{x})\end{pmatrix}^T$, $\underline{\boldsymbol{y}} = \begin{pmatrix}\mathrm{Re}(\boldsymbol{y}) & \mathrm{Im}(\boldsymbol{y})\end{pmatrix}^T$, $\underline{\boldsymbol{\zeta}} = \begin{pmatrix}\mathrm{Re}(\boldsymbol{\zeta}) & \mathrm{Im}(\boldsymbol{\zeta})\end{pmatrix}^T$ and $\underline{\boldsymbol{H}} = \begin{pmatrix}\mathrm{Re}(\boldsymbol{H}) & -\mathrm{Im}(\boldsymbol{H}) \\ \mathrm{Im}(\boldsymbol{H}) & \mathrm{Re}(\boldsymbol{H})\end{pmatrix}$.

Note that in the case of QAM constellation, the elements of the real vector $\underline{\boldsymbol{x}}$ belong to the associated PAM alphabet.

2.3 Maximum-Likelihood detection

At the receiver side, the mission is to detect the transmitted vector from the received signal. The optimal detector which minimizes the probability of detection error is ruled by the ML criterion which minimizes the Euclidean distance between the received vector $\boldsymbol{y}$ and the mixed vector $\boldsymbol{Hx}$ such that $\boldsymbol{x}$ belongs to the set $\boldsymbol{\beta}^N$. The ML solution reads then as:

$$\hat{\boldsymbol{x}}_{ML} = \underset{\boldsymbol{x}\in\boldsymbol{\beta}^N}{\arg\min} \|\boldsymbol{y} - \boldsymbol{Hx}\|_2^2 \tag{2.4}$$

To obtain the exact solution for the ML optimization criterion, we need an exhaustive search which requires exponential complexity proportional to N. Therefore, it can only be implemented for small values of N. When N is large the computation of the ML solution becomes infeasible due to the high complexity. Knowing the exact ML solution is desired in order to make it a benchmark helping designers to assess how proposed detectors perform compared to the optimal one. An alternative is the sphere decoding, but its complexity is also exponential especially in low and medium SNR values when the number of antennas N is high. To overcome this problem, designers choose to compare their algorithms taking the case of non-faded SISO AWGN for which a lower bound of ML performance can be computed.

2.3.1 Sphere decoding

The sphere decoding algorithm [15] is a detection method that performs close to the ML solution with reduced complexity. The ML real-valued system formulation reads as follows:

$$\underset{\boldsymbol{x}\in\mathcal{F}^N}{\arg\min} \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{Hx}}\|_2^2 \tag{2.5}$$

The idea behind the sphere decoding is to only consider lattice points that lie in a sphere of center $\underline{\boldsymbol{y}}$ and radius r instead of an exhaustive search over the entire lattice to reduce the search space and then the complexity of ML optimization problem. The problem here is the choice of the radius r. The larger r, the more points to test and hence the higher the complexity. However, a small value of r can lead to few points inside the sphere. The key question of the sphere decoding is to determine the optimal choice of lattice points inside the sphere. One observes that when lattice points along one-dimension are determined it is then easy to find those along the second dimension that lie in the two-dimensional sphere of the same radius r. It is thus proposed to find the lattice points successively from one dimension to the others. Let us now describe the algorithm that exploits the above observations. The lattice vector $\underline{\boldsymbol{Hx}}$ lies inside the sphere of radius r if and only if

$$\|\underline{\boldsymbol{y}} - \underline{\boldsymbol{Hx}}\|_2^2 \leq r^2 \tag{2.6}$$

One then considers the QR decomposition of the matrix $\underline{\boldsymbol{H}}$ in order to decompose the $2N$-dimensional problem into multiple one-dimensional subproblems as follows:

$$\underline{\boldsymbol{H}} = [\boldsymbol{Q}_1\boldsymbol{Q}_2]\left[\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}_{(2n-2N)\times 2N}\end{pmatrix}\right], \tag{2.7}$$

where $\boldsymbol{R}$ is a $2N \times 2N$ upper triangular matrix, $\boldsymbol{Q} = [\boldsymbol{Q}_1\boldsymbol{Q}_2]$ is a $2n \times 2n$ orthogonal matrix with $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ of dimensions $2N \times 2n$ and $2n \times (2n - 2N)$ respectively. Thereby, using the condition (2.6) we get:

$$\left\|\underline{\boldsymbol{y}} - [\boldsymbol{Q}_1\boldsymbol{Q}_2]\left[\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}_{(2n-2N)\times 2N}\end{pmatrix}\right]\underline{\boldsymbol{x}}\right\|_2^2 \leq r^2 \tag{2.8}$$

$$\|\boldsymbol{Q}_1^H\underline{\boldsymbol{y}} - \boldsymbol{R}\underline{\boldsymbol{x}}\|_2^2 + \|\boldsymbol{Q}_2^H\underline{\boldsymbol{y}}\|_2^2 \leq r^2 \tag{2.9}$$

Defining a new radius $\tilde{r}$ by $\tilde{r}^2 = r^2 - \|\boldsymbol{Q}_2^H\underline{\boldsymbol{y}}\|_2^2$, we then get:

$$\|\boldsymbol{Q}_1^H\underline{\boldsymbol{y}} - \boldsymbol{R}\underline{\boldsymbol{x}}\|_2^2 \leq \tilde{r}^2. \tag{2.10}$$

Let $\underline{\tilde{\boldsymbol{y}}} = \boldsymbol{Q}_1^H\underline{\boldsymbol{y}}$, the equation (2.10) is now equivalent to:

$$\sum_{i=1}^{2N}\left(\underline{\tilde{y}}_i - \sum_{j=1}^{2N} r_{i,j}\underline{x}_j\right)^2 \leq \tilde{r}^2, \tag{2.11}$$

where $\underline{\tilde{y}}_i$ is the ith element of $\underline{\tilde{\boldsymbol{y}}}$, $r_{i,j}$ denotes the $(i,j)^{th}$ entry of $\boldsymbol{R}$ and $\underline{x}_j$ is the jth element of $\underline{\boldsymbol{x}}$. Exploiting that $\boldsymbol{R}$ is upper triangular matrix, the expansion of (2.11) yields:

$$\left(\underline{\tilde{y}}_{2N} - r_{2N,2N}\underline{x}_{2N}\right)^2 + \left(\underline{\tilde{y}}_{2N-1} - r_{2N-1,2N}\underline{x}_{2N} - r_{2N-1,2N-1}\underline{x}_{2N-1}\right)^2 + \ldots \leq \tilde{r}^2, \tag{2.12}$$

where the first term depends only on $\underline{x}_{2N}$, the second term depends only on $\underline{x}_{2N}, \underline{x}_{2N-1}$ and so on. Therefore, to get the lattice points $\underline{\boldsymbol{H}\boldsymbol{x}}$ inside the sphere, a necessary condition is to have the first term inferior to the radius, That is to say:

$$\left(\underline{\tilde{y}}_{2N} - r_{2N,2N}\underline{x}_{2N}\right)^2 \leq \tilde{r}^2, \tag{2.13}$$

which amounts to:

$$\left\lceil\frac{-\tilde{r} + \tilde{y}_{2N}}{r_{2N,2N}}\right\rceil \leq \underline{x}_{2N} \leq \left\lfloor\frac{\tilde{r} + \tilde{y}_{2N}}{r_{2N,2N}}\right\rfloor. \tag{2.14}$$

Then, for every $\underline{x}_{2N}$ satisfying the condition (2.14), we define:

$$\tilde{r}_{2N-1}^2 = \tilde{r}^2 - \left(\underline{\tilde{y}}_{2N} - r_{2N,2N}\underline{x}_{2N}\right)^2 \tag{2.15}$$

and

$$\tilde{\underline{y}}_{2N-1}^{2N} = \tilde{\underline{y}}_{2N-1} - r_{2N,2N}\underline{x}_{2N}. \tag{2.16}$$

Therefore, we can define a stronger condition based on the two first terms in (2.12):

$$\left\lceil \frac{-\tilde{r}_{2N-1} + \tilde{y}_{2N-1}^{2N}}{r_{2N-1,2N-1}} \right\rceil \leq \underline{x}_{2N-1} \leq \left\lfloor \frac{\tilde{r}_{2N-1} + \tilde{y}_{2N-1}^{2N}}{r_{2N-1,2N-1}} \right\rfloor. \tag{2.17}$$

Based on this approach, we can continue in a similar manner for all the elements of $\underline{\boldsymbol{x}}$ and then we get all lattice points belonging to the sphere verifying (2.6). The sphere radius can be chosen proportionally to the variance σ^2 (large for low SNR ratio and small for high SNR) to ensure the true lattice point is the sphere with high probability.

The complexity of the SD detection algorithm stays exponential in N for low SNR values making it inadequate for large MIMO systems [10]. Several variants are proposed in order to reduce the complexity but the algorithm keeps impractical beyond $N = 32$ [16]. One of these variants is based on Lattice Reduction technique [17].

2.4 Linear detection

Contrary to ML detection methods, the algorithms based on linear detection apply a linear transformation $\boldsymbol{G}$ on the received vector in order to get soft estimates of the transmitted vector [18]. This explains their low polynomial complexity. Then, hard decision is taken as the nearest alphabet symbol to the soft estimate.

2.4.1 Matched filter (MF) detection

Equation (2.1) can be formulated as follows:

$$\begin{aligned} \boldsymbol{y} &= \sum_{i=1}^{N} \boldsymbol{h}_i x_i + \boldsymbol{\zeta} \\ &= \boldsymbol{h}_k x_k + \sum_{i=1, i\neq k}^{N} \boldsymbol{h}_i x_i + \boldsymbol{\zeta} \end{aligned} \tag{2.18}$$

where $\boldsymbol{h}_i$, $i = 1, 2, ..., N$ is the i-th column of the channel matrix $\boldsymbol{H}$. The first term in (2.18) is the component corresponding to the kth symbol and the second term is the interference due to the other symbols. The MF detection is based on simple linear transformation that ignores both second term and noise. The MF soft estimate of x_k is computed as:

$$\hat{x}_k = \boldsymbol{h}_k^H \boldsymbol{y}, \tag{2.19}$$

The MF output is thus given by:

$$\hat{\boldsymbol{x}}_{MF} = \boldsymbol{H}^H \boldsymbol{y}. \tag{2.20}$$

The MF linear transformation $\boldsymbol{G}^{MF} = \boldsymbol{H}^H$ has a complexity order of nN. It performs close to the optimum only when an overdetermined system is considered (i.e $n >> N$). However, its performance degrades when increases the number of sources N increases that is to pay as soon as the contribution of the second term in (2.18) becomes significant.

2.4.2 Zero forcing (ZF) detection

To enhance the performance of the MF detector, it is proposed to consider the interference contribution of the other symbols. The ZF detector is a linear detector which uses the pseudo-inverse of the channel matrix $\boldsymbol{H}$ to cancel the interference term. The transformation matrix $\boldsymbol{G}_{ZF}$ is then defined by:

$$\boldsymbol{G}^{ZF} = (\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H. \tag{2.21}$$

The ZF estimate of the symbol x_k is given by:

$$\begin{aligned} \hat{x}_k &= \boldsymbol{g}_k\boldsymbol{y} \\ &= \boldsymbol{g}_k(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\zeta}) \\ &= x_k + \boldsymbol{g}_k\boldsymbol{\zeta}, \end{aligned} \tag{2.22}$$

where $\boldsymbol{g}_k$ is the k-th column of the transformation matrix $\boldsymbol{G}$. The ZF technique completely cancels the interference from other symbols at the price of noise enhancement which is a weak point especially at low SNR. The SNR corresponding to the k-th ZF detection output equals:

$$SNR_k = \frac{|x_k|^2}{\sigma^2||\boldsymbol{g}_k||^2}. \tag{2.23}$$

Hence, the ZF technique enhances the noise variance by a factor of $||\boldsymbol{g}_k||^2$. As a result for low SNR values, the noise enhancement effect dominates and the ZF can represent a worse performance compared to the MF detector. Otherwise for high SNR, the interference cancellation dominates and the performance becomes better than the last one. The ZF detection output vector is thus given by:

$$\hat{\boldsymbol{x}}_{ZF} = \boldsymbol{G}^{ZF}\boldsymbol{y}, \tag{2.24}$$

Compared to the MF detector, the ZF detector has an increased complexity order of N^2, which remains however attractive for large-scale MIMO systems. Unfortunately, its performance degrades as N gets higher.

2.4.3 Minimum-mean square error (MMSE) detection

The MMSE detection algorithm aims at minimizing the mean square error between the transmitted vector and the transformed received vector. The transformation matrix is determined by solving the following optimization problem:

$$\arg\min_{\boldsymbol{G}} \mathbb{E}\left[\|\boldsymbol{x} - \boldsymbol{G}\boldsymbol{y}\|_2^2\right] \tag{2.25}$$

Deriving the equation (2.25), the optimal MMSE transformation matrix is given by:

$$\boldsymbol{G}^{MMSE} = (\boldsymbol{H}^H \boldsymbol{H} + \sigma^2 \boldsymbol{I}_N)^{-1} \boldsymbol{H}^H. \tag{2.26}$$

The MMSE output vector is then computed as:

$$\hat{\boldsymbol{x}}_{MMSE} = \boldsymbol{G}^{MMSE} \boldsymbol{y}, \tag{2.27}$$

The MMSE detection method performs better than the two previous linear detectors on the whole SNR range. It takes advantage of their strong points. At low SNR, it is equivalent to the MF detection ($\boldsymbol{H}^H \boldsymbol{H}$ negligible compared to $\sigma^2 \boldsymbol{I}_N$) whereas at high SNR it behaves the same as the ZF detector ($\boldsymbol{H}^H \boldsymbol{H}$ predominant compared to $\sigma^2 \boldsymbol{I}_N$). Note also that the MMSE detector needs the knowledge of the noise variance contrary to the other linear detectors described above. It represents the same order of complexity as the ZF technique due to the matrix inversion in (2.26).

2.5 Successive interference cancellation

Successive interference cancellation-based detectors [11] are considered as non-linear detectors. Symbols are estimated one after the other and their contribution to the interference is canceled immediately. The principle is to first compute the linear detection matrix and the post SINR. Then, sources are sorted according to the descending order arrangement of SINR. Sources with highest SINR are detected first to reduce the error propagation phenomenon.

An early test of MIMO wireless communication architecture using successive interference cancellation technique known as vertical BLAST (Bell Laboratories Layered Space-Time) or V-BLAST was implemented in real-time in the Bell Labs laboratory [1] in 1990s. The detection algorithm is now referred to as ZF-successive interference cancellation (ZF-SIC). The V-BLAST detection is detailed in following Algorithm 1.

Algorithm 1 V-BLAST algorithm

1: Input: $\boldsymbol{y}$, $\boldsymbol{H}$.
2: Initialization: $\boldsymbol{y}^{(0)} = \boldsymbol{y}$, $\boldsymbol{H}^{(0)} = \boldsymbol{H}$, $j = 0$.
3: Detection of the i-th symbol: $\hat{x}_i = \boldsymbol{g}_i^{(ZF)} \boldsymbol{y}$, $\boldsymbol{g}_i^{(ZF)}$ is the i-th column of $\boldsymbol{G}^{ZF}$ defined in equation (2.21).
4: Interference estimation: $\hat{\boldsymbol{a}}_i = \boldsymbol{h}_i \hat{x}_i$
5: Interference Cancellation: $\boldsymbol{y}^{(j+1)} = \boldsymbol{y}^{(j)} - \boldsymbol{h}_i \hat{x}_i$
6: Update the channel matrix: $\boldsymbol{H}^{(j+1)} = \boldsymbol{H}^{(j+1)}[1 : k-1, \boldsymbol{0}_{n\times 1}, k+1 : N]$, $j = j+1$.
7: Output: $\hat{\boldsymbol{x}}$. $=0$

A matrix inversion is done in each stage and the resulted complexity is about $O(N^4)$ with one more order compared to the original ZF algorithm.

2.6 Lattice Reduction-aided linear detection

Based on Lattice Reduction (LR)-techniques [12], we can derive new versions of linear detectors while preserving low complexity. The idea here is to transform the system model (2.1) into an equivalent system obtained by applying LR techniques. It is given by:

$$\boldsymbol{y} = \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \boldsymbol{\zeta}. \tag{2.28}$$

$\tilde{\boldsymbol{H}}$ is a transformed channel matrix designed to be better conditioned (orthogonality) than the original channel matrix $\boldsymbol{H}$. The vector $\tilde{\boldsymbol{x}}$ is a transformation of the original transmitted vector $\boldsymbol{x}$ using an unimodular matrix $\boldsymbol{T}$. The choice of $\tilde{\boldsymbol{H}}$ and $\boldsymbol{T}$ from $\boldsymbol{H}$ can be done based on a low-complexity iterative algorithm like the Seysen's Algorithm (SA) detailed in [19].
The columns of $\boldsymbol{H}$ can be interpreted as components of a lattice basis. Considering the transformed matrix $\tilde{\boldsymbol{H}} = \boldsymbol{HT}$, it generates the same lattice as $\boldsymbol{H}$ if and only if $\boldsymbol{T}$ is unimodular. The LR-technique aims to get the transformed matrix $\tilde{\boldsymbol{H}}$ better conditioned than the original one. The matrix $\tilde{\boldsymbol{H}}$ defines a new basis with vectors of shortest length being the more orthogonal possible. As the unimodular matrix has unitary determinant its inverse always exists. So, defining $\tilde{\boldsymbol{H}} = \boldsymbol{HT}$ and $\tilde{\boldsymbol{x}} = \boldsymbol{T}^{-1}\boldsymbol{x}$, the system model can be written as follows:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\zeta} \\ &= \boldsymbol{H}\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{x} + \boldsymbol{\zeta} \\ &= \tilde{\boldsymbol{H}}\tilde{\boldsymbol{x}} + \boldsymbol{\zeta}, \end{aligned} \tag{2.29}$$

Once the LR-technique is applied, the detector computes $\hat{\tilde{\boldsymbol{x}}}_{LR-ZF}$ as an estimate of $\tilde{\boldsymbol{x}}$ computed as:

$$\hat{\tilde{\boldsymbol{x}}}_{LR-ZF} = (\tilde{\boldsymbol{H}}^H\tilde{\boldsymbol{H}})^{-1}\tilde{\boldsymbol{H}}^H\boldsymbol{y}, \tag{2.30}$$

Thanks to the "higher orthogonality" of $\tilde{\boldsymbol{H}}$, the noise enhancement is reduced compared to the original ZF detector.

The ZF linear detection can be substituted with the MMSE detection which gives the following LR-MMSE detection output:

$$\hat{\boldsymbol{x}}_{LR-MMSE} = (\tilde{\boldsymbol{H}}^H\tilde{\boldsymbol{H}} + \sigma^2\boldsymbol{I}_N)^{-1}\tilde{\boldsymbol{H}}^H\boldsymbol{y}, \tag{2.31}$$

To further improve the performance, one can minimize a metric denoted by $q(\tilde{\boldsymbol{H}})$ which measures the orthogonality of the channel and which reads:

$$q(\tilde{\boldsymbol{H}}) = \sum_{i=1}^{N} ||\tilde{\boldsymbol{h}}_i||^2 ||\tilde{\boldsymbol{h}}_i'||^2 \tag{2.32}$$

The goal here is to find the matrix $\boldsymbol{T}$ that minimizes $q(\tilde{\boldsymbol{H}})$.
Note that the Lattice Reduction-based ZF/MMSE represent the same order of complexity as the ZF/MMSE algorithms that is to say $\mathcal{O}(N^3)$.

2.7 Local search-based detection

Local search technique is an heuristic method for solving NP-hard optimization problems. An important point for this technique is the choice of the neighborhood function. The idea is to start with an initial solution and then try to find a better solution on the defined neighbors. Previously linear detectors can be used to obtain the initial solution. Then the algorithm searches among the neighbors for a solution with less cost to replace the current solution and the search continues in this way. The key of this algorithm is the neighborhood definition and the stopping criteria choice. Both depend on the considered problem. The initial solution can also impact the accuracy of the final solution. There are many ways to define the neighborhood space. Let us consider a 4-QAM constellation as an example. The elements of the real-valued transmitted vector belong to the BPSK constellation with two elements and then the entire neighborhood space contains 2^{2N} vectors. This number grows exponentially with N which makes the consideration of all neighbors impossible. To overcome this problem, a simple way is to only consider the solutions that differ in only one element compared to the previous original solution. The number of neighbors to be considered is thus reduced to $2N$ and grows linearly in the system dimensions. It makes it attractive when the system dimensions are large especially for large-scale MIMO. Another way is to take into account D different elements instead of one element of the given solution to define the neighbors and we get the neighborhood size equal to $\binom{2N}{D}$. The problem of local search method is that sometimes we get stuck in a bad local optimum. Escape strategies can be considered. The main idea is to extend the solution space and to explore more solutions. A first way is to increase the value of D to define a better space. Another way is to move to the local optimum's best neighbor solution even it is worse than it and to continue the search. This can lead to another local optimum with better performance. This strategy is adopted by the Tabu search method.
From the system model (2.3) and the ML criterion, we define the function $C(\underline{\boldsymbol{x}}) = \underline{\boldsymbol{x}}^T \underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}} \underline{\boldsymbol{x}} - 2\underline{\boldsymbol{y}}^T \underline{\boldsymbol{H}} \underline{\boldsymbol{x}}$ as the ML cost which will be considered in the local search algorithms described hereinafter.

2.7.1 Likelihood ascent search (LAS)

2.7.1.1 Description

A basic version of LAS algorithm [20] is 1-LAS algorithm which just considers one element to define the neighbor space. When a local optimum is found it will be considered as a final solution. The LAS algorithm [21] starts first by estimating an initial solution $\underline{\boldsymbol{x}}^{(0)}$ which can be found by one of the linear detectors as (MF, ZF or MMSE...). The ML cost function C will be considered to refine the resulting solution at each iteration. For all $i = 1, ..., 2N$, the i-th element of the vector $\underline{\boldsymbol{x}}$ will be updated in the $(k+1)$-th iteration as follows:

$$\underline{\boldsymbol{x}}^{(k+1)} = \underline{\boldsymbol{x}}^{(k)} + \lambda_i^{(k)} \boldsymbol{e}_i, \tag{2.33}$$

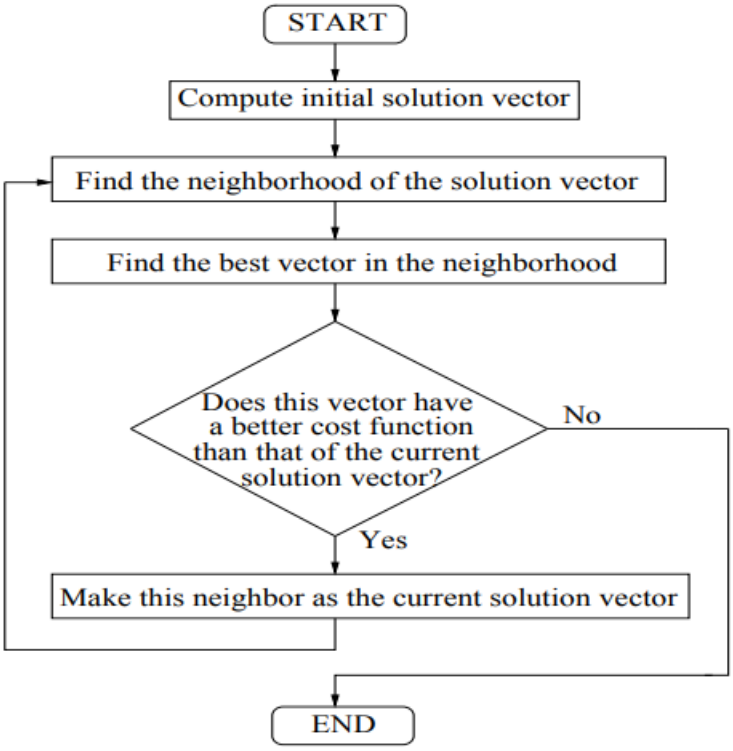

Figure 2.1: LAS algorithm flowchart

where all components of $\boldsymbol{e}_i$ equal zero except the i-th element which equals 1. In order to get $\underline{\boldsymbol{x}}^{(k+1)} \in \mathcal{F}^{2N}$, $\lambda_i^{(k)}$ can only take certain integer values that correspond to the value difference of the i-th symbol to the other symbols in $\mathcal{F}$. By denoting by $\Delta C^{(k+1)}$ the ML cost difference between solutions at iterations k and $(k+1)$, it can be written as follows:

$$\begin{aligned} \Delta C^{(k+1)} &= C(\underline{\boldsymbol{x}}^{(k+1)}) - C(\underline{\boldsymbol{x}}^{(k)}) \\ &= \left(\lambda_i^{(k)}\right)^2 (\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}})_{i,i} - 2\lambda_i^{(k)} (\underline{\boldsymbol{H}}^T (\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}^{(k)}))_i \end{aligned} \tag{2.34}$$

Let us define $\boldsymbol{z}^{(k)} = \underline{\boldsymbol{H}}^T (\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}^{(k)})$ and $z_i^{(k)}$ the i-th element of $\boldsymbol{z}^{(k)}$. Then, we get

$$\Delta C^{(k+1)} = \left(\lambda_i^{(k)}\right)^2 (\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}})_{i,i} - 2\lambda_i^{(k)} z_i^{(k)}. \tag{2.35}$$

In order to improve the solution from the k-th iteration to the $(k+1)$-th iteration, the quantity $\Delta C^{(k+1)}$ must be negative and then $\lambda_i^{(k)}$ and $z_i^{(k)}$ should have the same sign and the ML cost difference becomes:

$$\Delta C^{(k+1)} = \left(\lambda_i^{(k)}\right)^2 (\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}})_{i,i} - 2|\lambda_i^{(k)}||z_i^{(k)}|. \tag{2.36}$$

The necessary and sufficient condition to get $\Delta C^{(k+1)}$ negative is then given by:

$$|\lambda_i^{(k)}|< \frac{2|z_i^{(k)}|}{(\underline{\boldsymbol{H}}^T\underline{\boldsymbol{H}})_{i,i}} \tag{2.37}$$

A first solution to find $\lambda_i^{(k)}$ is to evaluate $\Delta C^{(k+1)}$ for all possible integer values of $\lambda_i^{(k)}$ that correspond to the moves to the other alphabet symbols but this process becomes expensive when the constellation size increases. To overcome this problem, exploiting the condition (2.37), the algorithm keeps $\boldsymbol{l}_i^{(k)} = |\lambda_i^{(k)}|$ as follows:

$$\boldsymbol{l}_{i,opt}^{(k)} = 2\left\lfloor \frac{|z_i^{(k)}|}{2(\underline{\boldsymbol{H}}^T\underline{\boldsymbol{H}})_{i,i}} \right\rfloor \tag{2.38}$$

The i-th symbol update is then done according to the following equation:

$$\underline{x}_i^{(k+1)} = \underline{x}_i^{(k)} + \boldsymbol{l}_{i,opt}^{(k)}\mathrm{sgn}\,(z_i^{(k)}), \tag{2.39}$$

The algorithm described above is called 1-LAS algorithm (one-coordinate update) because it only considers one element to define the neighbors. The final solution is a local optimum that can be improved by considering an escape strategy.

The escape strategy consists in using multiple LAS stages, each stage contributes to increase the likelihood function of the current solution. Each stage is itself splitted into different substages. The first substage updates the solution by applying the one-coordinate algorithm to reach a local optimum which will initiate next substage. The second substage uses the 2-symbols update algorithm and the solution neighbor space contains all vectors which differ in two symbols with respect to the current solution. At the end of this substage, if the likelihood function has increased, the process goes to next stage. Otherwise, it jumps to third substage where 3-symbols update algorithm is used. The procedure goes on in this way. The maximum number of substages equals D and the whole algorithm is referred to as D-LAS. The value of D is a trade-off between complexity and performance.

When the aim is to update P symbols, there are about $\binom{2N}{D}$ different ways to choose the symbols to be updated. Defining $j_1, j_2, ..., j_D$ the indices of the chosen symbols in the resulted vector at k-th iteration to be updated in the $(k+1)$-th iteration, we get:

$$\underline{\boldsymbol{x}}^{(k+1)} = \underline{\boldsymbol{x}}^{(k)} + \sum_{p=1}^{D} \lambda_{j_d}^{(k)} \boldsymbol{e}_{j_d}. \tag{2.40}$$

The elements of $\underline{\boldsymbol{x}}^{(k+1)}$ must be in $\mathcal{F}$. Then, the values of $\lambda_{j_d}^{(k)}$ should be integers for all iterations and all indices d. The ML cost difference function can be written as follows:

$$\begin{aligned}\Delta C^{(k+1)} &= \sum_{d=1}^{D}\left(\lambda_{j_d}^{(k)}\right)^2 (\underline{\boldsymbol{H}}^T\underline{\boldsymbol{H}})_{j_d,j_d} \\ &+ 2\sum_{\ell=1}^{D}\sum_{m=\ell+1}^{D} \lambda_{j_m}^{(k)}\lambda_{j_\ell}^{(k)}(\underline{\boldsymbol{H}}^T\underline{\boldsymbol{H}})_{j_m,j_\ell} - 2\sum_{d=1}^{D}\lambda_{j_d}^{(k)} z_{j_d}^{(k)}.\end{aligned} \tag{2.41}$$

The objective now is to get $\Delta C^{(k+1)}$ negative. Herein, it is possible to have multiple D-tuples $(\lambda_{j_1}^{(k)}, \lambda_{j_2}^{(k)}, ..., \lambda_{j_D}^{(k)})$ with negative difference cost so the best choice is to take into account the P-tuple with the most negative $\Delta C^{(k+1)}$.

This task is very difficult. That's why approximate methods are adopted to get lower complexity. The difference cost can be reformulated by using matrices as follows:

$$\Delta C^{(k+1)} = \boldsymbol{\Lambda}^{(k)^T} \boldsymbol{F} \boldsymbol{\Lambda}^{(k)} - 2\boldsymbol{\Lambda}^{(k)^T} \boldsymbol{z}^{(k)}, \tag{2.42}$$

where $\boldsymbol{\Lambda}^{(k)} = [\lambda_{j_1}^{(k)}, \lambda_{j_2}^{(k)}, ..., \lambda_{j_D}^{(k)}]^T$, $\boldsymbol{z}^{(k)} = [z_{j_1}^{(k)}, z_{j_2}^{(k)}, ..., z_{j_D}^{(k)}]^T$ and $\boldsymbol{F} \in \mathbb{R}^{D \times D}$ with $(\boldsymbol{F})_{m,\ell} = (\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}})_{j_m, j_\ell}$. To reduce the complexity, the optimum solution can be found by applying ZF technique which gives:

$$\boldsymbol{\Lambda}_{ZF}^{(k)} = \boldsymbol{F}^{-1} \boldsymbol{z}^{(k)}. \tag{2.43}$$

In order to get the solution verifying the necessary and sufficient conditions, the rounding function is applied to $\boldsymbol{\Lambda}_{ZF}^{(k)}$ and we get:

$$\boldsymbol{\Lambda}_{opt}^{(k)} = 2\lfloor \frac{\boldsymbol{\Lambda}_{opt}^{(k)}}{2} \rceil \tag{2.44}$$

The LAS algorithm is summarized in Fig. 2.1.

2.7.1.2 Complexity study

The complexity of the LAS algorithm is dominated by three operations. The first one is the computation of the initial solution which can be found by ZF or MMSE algorithms and induces a complexity order of $\mathcal{O}(N^2)$ per-symbol due to the matrix inversion. The second one is the calculation of $\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}}$ which represents also a complexity order of $\mathcal{O}(N^2)$ per-symbol. However, the final one which is the search operation requires a complexity order of about $\mathcal{O}(N)$ per-symbol. Therefore, the total complexity is about $\mathcal{O}(N^2)$ per-symbol dominated by the two first steps.

2.7.2 Reactive tabu search (RTS)

2.7.2.1 Description

The RTS algorithm is also a local search algorithm [22, 23]. It begins similarly to LAS algorithm by looking for an initial solution vector to be improved. However, it defines the neighbor vectors based on another neighborhood criterion. It replaces the current vector by the best neighbor vector even if it is worse in terms of ML cost function. The process continues for a number of iterations and then the best swept solution over all iterations is chosen as the final solution. When, we increase the number of iterations, there is a risk to fall into a solution already treated which is called cycling problem. To overcome this problem, RTS technique proposes to record the last moves defined as tabu and the number of iterations considered is defined as tabu period to be parametrized.

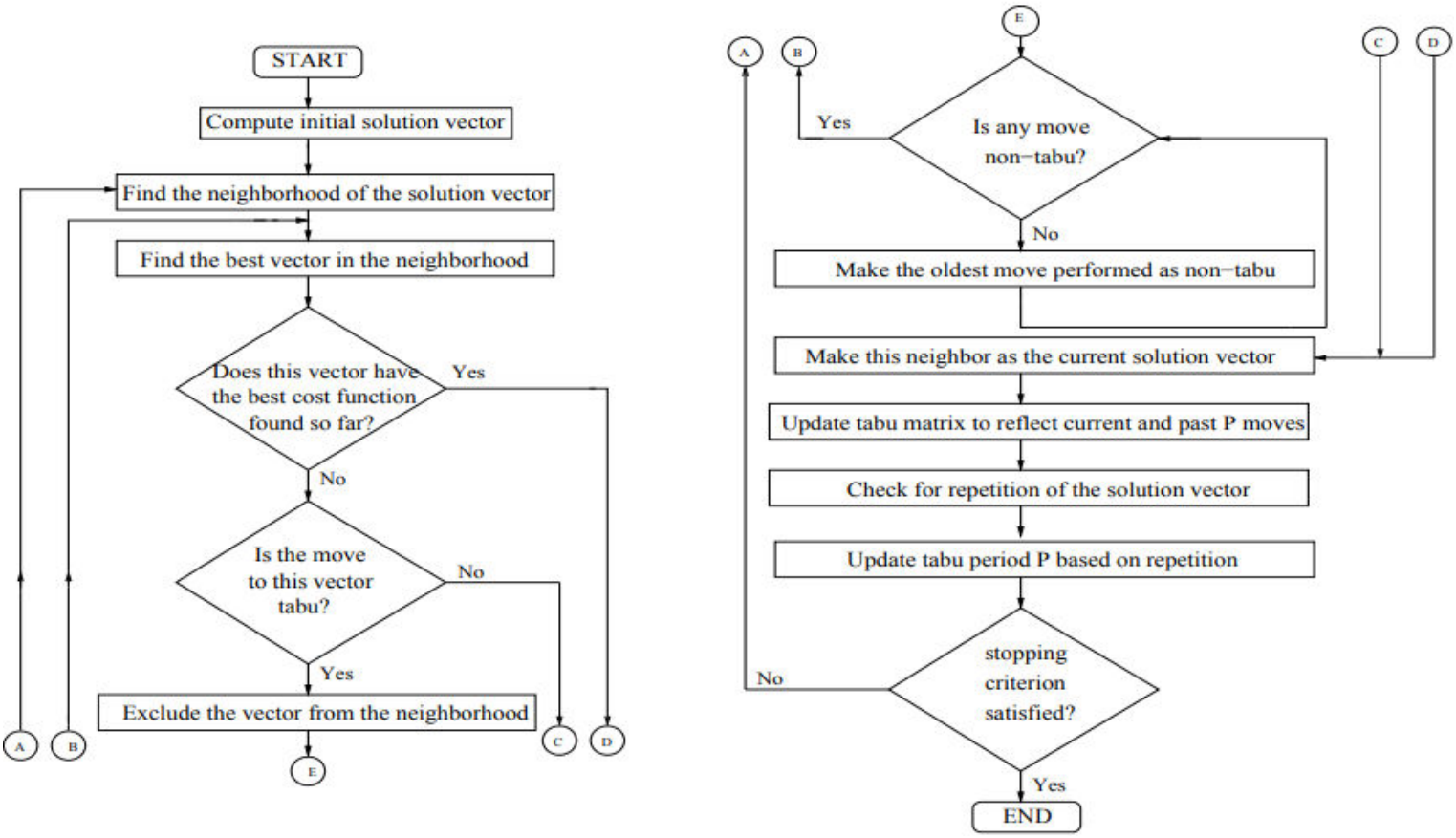

Figure 2.2: RTS algorithm flowchart

Let us consider the real valued-system model. The searched vector $\underline{\boldsymbol{x}}$ belongs to the real modulation alphabet $\mathcal{F} = \{\alpha_1, \alpha_2, .., \alpha_p\}$ with cardinality p. Define $\mathcal{N}(\alpha_i)$, $i = 1, ..., p$, the set of neighbors of the symbol α_i. Denoting its cardinality by $|\mathcal{N}(\alpha_i)| = S$ the set will contain the S nearest elements in the real-valued modulation alphabet $\mathcal{F}$ to the symbol α_i based on the Euclidean distance. Then, the neighborhood space contains the $2CN$ vectors which differ from the current in only one element. Let $w_s(\alpha_i)$, $s = 1, .., S$, denote the s-th neighbor symbol of α_i, the (r, s)-th neighbor vector of $\underline{\boldsymbol{x}} = [\underline{x}_1, \underline{x}_2, ..., \underline{x}_{2N}]$ is given by $\boldsymbol{n}(r, s) = [n_1(r, s), n_2(r, s), ..., n_{2N}(r, s)]$, $r = 1, ..., 2N$, $s = 1, ..., S$ where for $j = 1, ..., 2N$, $n_j(r, s) = \underline{x}_j$ if $j \neq r$ and $n_j(r, s) = w_s(\underline{x}_r)$ if $j = r$.

The operation of replacing the current vector $\underline{\boldsymbol{x}}^{(k)}$ by $\underline{\boldsymbol{x}}^{(k+1)} = n_j^{(k)}(r, s)$ which belongs to the neighbor set of the solution at the k-th iteration is called a (r, s) move. In order to avoid the cycling problem, a tabu matrix $\boldsymbol{T}$ of dimensions $(2Np \times S)$ is proposed. It considers all the possible moves (r, s) at any iteration and takes a positive value at some (r, s) coordinates considered as tabu values of move to designate an already treated move which must be ignored. For each element of the solution vector, there are p rows in $\boldsymbol{T}$. The rows of indices $(r - 1)p + 1$ to rp

correspond to the neighbors differing by the r-th element. The (m,ℓ)-th entry of $\boldsymbol{T}$ corresponds to the move (r,s) when $r = \lfloor (m-1)/p \rfloor + 1$, $s = \ell$ and $\underline{x}_r^{(k)} = \alpha_j$, where $j = \mod(m-1,p)+1$.

Hereinafter, we describe the RTS algorithm. First, it starts with an initial solution given for example by a linear detector such as ZF or MMSE. Let us denote by $\underline{\boldsymbol{x}}_{opt}^{(k)}$ the best solution at the k-th iteration in terms of likelihood function $C^{(k)}$ defined in (2.40). The algorithm consists of different steps.

The first step consists in determining the current solution neighbors $\boldsymbol{n}^{(k)}(r,s)$, $r = 1,...,2N$, $s = 1,...,S$ described above. Then, their ML costs $C^{(k)}$ are calculated to find the optimal move (r_*, s_*). This move should verify the following two conditions (2.45) and (2.46) to be accepted.

$$C^{(k)}(\boldsymbol{n}^{(k)}(r_*,s_*)) < C^{(k)}(\underline{\boldsymbol{x}}_{opt}^{(k)}) \tag{2.45}$$

$$\boldsymbol{T}((r_*-1)p+i,s_*) = 0. \tag{2.46}$$

When the move (r_*, s_*) does not verify these conditions, it is not accepted and another move is calculated minimizing the ML cost on the neighbor space excluding the previously found moved. Then, its acceptance is checked. When, it is also not accepted the procedure continues until a move (r_*, s_*) verifies the above two conditions. In the end, the new solution obtained at $(k+1)$-th iteration is defined by:

$$\underline{\boldsymbol{x}}_*^{(k+1)} = \boldsymbol{n}^{(k)}(r_*,s_*). \tag{2.47}$$

Let us define $i^{(k)}$, $i^{(k+1)}$, $s^{(k+1)}$ by $\alpha_{i^{(k)}} = \underline{x}_{r_*}^{(k)} = w_{s^{(k+1)}}(\underline{x}_{r_*}^{(k+1)})$ and $\alpha_{i^{(k+1)}} = \underline{x}_{r_*}^{(k+1)}$. Let d_{rep} stand for the average of iteration number between two repetitions of the same solution. It gives an idea about the number of iterations needed to get a solution for a second time. P_{tabu} is the tabu period fixed initially by the user and updated by the algorithm. It represents, when a move is considered as a tabu in an iteration, the number of next iterations until a move is accepted. Once the solution is found by (2.47), it will be checked for repetition by comparing its ML cost to those of previous solutions. When a repetition exists, d_{rep} is updated based on the number of iterations to get the found repetition. The tabu period P_{tabu} is incremented by 1. Then, we check if the number of iterations needed to change P_{tabu} from its last value exceeds γd_{rep}, where γ is a constant fixed by the user, the new value of the tabu period is $P_{tabu} = \max(1, P_{tabu}-1)$. The modifications of the tabu matrix $\boldsymbol{T}$ involved by an accepted move (r_*, s_*) are the following:
If $C^{(k)}(\underline{\boldsymbol{x}}_*^{(k+1)}) < C^{(k)}(\underline{\boldsymbol{x}}_{opt}^{(k)})$, make:

$$\boldsymbol{T}((r_*-1)p+i^{(k)},s_*) = \boldsymbol{T}((r_*-1)p+i^{(k+1)},s^{(k+1)}) = 0. \tag{2.48}$$

$$\underline{\boldsymbol{x}}_{opt}^{(k+1)} = \underline{\boldsymbol{x}}_*^{(k+1)} \tag{2.49}$$

else

$$\boldsymbol{T}((r_*-1)p+i^{(k)},s_*) = \boldsymbol{T}((r_*-1)p+i^{(k+1)},s^{(k+1)}) = P_{tabu}+1. \tag{2.50}$$

$$\underline{\boldsymbol{x}}_{opt}^{(k+1)} = \underline{\boldsymbol{x}}_{opt}^{(k)} \tag{2.51}$$

Next, the matrix $\boldsymbol{T}$ is modified again for $m = 1, ..., 2Np$, $\ell = 1, ..., S$ as:

$$\boldsymbol{T}(m, \ell) = \max\left(\boldsymbol{T}(m, \ell) - 1, 0\right) \tag{2.52}$$

Finally, the algorithm stops when the number of iterations fixed by the user is achieved or when the number of repetitions of solutions exceeds a maximum predefined value and the final solution is the best one found along the algorithm. The RTS algorithm is summarized in Fig. 2.2.

2.7.2.2 Complexity study

Like the LAS algorithm, the total complexity of the RTS algorithm mainly comes from two operations. The first one is the detection of the initial solution. The MMSE detection has for example a complexity order $\mathcal{O}(N^2 n)$. Second, the computation of $\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}}$ is also about $\mathcal{O}(N^2 n)$. The difference with the LAS algorithm is the search operation whose complexity is not deterministic. The overall complexity is dominated by the two parts to have a whole complexity of $\mathcal{O}(N^2 n)$. It makes it attractive for Large-scale MIMO systems. However, compared to the LAS algorithm, it represents an extra complexity due to the implemented escape strategy.

2.7.3 Comparison of selected local search algorithms

The LAS is a local search-based algorithm where the definition of the neighbor space is static. The first detected minimum is declared as the final solution. To enhance its performance, a multistage LAS is proposed where an escape strategy is proposed and the algorithm uses more than one coordinate for the neighborhood definition. It considers all vectors that differ in more than one element of the found solution to form the neighbor space and a better local minimum can thus be found. However, the RTS algorithm uses a dynamic neighborhood definition where some candidate vectors are removed in order to avoid the cycling problem while searching for a better solution than the previous one. Compared to LAS-based algorithms, it represents a significant performance improvement thanks to the implemented escape strategy which accepts moves to neighbors even if they imply worse performance than the current solution.

2.8 Conclusion

In this chapter, we have described the detection problem in MIMO and large-scale MIMO systems. Then, we have detailed the well-known algorithms which can be considered for large-scale MIMO communications. We have shown that they can be classified into different families. The first one contains the ML-based algorithms such as the sphere decoding (SD) which performs close to the optimum but can not be implemented in the case of large-scale MIMO due to its exponential complexity.

Then, the family of linear detection algorithms (MF, ZF and MMSE) has been presented by stressing their weak and strong points. We have exhibited large-scale MIMO contexts for which they can be attractive. Next, we have evoked that linear detectors can be improved either by using Successive Interference Cancellation or by using combinatorial analysis-based techniques such as the local search method. In next chapter, we will show how Compressive Sensing (CS) techniques can be exploited to design a detection algorithm suited to large-scale MIMO with promising performance compared to the state-of-the-art algorithms.

Chapter 3

Compressive sensing-based detection for large-scale MIMO systems

Contents

3.1 Introduction

In this chapter, we address the problem of large-scale MIMO detection. We first propose a recovery scheme for noiseless mixing systems of finite alphabet signals exploiting their *simplicity* in order to determine the limits of such systems. The term *simplicity* was first introduced by Donoho *et al.* in [24]. It is demonstrated that underdetermined mixing systems where the number of sources exceeds the

[0]This chapter was partially published in: Z. Hajji, and A. Aïssa-El-Bey, and K. Amis, "Simplicity-based recovery of finite-alphabet signals for large-scale MIMO systems" *Digital Signal Processing,vol. 80, pp. 70-82, September 2018*

number of observations can be successfully recovered exceeding the limits of the well-known Shannon-Nyquist sampling theorem. We show that this property can be exploited as the sparsity property of such signals to propose novel algorithms for Compressive Sensing (CS) applications. We then extend the proposed simplicity-based recovery scheme to the noisy case to design a low-complexity detector for large-scale MIMO systems. To our knowledge, it is the first proposed algorithm that can behave efficiently in determined and underdeterminded large-scale MIMO systems configurations. Such configuration can be expected in uplink communications, as the number of connected users times their transmit antenna number could be much higher than the base station receive antenna number.

Afterwards, we show the efficiency of the proposed scheme compared to the state-of-the-art detection algorithms by investigating the success detection conditions, the error rate performance and the computational complexity.

3.2 Overview of CS recovery and detection schemes and first tracks

3.2.1 Noise-free large-scale MIMO systems

We first consider the noise-free communications. The real-valued system model (2.3) becomes as follows:

$$\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}, \qquad \underline{\boldsymbol{x}} \in \mathcal{F}^{2N}. \tag{3.1}$$

We assume that the elements of $\mathcal{F}$ are equiprobable under the realization of $\underline{\boldsymbol{x}}$ respectively. Then, our problem is the recovery of $\underline{\boldsymbol{x}}$ from $\underline{\boldsymbol{y}}$ given $\underline{\boldsymbol{H}}$ and $\mathcal{F}$.

A special case was introduced by Mangasarian *et al.* in [25]. They considered a real-valued problem $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ with $\boldsymbol{H}$ an $n \times N$ real-valued generic random matrix[1] and the vector $\boldsymbol{x}$ belonging to the real-valued finite alphabet $\{-1, 1\}$. In this case, $\boldsymbol{x}$ can be recovered by solving the ℓ_∞-norm minimization

$$(P_\infty) : \arg\min_x \|\boldsymbol{x}\|_\infty \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}. \tag{3.2}$$

This optimization system was reformulated by a linear programming problem and the authors proved that the probability of successful recovery equals the probability that all of the columns of the generic random matrix lie in the same hemisphere. This probability is determined by the following theorem.

Theorem 3.2.1 ***Wendel [26]*** *Let $\boldsymbol{H}$ be a $n \times N$ real-valued generic random matrix. The probability that all of its columns lie in the same hemisphere is precisely equal to*

$$P_{n,N} = 2^{-N+1} \sum_{i=0}^{n-1} \binom{N-1}{i}. \tag{3.3}$$

[1] A matrix $\boldsymbol{H}$ is a generic random matrix if all sets of ℓ columns are linearly independent with probability 1 and each column is symmetrically distributed about the origin. [25]

As an extension of this work, the authors in [27] generalized the problem to all size-2 constellations $[\alpha_1, \alpha_2]$ thanks to a simple translation.

In the case of real-transformed system model given by Eq. (3.1), we demonstrate in Appendix 7.1 that given the properties of the complex-valued matrix $\boldsymbol{H}$, its real-valued matrix version $\underline{\boldsymbol{H}}$ is random generic. Then, the probability of successful recovery is equal to the probability that all of the columns of $\boldsymbol{H}$ lie in the first quadrant of the complex plane, that is to say the probability that all of the columns of $\underline{\boldsymbol{H}}$ lie in the same hemisphere. According to Wendel's theorem, this probability denoted by $Q_{n,N}$ equals $P_{2n,2N}$:

$$Q_{n,N} = P_{2n,2N} = 2^{-2N+1} \sum_{i=0}^{2n-1} \binom{2N-1}{i}. \tag{3.4}$$

In the context of underdetermined systems where the number of observations is less than the number of sources, the CS is a good candidate to separate the sources, provided the source vector is sparse. In the case of interest, the source vector isn't sparse and the symbols belong to a finite constellation with non-null elements. In order to apply recovery techniques similar to the Basis Pursuit (BP), the authors proposed in [28] a solution based on a suitable sparse transform to benefit from the combination of sparsity and finite-alphabet constraints. They succeeded in decomposing any element of the set $\mathcal{F}^{2N}$ as a sparse vector in $\mathbb{R}^{2Np}$. The sparse vector is composed of $2N$ consecutive p-uples, such that each p-uple contains one 1 and $p-1$ zeros. By proceeding so, the problem of detection becomes equivalent to a problem of sparse recovery from incomplete measurements. This problem can be seen as minimization of the ℓ_0-norm of the sparse-transformed vector subject to two constraints. The first is $\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\boldsymbol{B}_f\boldsymbol{s}$ where $\boldsymbol{s}$ is the sparse-transform of $\boldsymbol{x}$ and $\boldsymbol{B}_f = \boldsymbol{I}_{2N} \otimes \boldsymbol{f}^T$ is the transformation matrix which is defined as the Kronecker product of the identity matrix and the real-valued alphabet vector $\boldsymbol{f} = [\alpha_1, \alpha_2, ..\alpha_p]^T$. The second is the uniqueness constraint which reads $\boldsymbol{B}_1\boldsymbol{s} = \boldsymbol{1}_N$ where $\boldsymbol{B}_1 = \boldsymbol{I}_{2N} \otimes \boldsymbol{1}_p^T$. It imposes the sparse reconstruction of the searched vector. However, an ℓ_0-minimization problem is NP-hard. Therefore, to exploit the sparsity to solve the recovery and to have a problem with feasible complexity, the ℓ_0-minimization is relaxed to an ℓ_1-minimization, by mimicking literature on sparse reconstruction [29]. The optimization problem now reads

$$(P_{SA,1}) : \arg\min_{s} \ \|\boldsymbol{s}\|_1 \quad \text{subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\boldsymbol{B}_f\boldsymbol{s}, \quad \boldsymbol{B}_1\boldsymbol{s} = \boldsymbol{1}_{2N}, \tag{3.5}$$

where $\boldsymbol{s}$ is the resulted sparse vector which contains $2N$ p-tuples, each with a single element different from zero.

The main drawback of $(P_{SA,1})$ is its complexity which highly depends on the alphabet size. This makes it less interesting for higher sizes. To address the complexity issue, we have proposed another sparse decomposition which is done in two steps [30]. The first is a binary decomposition as proposed in [31] which transforms the elements of a vector in $\mathcal{F}$ into a size-$4N\log_2(p)$ vector of binary elements $\{-1, 1\}$. The second step is the application of the previous sparse decomposition to the resulting

binary vector. The problem becomes the recovery of a half-sparse vector with half of null elements. The resulting problem, denoted by $(P_{HSA,1})$, reads

$$(P_{HSA,1}) : \arg\min_{s} \|\boldsymbol{s}\|_1 \quad \text{subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\,\boldsymbol{B}_\beta\,\boldsymbol{B}_p\,\boldsymbol{s}, \quad \boldsymbol{B}_1\boldsymbol{s} = \mathbf{1}_{2\ell N}, \tag{3.6}$$

where $\ell = \log_2(M) = 2k$, $\boldsymbol{B}_1 = \boldsymbol{I}_{\ell N} \otimes \mathbf{1}_2^T$, $\boldsymbol{B}_\beta = \boldsymbol{I}_N \otimes \boldsymbol{\beta}$ with $\boldsymbol{\beta} = [2^{k-1}, ..., 2^1, 2^0]$ and $\boldsymbol{B}_\rho = \boldsymbol{I}_{\ell N} \otimes \boldsymbol{\rho}$ with $\boldsymbol{\rho} = [-1, 1]$. $\boldsymbol{B}_\beta$ defines the binary decomposition and $\boldsymbol{B}_\rho$ the half-sparse decomposition.
$(P_{HSA,1})$ is less complex than $(P_{SA,1})$ while achieving the same successful recovery probability. It reduces by about $\left(\frac{2\log_2(p)}{p}\right)^2$ the computation cost [30].

3.2.2 Noisy large-scale MIMO systems

Let us now consider the real-valued noisy large-scale MIMO system model (2.3). The main objective is to estimate $\underline{\boldsymbol{x}}$ from $\underline{\boldsymbol{y}}$ given $\underline{\boldsymbol{H}}$ and $\mathcal{F}$ by exploiting the sparse decomposition of $\underline{\boldsymbol{x}}$. This objective can be achieved by an ℓ_1-minimization problem that involves ϵ as a variable parameter depending on the current signal-to-noise ratio (SNR) value to ensure that the estimated vector is close to the emitted one. The authors proposed in [32] to apply the sparse decomposition and solve the noisy MIMO recovery problem by the following constrained ℓ_1-minimization:

$$\arg\min_{s} \|\boldsymbol{s}\|_1 \quad \text{subject to} \quad \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\boldsymbol{B}_f\boldsymbol{s}\|_2 \leq \epsilon, \quad \boldsymbol{B}_1\boldsymbol{s} = \mathbf{1}_{2N}. \tag{3.7}$$

The efficiency of the algorithm depends on the choice of ϵ. To counterbalance the critical choice of the parameter, they proposed another quadratic optimization system which can be seen as relaxation of the (ML) in another quadratic system with ℓ_0-equality as a constraint to ensure the sparsity of the searched vector. The ℓ_1 constraint is equivalent to a positivity constraint. The result is a quadratic programming model with linear equality constraints and non-negative variables. It can be resolved by polynomial-complexity algorithms. In the end, the optimization problem reads

$$(P_{SA,2}) : \arg\min_{s} \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\boldsymbol{B}_f\boldsymbol{s}\|_2 \quad \text{subject to} \quad \boldsymbol{B}_1\boldsymbol{s} = \mathbf{1}_{2N}, \quad \boldsymbol{s} \geq 0. \tag{3.8}$$

Like $(P_{SA,1})$, the complexity of $(P_{SA,2})$ highly depends on the constellation size. The same decomposition as used in $(P_{HSA,1})$ can be applied to obtain the reduced-complexity problem $(P_{HSA,2})$ which reads [30]

$$(P_{HSA,2}) : \arg\min_{s} \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\boldsymbol{B}_\beta\boldsymbol{B}_\rho\boldsymbol{s}\|_2 \quad \text{subject to} \quad \boldsymbol{B}_1\boldsymbol{s} = \mathbf{1}_{2lN}, \quad \boldsymbol{s} \geq 0. \tag{3.9}$$

In this chapter, we present a new method for compressive sensing that does not require the sparsity of the signal to be recovered. It exploits the alphabet properties and looks for a solution in a convex space containing the alphabet elements. Compared to previously described methods $(P_{HSA,i})$ and $(P_{SA,i})$, it brings further complexity reduction to adapt to high finite-alphabet size with recovery performance conservation.

3.3 Simplicity property exploitation to solve the noise-free recovery

3.3.1 Proposed method definition and theoretical study

In this section, we consider the real-valued noise-free system (3.1) and we propose a new recovery scheme. The maximum likelihood (ML) detector requires an exhaustive search over all possible mixed symbol vectors and selects the solution that corresponds to the closest point to the searched signal in the known alphabet [15]. In other words, it selects the vector with elements in the alphabet that satisfies the equality $\underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}$.

$$(P_{ML,1}) : \arg\min_{\underline{\boldsymbol{x}}} \mathbf{1}_{2N}^T \underline{\boldsymbol{x}} \quad \text{subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}, \quad \underline{\boldsymbol{x}} \in \mathcal{F}^{2N}. \tag{3.10}$$

The main drawbacks of this detector are twofold: first, the $(P_{ML,1})$ criterion is not convex and second, it suffers from high computational complexity caused by the exhaustive search over the set $\mathcal{F}^{2N}$. Herein, we propose a new detection scheme which is based on a relaxation of the ML detector constraint. We relax the solution space $\{\underline{\boldsymbol{x}} \in \mathcal{F}^{2N} | \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}\}$ by substituting it with the convex set $\{\underline{\boldsymbol{x}} \in [\alpha_1, \alpha_p]^{2N} | \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}\}$. Exploiting this relaxation, the new resulting optimization problem can be resolved by polynomial algorithms for convex optimization using the following proposition.

Proposition 3.3.1 *$\underline{\boldsymbol{x}} \in [\alpha_1, \alpha_p]^{2N}$ is the unique solution to the problem*

$$\arg\min_{\underline{\boldsymbol{x}}} \ \mathbf{1}_{2N}^T \underline{\boldsymbol{x}} \quad \text{subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}, \quad \underline{\boldsymbol{x}} \in \mathcal{F}^{2N}$$

if and only if its corresponding vector $\boldsymbol{r} \in \mathbb{R}^{4N}$ is the unique solution to the optimization problem:

$$\arg\min_{\boldsymbol{r}} \ \mathbf{1}_{4N}^T \boldsymbol{r} \quad \text{subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{B}_\alpha}} \boldsymbol{r}, \quad \boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}, \quad \boldsymbol{r} \geq 0, \tag{3.11}$$

where $\boldsymbol{B}_\alpha$ is defined as $\boldsymbol{B}_\alpha = \boldsymbol{I}_{2N} \otimes [\alpha_1, \alpha_p]$.

[Proof of Proposition 3.3.1] *Let $\mathcal{G} = \{\underline{\boldsymbol{x}} \in \mathbb{R}^{2N} | \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}, \underline{\boldsymbol{x}} \in [\alpha_1, \alpha_p]^{2N}\}$ and $\mathcal{H} = \{\underline{\boldsymbol{x}} \in \mathbb{R}^{2N} | \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{x}}}, \underline{\boldsymbol{x}} = \boldsymbol{B}_\alpha \boldsymbol{r}; \boldsymbol{r} \in \mathbb{R}^{4N}, \boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}$ and $\boldsymbol{r} \geq 0\}$. $\mathcal{H}$ stands for the feasible set of the problem defined in (3.11). Then, it suffices to show the equality between $\mathcal{G}$ and $\mathcal{H}$.*
Suppose that $\underline{\boldsymbol{x}} \in \mathcal{G}$. Then the i-th element of $\underline{\boldsymbol{x}}$ can be written as $\underline{x}_i = r_{2i}\alpha_1 + r_{2i+1}\alpha_p$ where $r_{2i} + r_{2i+1} = 1$, $0 \leq r_{2i}, r_{2i+1} \leq 1$. Thus $\underline{\boldsymbol{x}} \in \mathcal{H}$. Reciprocally $\mathcal{H} \subset \mathcal{G}$. We deduce that $\mathcal{G} = \mathcal{H}$. It is important to mention that due to the positivity constraint and the fact that the ℓ_1-norm of a vector is the sum of the absolute values of its elements, the proposed optimization system is equivalent to the following ℓ_1-minimization problem:

$$(P_{SI,1}) : \arg\min_{r} \ \|\boldsymbol{r}\|_1 \text{ subject to} \quad \underline{\boldsymbol{y}} = \underline{\boldsymbol{\boldsymbol{H}\boldsymbol{B}_\alpha}} \boldsymbol{r}, \quad \boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}. \tag{3.12}$$

The new optimization problem $(P_{SI,1})$ is a linear programming model with linear equality constraints. It can be solved by the simplex [33] or the interior point methods [34]. On the whole of the Phd dissertation, we take an interest in the algorithms based on the interior point methods. These algorithms start by finding an interior point of the polytope defined by the constraints and then proceed to the optimal solution by moving inside the polytope.

In order to study the necessary and sufficient conditions of the solution uniqueness of the proposed optimization problem, we exploit the geometry of the system model and we utilize a face counting technique [24]. The following theorem gives the solution uniqueness probability from which we can derive the conditions of successful recovery.

Theorem 3.3.1 **(*i*)** *Given the alphabet size $p \geq 2$, if $\underline{\boldsymbol{H}}$ is a $2n \times 2N$ generic random complex matrix, the probability that $(P_{SI,1})$ has a unique solution is given by:*

$$Q_{n,N}(p) = \sum_{k=0}^{2n-1} \binom{2N}{k} \left(\frac{2}{p}\right)^{2N-k} \left(\frac{p-2}{p}\right)^{k} P_{2n-k,2N-k}. \tag{3.13}$$

(*ii*) *By assuming that (n, N) grows proportionally, $Q_{n,N}(p)$ tends to 0 when $\frac{n}{N} < \frac{p-1}{p}$ and tends to 1 when $\frac{n}{N} > \frac{p-1}{p}$.*

[**Proof of Theorem 3.3.1**] The proof of Statement **(*i*)** of Theorem 3.3.1 requires the introduction of the simplicity concept defined herinafter.

Definition 3.3.1 ***Simplicity [35]*** *A given vector $\underline{\boldsymbol{x}} \in [\alpha_1, \alpha_p]^{2N}$ is called k-simple if it has exactly k entries different from α_1 and α_p.*

According to Theorem 3.3.1, calculating the solution uniqueness probability of the optimization problem amounts to calculate the probability that the equation $\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}; \underline{\boldsymbol{x}} \in [\alpha_1, \alpha_p]^{2N}$ has only one root. It can also be formulated as: $\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}} \in \mathcal{P}$ with $\mathcal{P} = \{\underline{\boldsymbol{x}} \in \mathbb{R}^{2N} | \underline{\boldsymbol{x}} = \boldsymbol{B}_\alpha \boldsymbol{r};\ \boldsymbol{r} \in \mathbb{R}^{4N},\ \boldsymbol{B}_1 \boldsymbol{r} = \boldsymbol{1}_{2N} \text{ and } \boldsymbol{r} \geq 0\}$ where $\boldsymbol{B}_\alpha = \boldsymbol{I}_{2N} \otimes \boldsymbol{\alpha}$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_p]$ and $\boldsymbol{B}_1 = \boldsymbol{I}_{2N} \otimes \boldsymbol{1}_2^T$. The convex hull of this polytope which is the minimal convex set containing all the elements of $\mathcal{P}$ is the set $\{\boldsymbol{h} \in \mathbb{R}^{2N} | h_i = \alpha_1 \text{ or } h_i = \alpha_p, 1 \leq i \leq 2N\}$. It contains all the vectors with entries equal to the bounds. Let $\underline{\boldsymbol{x}} \in \mathbb{R}^{2N}$ be a k-simple vector in $\mathcal{P}$, in other words, with exactly k entries strictly different from the elements of the vectors of the convex hull of $\mathcal{P}$. Let F denote the associated k-face of $\mathcal{P}$. Given the system $\underline{\boldsymbol{y}} = \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}} \in \mathcal{P}$, it is showed in [35, Lemma 5.2] that the condition of solution unicity is equivalent to the condition that $\underline{\boldsymbol{H}}F$ is a k-face of $\underline{\boldsymbol{H}}\mathcal{P}$. In [35, Theorem 1.10], exploiting the fact that $\underline{\boldsymbol{H}}$ is completely general[2], it is demonstrated that this condition is satisfied with probability $1 - P_{2(N-n),2N-k} = P_{2n-k,2N-k}$. As the elements of $\underline{\boldsymbol{x}}$ take on values with equal probability in the set $\mathcal{F} = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$, the probability that

[2]For definition and proof see Appendix 7.1

	Dimension	*Cost per iteration*	*Total*
$(P_{SA,1})$	MN	$\mathcal{O}(M^2N^2(N+n))$	$\mathcal{O}(M^2N^2(N+n)^{3/2})$
$(P_{HSA,1})$	$\log_2(M)N$	$\mathcal{O}((\log_2(M))^2N^2(N+n))$	$\mathcal{O}((\log_2(M))^2N^2(N+n)^{3/2})$
(P_{∞})	$2N$	$\mathcal{O}(N^2(N+n))$	$\mathcal{O}(N^2(N+n)^{3/2})$
$(P_{SI,1})$	$2N$	$\mathcal{O}(N^2(N+n))$	$\mathcal{O}(N^2(N+n)^{3/2})$

Table 3.1: Computation cost with the interior point method.

it is k-simple is $\binom{2N}{k}\left(\frac{2}{p}\right)^{2N-k}\left(\frac{p-2}{p}\right)^{k}$. According to Bayes' axiom the probability that $\underline{\boldsymbol{x}}$ is the unique k-simple solution is equal to $\binom{2N}{k}\left(\frac{2}{p}\right)^{2N-k}\left(\frac{p-2}{p}\right)^{k} P_{2n-k,2N-k}$. Hence, the proof of Statement $(\boldsymbol{i})$. Now that Statement $(\boldsymbol{i})$ is established, the proof of Statement $(\boldsymbol{ii})$ can be obtained by following the same reasoning as in [27, Proof of Theorem 3, page 2012].

3.3.2 Complexity Analysis

The interest of the proposed detection scheme comes from its complexity order. The CVX toolbox relies on the interior point method whose complexity is a function of the number of constraints and the dimension of the searched vector [36, 37]. A convex optimization problem defined over $\mathbb{R}^m$ subject to d constraints requires, in the worst case, $\mathcal{O}(\sqrt{d})$ iterations for a computation cost order of $\mathcal{O}(m^2d)$ per iteration and yields a total computation cost order equal to $\mathcal{O}(m^2d^{3/2})$ [38]. Applied to the different convex optimization problems dealt with in this section, we obtain the computation costs reported in Table 3.1. According to these estimations, the half-sparse decomposition enables to reduce the computation cost by $\left(\frac{p}{\log_2(p)}\right)^2$. However, in the case of binary alphabets, (P_{∞}) is the most interesting because its complexity order is the lowest. For higher-size alphabets, (P_{∞}) does not apply and $(P_{SI,1})$ becomes the most relevant problem to solve. Its computation cost is the same as (P_{∞}) and keeps constant whatever the alphabet size.

3.3.3 Simulation results

The following simulation results illustrate the theoretical framework exposed above. The experimental setup is common to all simulations. We use even values of p and choose $\mathcal{F} = \{\pm(2k-1) : k = \{1, 2, ..., p/2\}\}$. For each simulation, we fix $N \in \{64, 128, 256\}$ and make n vary so as to assess a significant number of values for the ratio of n/N. For each pair (n, N), 1000 iterations are carried out. For each iteration, we generate a realization of the complex-valued generic random matrix of size $n \times N$. The matrix coefficients are independent and identically Gaussian distributed with zero mean and unit variance. We then transform it in a real-valued formulation with size $2n \times 2N$. We generate $\underline{\boldsymbol{x}}$ with $2N$ entries drawn uniformly from $\mathcal{F}$. We solve the optimization problem by using the Matlab CVX toolbox [36]. The simulation results are obtained by using a PC equipped with Linux Ubuntu

14.04 OS, Intel Core i3-2350M processor (2.3 GHz) and 8GB RAM. A solution $\hat{\underline{\boldsymbol{x}}}$ is returned and we assume that the recovery is correct if the relative error $\frac{||\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{x}}||_2}{||\underline{\boldsymbol{x}}||_2}$ is less than 10^{-6}.

Fig. 3.1 is the phase diagram in the case $p = 2$ for the proposed simplicity-based approach. The simulated successful recovery probability corroborates Theorem 3.3.1 and it coincides with the analytical expression. In particular, we observe that the breakpoint occurs when $n/N = 1/2$ with a successful recovery probability equal to one half, as established theoretically. We also recall that the proposed simplicity-based recovery method performs the same as Mangasarian approach.

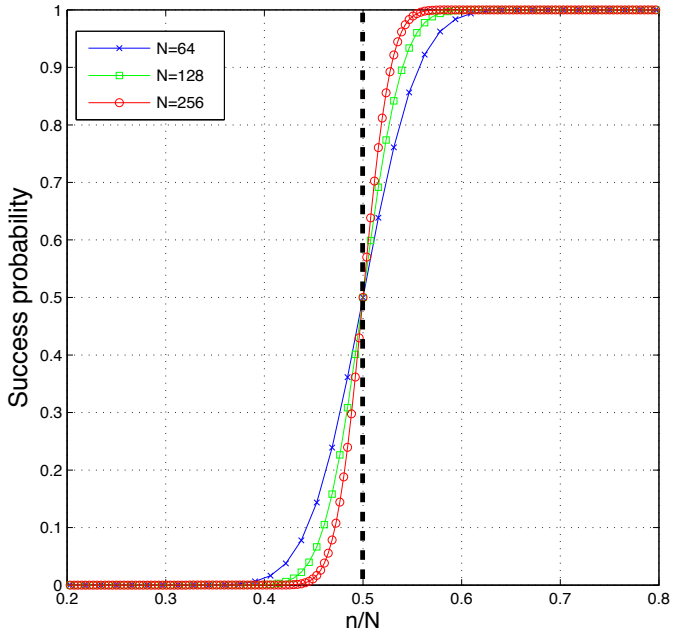

Figure 3.1: Phase diagrams of the proposed method, for $p = 2$ and $N \in \{64, 128, 256\}$.

We now address the case $p > 2$, for which the Mangasarian *et al.* approach is not applicable. Fig. 3.2 provides the phase diagrams of the proposed approach ($P_{SI,1}$) for $p = 4$ and $p = 8$ and different values of N. The simulated successful recovery probability coincides with the analytical expression. The simulation results confirm the theoretical study presented above. The breakpoint occurs when $\frac{n}{N}$ equals the value $\frac{p-1}{p}$, that is to say $\frac{3}{4}$ for $p = 4$ and $\frac{7}{8}$ for $p = 8$.

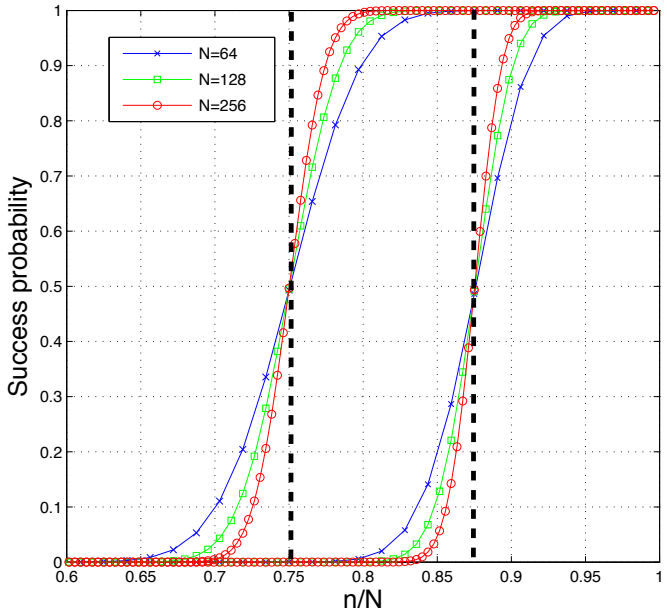

Figure 3.2: Phase diagrams of the proposed method, for $p = 4$, $p = 8$ and $N \in \{64, 128, 256\}$.

3.4 Application of the simplicity principle to noisy large-scale MIMO systems

3.4.1 Proposed method definition and theoretical analysis

We consider in this section the real scenario of large MIMO communication when a noisy channel is considered as (2.3). We propose to recover the vector $\underline{\boldsymbol{x}}$ thanks to the previously introduced decomposition that exploits the fact that the real-transformed symbols belong to the interval $[\alpha_1, \alpha_p]$.
We proceed in the same way as in the noise-free case and based on the Theorem 3.3.1, we propose to solve the following optimization problem:

$$FAS : \arg\min_{\boldsymbol{r}} \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\boldsymbol{B}_\alpha \boldsymbol{r}\|_2 \quad \text{subject to} \quad \boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}, \quad \boldsymbol{r} \geq 0.$$

The resulted detector is referred to as Finite Alphabet Simplicity (FAS) detector in the remaining of the PhD dissertation. To evaluate its performance, we search for the conditions for a stationary point and we investigate the statistical distribution of the detection output. Let us define the necessary sets used to establish the analytical results.

Definition 3.4.1 *Let Ω, the set of active constraints, defined by $\Omega = \{i | \underline{x}_i = \alpha_1$ or $\underline{x}_i = \alpha_p\}$ and Λ the set of binding constraints defined by $\Lambda = \{i | \underline{x}_i = \alpha_1$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \geq 0$ or $\underline{x}_i = \alpha_p$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \leq 0\}$. The cardinality of Λ*

defines the simplicity order of the searched vector.
The complementary set $\bar{\Lambda}$ *corresponds to the non-binding constraint set. Its cardinality is denoted by* $\mathcal{C} = \mathrm{Card}(\bar{\Lambda})$.

The following theorem gives the conditions for a solution to be a stationary point for FAS.

Theorem 3.4.1 ***Stationary point condition [39]***
$\hat{\boldsymbol{r}}$ *is a stationary point for FAS (a point satisfying the first order necessary conditions for optimality) if and only if* $\underline{\hat{\boldsymbol{x}}} = \boldsymbol{B}_\alpha \hat{\boldsymbol{r}}$ *is feasible and* $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\underline{\hat{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_{\bar{\Lambda}} = \boldsymbol{0}$.

The output solution verifies the stationary point condition. In order to define the theoretical performance and to enhance the recovery performance either by an iterative scheme or by the addition of a forward error correction code soft-decision decoder, we need the probability density function of the detector output. The following theorem defines its theoretical statistical distribution.

Theorem 3.4.2 ***Statistical distribution of the detection output***
Let $\hat{\boldsymbol{r}}$ *the solution of FAS. Then the components of* $\underline{\hat{\boldsymbol{x}}} = \boldsymbol{B}_\alpha \hat{\boldsymbol{r}}$ *follow a censored normal distribution given by*

$$f_{\underline{\hat{x}}_k}(x) = \frac{1}{p}\sum_{j=1}^{p} f_{\underline{\hat{x}}_k | x_k = \alpha_j}(x), \tag{3.14}$$

with

$$\begin{aligned} f_{\underline{\hat{x}}_k | \underline{x}_k = \alpha_j}(x) = \Bigg(& \frac{1}{2}\mathrm{erfc}\left(\frac{\alpha_j - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\delta_{\alpha_1}(x) + \frac{1}{2}\mathrm{erfc}\left(\frac{\alpha_p - \alpha_j}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\delta_{\alpha_p}(x) \\ & + \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}}\exp\left(-\frac{(x-\alpha_j)^2}{2\sigma^2_{\underline{\hat{x}}}}\right)\mathbb{1}_{[\alpha_1,\alpha_p]}(x)\Bigg) \end{aligned} \tag{3.15}$$

and

$$\sigma^2_{\underline{\hat{x}}} = \sum_{k=0}^{2n-2}\binom{2N}{k}\left(\frac{1}{p}\right)^{2N-k}\left(\frac{p-1}{p}\right)^k \frac{2n\sigma^2}{2n-k-1}, \tag{3.16}$$

where $\sigma^2 = \mathbb{E}[\underline{\zeta}_i^2]$, $\forall i = 1, \ldots, 2n$.

[Proof of Theorem 3.4.2] Let us consider $\hat{\boldsymbol{r}}$ a stationary point of FAS problem. Then $\underline{\hat{\boldsymbol{x}}} = \boldsymbol{B}_\alpha \hat{\boldsymbol{r}}$ becomes feasible and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\underline{\hat{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_{\bar{\Lambda}} = \boldsymbol{0}$ (see Theorem 3.4.1). Let $\underline{\boldsymbol{H}}_{\bar{\Lambda}} = [\underline{\boldsymbol{H}}_{:,k}]_{k\in\bar{\Lambda}}$ with $\underline{\boldsymbol{H}}_{:,k}$ the k^{th} column of $\underline{\boldsymbol{H}}$. We first assume that $\mathcal{C} = \mathrm{Card}(\bar{\Lambda})$ satisfies $\mathcal{C} < 2n-1$, so that the Moore-Penrose pseudo-inverse of $\underline{\boldsymbol{H}}_{\bar{\Lambda}}$ exists and equals $\underline{\boldsymbol{H}}^\dagger_{\bar{\Lambda}} = (\underline{\boldsymbol{H}}^T_{\bar{\Lambda}}\underline{\boldsymbol{H}}_{\bar{\Lambda}})^{-1}\underline{\boldsymbol{H}}^T_{\bar{\Lambda}}$. Therefore, the restriction of $\underline{\hat{\boldsymbol{x}}}$ to the index set $\bar{\Lambda}$ reads

$$\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}} = \underline{\boldsymbol{H}}^\dagger_{\bar{\Lambda}}(\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}_{\Lambda}\underline{\hat{\boldsymbol{x}}}_{\Lambda}). \tag{3.17}$$

From Eq. (3.17) and exploiting the property that the set of binding constraints Λ can be seen as the set of indexes of entries of $\underline{\boldsymbol{x}}$ which were correctly estimated,

according to the central limit theorem, given $\underline{\boldsymbol{x}}$ and $\mathcal{C}$, $\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}}$ is normally distributed with mean $\underline{\boldsymbol{x}}_{\bar{\Lambda}}$. To compute the covariance matrix $\boldsymbol{\Sigma}_{\underline{\hat{\boldsymbol{x}}}_{\bar{\boldsymbol{\Lambda}}}}$, we exploit the fact that the number of non-binding constraints $\mathcal{C}$ is a random variable. Therefore, $\boldsymbol{\Sigma}_{\underline{\hat{\boldsymbol{x}}}_{\bar{\boldsymbol{\Lambda}}}}$ is given by:

$$\boldsymbol{\Sigma}_{\underline{\hat{\boldsymbol{x}}}_{\bar{\boldsymbol{\Lambda}}}} = \mathbb{E}\left[\mathbb{E}[(\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}} - \mathbb{E}[\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}}])(\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}} - \mathbb{E}[\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}}])^T|\mathcal{C}=k]\right], \tag{3.18}$$

with

$$\begin{aligned}\mathbb{E}[(\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}} - \mathbb{E}[\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}}])(\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}} - \mathbb{E}[\underline{\hat{\boldsymbol{x}}}_{\bar{\Lambda}}])^T|\mathcal{C}=k] &= \sigma^2\mathbb{E}[(\underline{\boldsymbol{H}}_{\bar{\Lambda}}^T\underline{\boldsymbol{H}}_{\bar{\Lambda}})^{-1}|\mathcal{C}=k] \\ &= \sigma^2\frac{2n}{2n-k-1}\boldsymbol{I}_k,\end{aligned} \tag{3.19}$$

where we have used that, given $\mathcal{C}=k$, the matrix $(\underline{\boldsymbol{H}}_{\bar{\Lambda}}^T\underline{\boldsymbol{H}}_{\bar{\Lambda}})^{-1}$ follows an inverse Wishart distribution and then $\mathbb{E}[(\underline{\boldsymbol{H}}_{\bar{\Lambda}}^T\underline{\boldsymbol{H}}_{\bar{\Lambda}})^{-1}|\mathcal{C}=k] = \frac{2n}{2n-k-1}\boldsymbol{I}_k$ (see [40]). The distribution of $\mathcal{C}$ is provided by the following Proposition 4.2.1.

Proposition 3.4.1 *The number of non binding constraints introduced in Definition 3.4.1 follows the binomial distribution with parameters $2N$ and $\frac{p-1}{p}$:*

$$\mathcal{C} = \mathrm{Card}(\bar{\Lambda}) \sim \mathcal{B}(2N, \frac{p-1}{p})$$

The proof of Proposition 4.2.1 is given in Appendix 7.2. From Proposition 4.2.1, we observe that the probability of event $\mathcal{C} \geq 2n-1$ is not significant and these events will thus be neglected in the computation of $\boldsymbol{\Sigma}_{\underline{\hat{\boldsymbol{x}}}_{\bar{\boldsymbol{\Lambda}}}}$. We then obtain

$$\boldsymbol{\Sigma}_{\underline{\hat{\boldsymbol{x}}}_{\bar{\boldsymbol{\Lambda}}}} = \sigma^2_{\underline{\hat{\boldsymbol{x}}}}\boldsymbol{I}_{\mathcal{C}}, \tag{3.20}$$

with

$$\sigma^2_{\underline{\hat{\boldsymbol{x}}}} = \sum_{k=0}^{2n-2}\Pr\left(\mathcal{C}=k\right)\frac{2n\sigma^2}{2n-k-1} = \sum_{k=0}^{2n-2}\binom{2N}{k}\left(\frac{1}{p}\right)^{2N-k}\left(\frac{p-1}{p}\right)^k\frac{2n\sigma^2}{2n-k-1}.$$

Finally, the constraints of FAS impose that $\underline{\hat{x}}_i \in [\alpha_1, \alpha_p]$ and we deduce that, given $\underline{\boldsymbol{x}}$, the components of $\underline{\hat{\boldsymbol{x}}}$ corresponding to the non-binding constraints follow a truncated normal distribution with mean $\underline{\boldsymbol{x}}$ and variance $\sigma^2_{\underline{\hat{\boldsymbol{x}}}}$.

As for the components of $\underline{\hat{\boldsymbol{x}}}$ corresponding to the binding constraints, they satisfy either $\underline{x}_i = \alpha_1$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\underline{\hat{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \geq 0$, or $\underline{x}_i = \alpha_p$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\underline{\hat{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \leq 0\}$. We thus conclude that they follow a binary distribution with probability $\frac{1}{2p}$ (see 7.2 for the justification of $\Pr(\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\underline{\hat{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \geq 0) = \frac{1}{2}$).
Consequently, the conditional distribution of $\underline{\hat{x}}_k$ given x_k reads

$$\begin{aligned}f_{\underline{\hat{x}}_k|\underline{x}_k=\alpha_j}(x) = &\left(\frac{1}{2}\mathrm{erfc}\left(\frac{\alpha_j-\alpha_1}{\sqrt{2}\sigma_{\underline{\hat{\boldsymbol{x}}}}}\right)\delta_{\alpha_1}(x) + \frac{1}{2}\mathrm{erfc}\left(\frac{\alpha_p-\alpha_j}{\sqrt{2}\sigma_{\underline{\hat{\boldsymbol{x}}}}}\right)\delta_{\alpha_p}(x)\right. \\ &\left. + \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}}\exp\left(-\frac{(x-\alpha_j)^2}{2\sigma^2_{\underline{\hat{\boldsymbol{x}}}}}\right)1_{[\alpha_1,\alpha_p]}(x)\right).\end{aligned}$$

As demonstrated, we can see the distribution of the output vector as a mix of a binary distribution due to the simplicity of the vector and a truncated normal distribution with variance depending on the system dimensions and the noise variance. Depending on the system dimensions, the exact variance computation may be too complex in practice. The following lemma provides an approximation of the variance which can be simply calculated. Its accuracy will be studied in the simulation section.

Lemma 3.4.1 ***Variance approximation***
Let $\hat{\boldsymbol{r}}$ *the solution of FAS. The variance of the output vector* $\underline{\hat{\boldsymbol{x}}} = \boldsymbol{B}_\alpha \hat{\boldsymbol{r}}$ *can be approximated for* $n \geq N\left(\frac{p-1}{p}\right) + 1$ *as*

$$\sigma^2_{\underline{\hat{\boldsymbol{x}}}} \approx \frac{2n\sigma^2}{2n - 2N\left(\frac{p-1}{p}\right) - 1}. \tag{3.21}$$

From the statistical distribution of the detection output and by exploiting the general results available in [41], a lower bound of the symbol error probability can be obtained. This bound is asymptotically reached when the SNR gets high, which provides an approximation of the symbol error probability. It can be used to predict performance without simulating the whole system and its accuracy will be checked in the simulation part.

Theorem 3.4.3 ***Symbol Error Probability upper-bound***
The symbol error probability in the case of a M-ary QAM constellation can be upper-bounded by:

$$P_s \leq \frac{1}{2p} \sum_{k=1}^{p} \sum_{\substack{j=1 \\ j \neq k}}^{p} \operatorname{erfc}\left(\frac{\alpha_j - \alpha_k}{2\sqrt{2}\sigma_{\underline{\hat{\boldsymbol{x}}}}}\right) + \frac{p-1}{2p} \sum_{i=1}^{p} \operatorname{erfc}\left(\frac{\alpha_p - \alpha_i}{\sqrt{2}\sigma_{\underline{\hat{\boldsymbol{x}}}}}\right). \tag{3.22}$$

This upper bound can be used as a tight approximation of the symbol error probability. For high SNR, the symbol error probability can be further approximated by:

$$P_s \approx \frac{p-1}{p} \operatorname{erfc}\left(\frac{\alpha_2 - \alpha_1}{2\sqrt{2}\sigma_{\underline{\hat{\boldsymbol{x}}}}}\right). \tag{3.23}$$

3.4.2 Complexity Analysis

Table 3.2 summarizes the complexity order of the decomposition-based detectors including the proposed one referred to as FAS, the MMSE, the MMSE-SIC, the MMSE-SIC-LR, and the SD detector. The SD detector is a high-complexity detector especially when the modulation order or the number of antennas increase, it is the least cost efficient. The MMSE-based detector consists of one matrix inversion and some matrix multiplications and additions. The MMSE-SIC adds some order of complexity. According to the complexity analysis in [12], the additional complexity order involved in the MMSE-SIC-LR due to lattice reduction is equal to $\mathcal{O}(N^2 \log B)$

	Iteration number	*Cost per iteration*	*Total*
MMSE	1	$\mathcal{O}(N^3)$	$\mathcal{O}(N^3)$
MMSE-SIC	1	$\mathcal{O}(N^3)+\mathcal{O}(MN^2)$ $+\mathcal{O}(M^2N)$	$\mathcal{O}(N^3)+\mathcal{O}(MN^2)$ $+\mathcal{O}(M^2N)$
MMSE-SIC-LR	1	$\mathcal{O}(N^3)+\mathcal{O}(MN^2)$ $+\mathcal{O}(M^2N)+\mathcal{O}(N^2\log B)$	$\mathcal{O}(N^3)+\mathcal{O}(MN^2)$ $+\mathcal{O}(M^2N)+\mathcal{O}(N^2\log B)$
FAS	$\mathcal{O}(\sqrt{N})$	$\mathcal{O}(N^{2.5})$	$\mathcal{O}(N^3)$
$(P_{HSA,2})$	$\mathcal{O}(\sqrt{2\log_2(M)N})$	$\mathcal{O}(N^{2.5})$	$\mathcal{O}(\sqrt{2\log_2(M)N^3})$
$(P_{SA,2})$	$\mathcal{O}(\sqrt{MN})$	$\mathcal{O}(N^{2.5})$	$\mathcal{O}(\sqrt{MN^3})$
SD	1	$\mathcal{O}(\sqrt{M^N})$	$\mathcal{O}(\sqrt{M^N})$

Table 3.2: Computational cost with the interior point method.

where B is is the norm of the longest basis vector. In the case of determined MIMO systems, FAS achieves the same order of complexity compared to the MMSE-based methods.

3.4.3 Simulation results

In this section, we assume perfect channel state information and we evaluate the error rate achieved by the proposed detector based on FAS. We consider $n \times N$ MIMO systems, where N and n are the number of symbols to be recovered and the number of observations, respectively. $\boldsymbol{H}$ is a complex valued generic random matrix of size $n \times N$. We transform it in a real-valued formulation $\underline{\boldsymbol{H}}$ with size $2n \times 2N$. The channel coefficients are independent and identically Gaussian distributed with zero mean and unit variance, and the data symbols belong to a finite QAM alphabet. We use the Matlab CVX toolbox again. The quadratic minimization problem FAS is solved by the Gurobi optimizer [42].

3.4.3.1 Comparison of the simulation results with the theoretical analysis

We first check that the simulated histogram of detection output is in accordance with the theoretical statistical distribution given in Theorem 3.4.2 and Lemma 4.2.1. Fig. 3.3, Fig. 3.4, Fig. 3.5 and Fig. 3.6 give the results for a 64×64 system ($n = N = 64$) and 16-QAM with SNR = 15dB and SNR=30dB and 64-QAM with SNR=20dB and SNR=35dB respectively. From these figures, we observe that the theoretical distribution (exact as well as approximate) coincides with the simulated histogram for both low and high SNRs and different modulation orders.

In Fig. 3.7, both simulated SER and theoretical symbol error probability approximations are plotted for $n = N = 32$ and 16-QAM. We observe that the approximation given by Eq. (3.22) coincides with the simulated SER. The other one given by Eq. (3.23) is slightly more optimistic than simulated SER for very low SNR values. These observations validate the theoretical analysis.

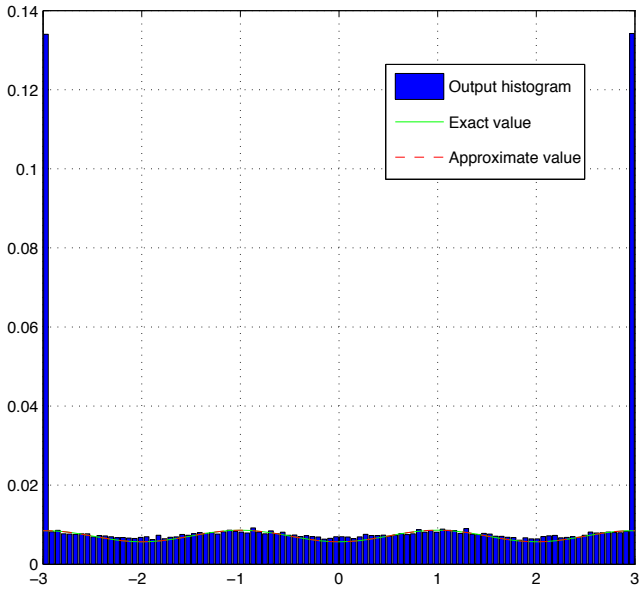

Figure 3.3: Output statistics for 64×64 systems with 16-QAM and SNR=15dB (low SNR).

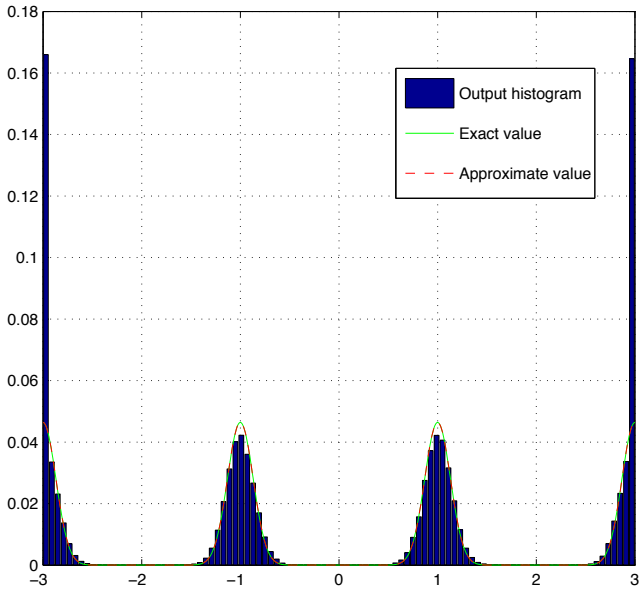

Figure 3.4: Output statistics for 64×64 systems with 16-QAM and SNR=30dB (high SNR).

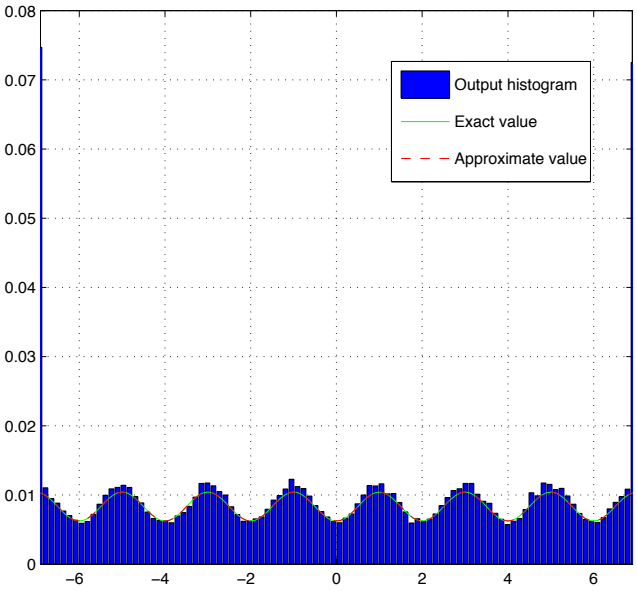

Figure 3.5: Output statistics for 64×64 systems with 64-QAM and SNR=20dB (low SNR).

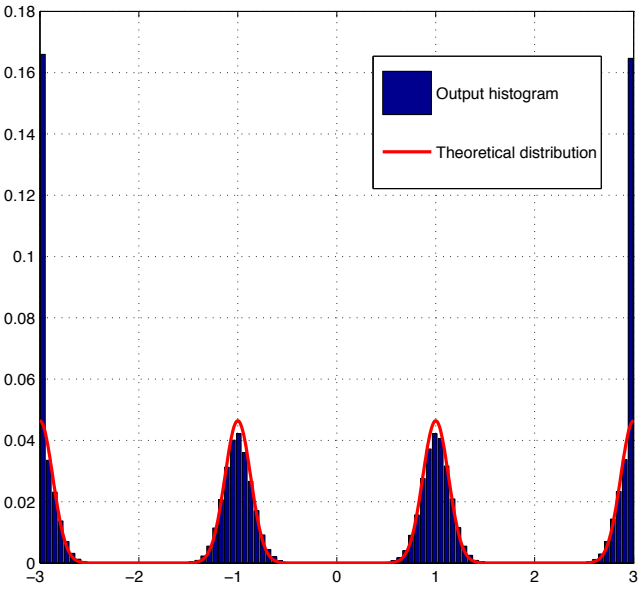

Figure 3.6: Output statistics for 64×64 systems with 64-QAM and SNR=35dB (high SNR).

3.4.3.2 Comparison with the SD (underdetermined case)

Fig. 3.8 shows the performance of the proposed scheme FAS for the underdetermined MIMO system of size 24×18 with 4-QAM. We observe that it achieves a BER under 10^{-3} for the SNR values higher than 20dB. Beyond 8dB, the SD outperforms the proposed scheme, e.g., at BER 10^{-3}, the gain is about 6.3dB. However, as the MIMO system dimensions increase, the SD computation cost will rapidly become too high to be implemented in practice, making the SD inadequate for large-scale MIMO applications.

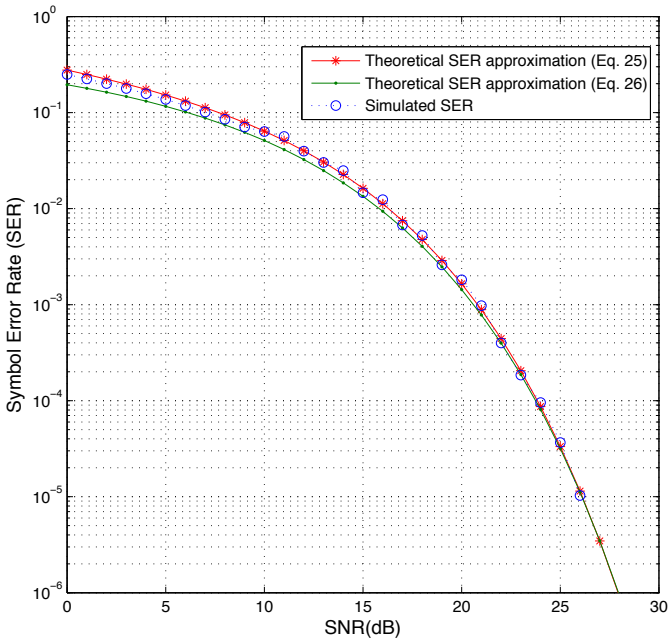

Figure 3.7: SER performance for 32×32 systems with 16-QAM.

3.4.3.3 Comparison with MMSE-based detection schemes (determined case)

Fig. 3.9 considers a determined system with $N = n = 64$ and 4-QAM. We compare simplicity-based detection FAS to MMSE, MMSE-SIC and MMSE-SIC-LR in terms of BER. We observe that the proposed detector outperforms both MMSE and MMSE-SIC over the whole SNR region and better exploits the receive diversity thanks to joint detection and box constraint effect which reduce the error propagation. At BER 10^{-3}, the proposed detector outperforms the MMSE by about 7 dB and the MMSE-SIC by about 1.5 dB. This gain increases with the growth of the SNR values to achieve about 2 dB at BER 10^{-4} compared to the MMSE-SIC. FAS outperforms the MMSE-SIC-LR for medium SNR values (a gain of 0.6 dB is observed for BER 10^{-3}). The advantage over the MMSE-SIC-LR decreases with

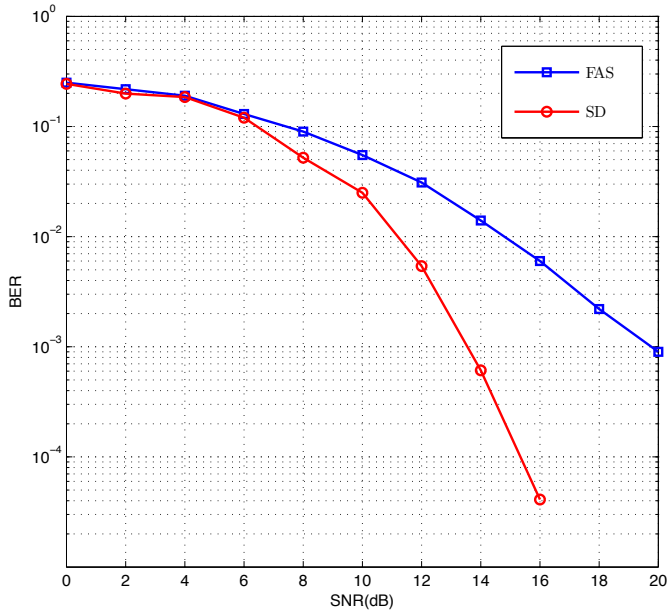

Figure 3.8: BER performance comparison for 24×18 systems ($\frac{n}{N} = 0.75$) with 4-QAM.

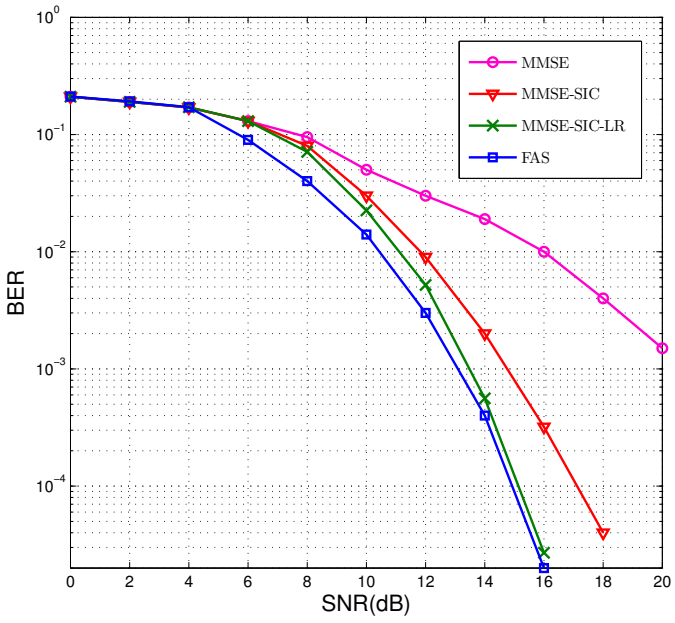

Figure 3.9: BER performance comparison for 64×64 systems and 4-QAM.

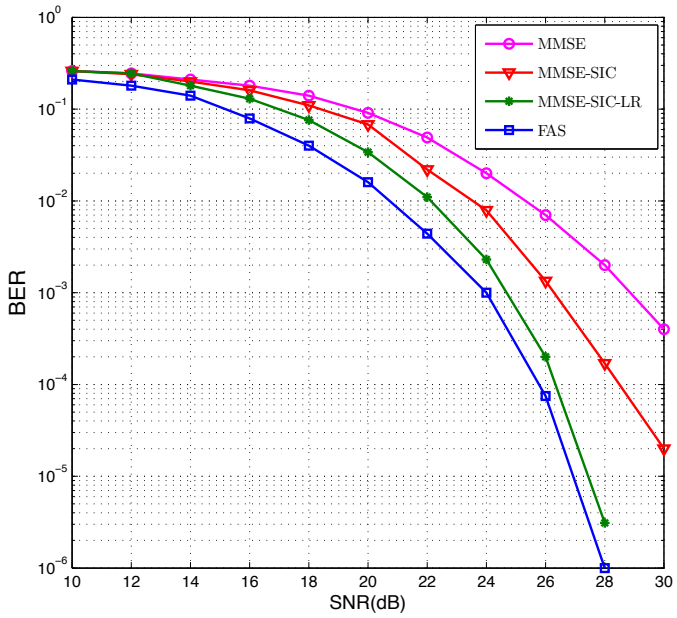

Figure 3.10: BER performance comparison in 64×64 systems and 16-QAM.

increasing SNR and both schemes perform the same for high SNR values. In Fig. 3.10, we also show that the gain of the proposed scheme FAS over MMSE-based schemes is maintained for higher order modulation. For 16-QAM, the proposed detector outperforms the MMSE-SIC by about 2 dB for BER 10^{-4}.

3.4.3.4 Comparison with local search-based detection schemes

Fig. 3.11 considers a determined system with $N = n = 32$ and 4-QAM. We compare simplicity-based detection FAS to LAS and RTS algorithms described previously in Section 2.7 in terms of BER. The initial solution for LAS and RTS algorithms is chosen as the MMSE output. We observe that for 4-QAM modulation, the proposed detector performs close to the LAS algorithm with slight loss for low SNR values and slight gain that starts from $SNR = 14$ dB and continues to increase slowly for high SNR values. However, we show that the RTS outperforms the proposed FAS and LAS for the whole SNR range. Increasing the modulation order (16-QAM, 64-QAM), we show that the FAS significantly outperforms the LAS algorithm for medium and high SNR values and becomes more efficient than RTS for high SNR. At BER 10^{-3}, the RTS outperforms the proposed detector FAS by about 2.5 dB. This gain increases with the growth of the SNR values to achieve about 3 dB at BER 10^{-4} compared to FAS for 4-QAM. In the case of 16-QAM, we show that for BER 10^{-2} the RTS still outperform the FAS by about 3dB, this gain decreases until vanishing at $SNR = 25$ dB for BER 10^{-3}. From $SNR = 25$ dB, the FAS outperforms the RTS algorithm and the gain increases when SNR increases. The

same holds for the 64-QAM with inflection point at $SNR = 37$ dB.

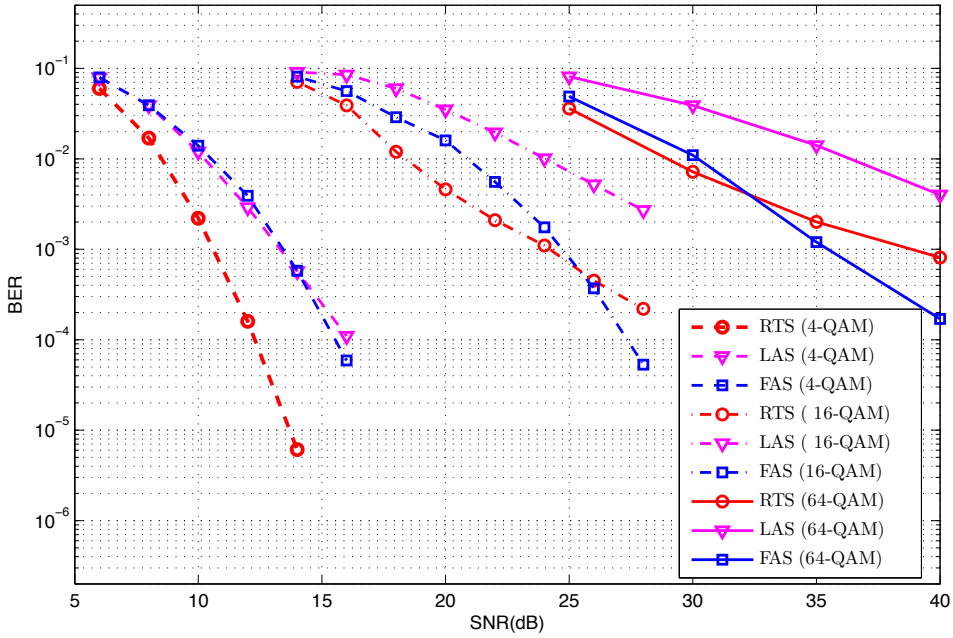

Figure 3.11: BER performance comparison in 32×32 systems with 4-QAM,16-QAM and 64-QAM.

3.5 Conclusion

This chapter focused on finite-alphabet source signal recovery in large-scale MIMO systems. We first proposed a simplicity-based ℓ_1-minimization combined with box constraints to solve the noise-free case. For the proposed criterion, we investigated the necessary condition of uniqueness and existence of a solution which is given by $\frac{n}{N} > \frac{p-1}{p}$ (see Statement (ii) of Theorem 3.3.1). This condition covers the determined case and partially the underdetermined case. Compared to previous existing sparsity-based techniques, we obtained a sufficient computation cost reduction with recovery success rate preservation. Simulation results corroborated the theoretical analysis. By exploiting the necessary condition of successful recovery on the problem parameters, we studied performance of the proposed criterion in the case of large-scale MIMO systems. The low-complexity resulting algorithm is well-adapted to such applications and its computation cost doesn't depend on the constellation size. Compared to LAS and RTS algorithms, the proposed FAS outperforms LAS in all studied cases and surpasses RTS below a BER threshold which increases as the modulation order gets higher. The theoretical distribution of the detector output was then validated through simulations. In next chapter, the analytical results will

be used to define an iterative shadow area-based detection to further improve the performance and to define a turbo-like iterative receiver to take into account an outer FEC code.

Chapter 4

Simplicity-based Iterative receivers for large-scale MIMO systems

Contents

4.1 Introduction

In this chapter, we consider large-scale MIMO detection and we define iterative receivers which use the simplicity-based detection algorithm (FAS) proposed in previous Chapter 3. In a first part, we focus on uncoded large-scale MIMO systems. We propose a novel successive interference cancellation algorithm with an iterative processing based on the shadow area principle and we optimize its parameters by exploiting the theoretical analysis of the detector output. In a second part, we assume FEC-encoded large-scale MIMO systems and our purpose is the definition of turbo-like iterative receivers. We propose an iterative receiver based on a ML-like detection whose restricted candidate subset is defined by the FAS detection output. We also introduce another receiver based on FAS algorithm whose criterion is penalized with the mean absolute error function.

[0]This chapter was partially proposed for publication in *IEEE Transactions on Communications*: Z. Hajji, and K. Amis, and A. Aïssa-El-Bey, "Iterative receivers for large-scale MIMO systems with finite-alphabet simplicity-based detection". (Major revision)

This chapter is organized as follows. Section 4.2 deals with the iterative detection problem in the uncoded case solved thanks to the shadow area principle applied with FAS algorithm. Section 4.3 focuses on the design of turbo-like iterative receivers based on FAS detection. Finally, Section 4.4 concludes the chapter.

4.2 Iterative Detection Based on the Shadow area principle

In this section, our purpose is to improve the FAS detection performance by including it within an iterative detection procedure. To that purpose, we consider shadow area constraints (SAC) used in [43] to limit error propagation [44, 45] in successive interference cancellation (SIC) schemes. Contrary to usual SIC, multiple feedback SIC with shadow area constraints (MF-SIC-SAC) feeds back more than one constellation point to the IC. Symbols are selected according to their belonging to a shadow area or not. In [43], the parameter which defines shadow areas is fixed empirically. Herein, we propose to apply a similar approach to FAS detection and to exploit the theoretical distribution of its output (Theorem. 3.4.2) to fix the optimal parameter that limits the shadow areas.

4.2.1 Shadow area and detection reliability

In the detection method described in previous section, all sources are detected at once and some decisions may be less reliable than others. In this section, we propose a reliability measure based on the shadow area principle [46–48] that exploits the output statistics reminded in Section 3.4. We first define the *centers* as the elements of $\mathcal{F}$. The principle is to take decision on components $\underline{x}_k$ such that $\hat{\underline{x}}_k$ is close enough to one center and cancel their contribution in the observation $\underline{\boldsymbol{y}}$ so as to proceed a novel detection iteration. To do so, we propose to take into account the reliabilities of the output $\hat{\underline{x}}_k$. According to Theorem. 3.4.2, the distribution of $\hat{\underline{x}}_k$ given $\underline{x}_k = \alpha_i$ has a Gaussian shape centered on α_i and moving away from the center makes the symbol less reliable. From this observation, we define shadow areas as intervals whose middle isn't a *center* and whose width depends on a threshold to be fixed hereinafter. $\hat{\underline{x}}_k$ is considered either as unreliable when it falls in a shadow area, or as reliable otherwise. We take decisions on reliable $\hat{\underline{x}}_k$, cancel their contribution from $\underline{\boldsymbol{y}}$ and proceed another detection. Adjacent to shadow areas, the high-reliability intervals are defined as intervals of length 2η and are centered on the different symbols of $\mathcal{F}$. Let us denote by $\mathcal{A}$ the set of indices k such that $\hat{\underline{x}}_k$ is considered as reliable. The decision on $\hat{\underline{x}}_k, k \in \mathcal{A}$ is taken as the nearest symbol value in $\mathcal{F}$. We denote by $\tilde{\underline{\boldsymbol{x}}}_{\mathcal{A}}$ the resulting decision vector. The equivalent notations for unreliable elements (falling in shadow areas) are respectively $\overline{\mathcal{A}}$ for the set of indices and v_N for its cardinality. The observation after interference cancellation is denoted by $\tilde{\underline{\boldsymbol{y}}}$ and equals

$$\begin{aligned} \underline{\tilde{\boldsymbol{y}}} &= \underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}_{\mathcal{A}}\underline{\tilde{\boldsymbol{x}}}_{\mathcal{A}} \\ &= \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}\underline{\boldsymbol{x}}_{\overline{\mathcal{A}}} + \tilde{\boldsymbol{\zeta}}, \end{aligned} \tag{4.1}$$

where $\tilde{\boldsymbol{\zeta}} = \underline{\boldsymbol{H}}_{\mathcal{A}}(\underline{\boldsymbol{x}}_{\mathcal{A}} - \underline{\tilde{\boldsymbol{x}}}_{\mathcal{A}}) + \boldsymbol{\zeta}$. The task is to estimate the vector $\underline{\boldsymbol{x}}_{\overline{\mathcal{A}}}$ which can be recovered by the following problem:

$$\arg\min_{\tilde{\boldsymbol{r}}} \|\underline{\tilde{\boldsymbol{y}}} - \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}\tilde{\boldsymbol{B}}_{\alpha}\tilde{\boldsymbol{r}}\|_2 \quad \text{subject to} \quad \tilde{\boldsymbol{B}}_1\tilde{\boldsymbol{r}} = \mathbf{1}_{v_N} \quad \text{and} \quad \tilde{\boldsymbol{r}} \geq 0. \tag{4.2}$$

where $\tilde{\boldsymbol{B}}_{\alpha} = \boldsymbol{I}_{v_N} \otimes [\alpha_1, \alpha_p]$, $\tilde{\boldsymbol{B}}_1 = \boldsymbol{I}_{v_N} \otimes \mathbf{1}_2^T$ and $\tilde{\boldsymbol{r}} \in [0,1]^{2v_N}$.

Algorithm 2 Shadow Area Constrained (SAC) - FAS detection

1: Input: $\boldsymbol{H}$, $\boldsymbol{y}$
2: $\boldsymbol{r} = \arg\min_{\boldsymbol{r}} \|\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}\boldsymbol{B}_{\alpha}\boldsymbol{r}\|_2 \quad \text{subject to} \quad \boldsymbol{B}_1\boldsymbol{r} = \mathbf{1}_{2N} \quad \text{and} \quad \boldsymbol{r} \geq 0.$
3: Compute $\underline{\hat{\boldsymbol{x}}} = \boldsymbol{B}_{\alpha}\,\boldsymbol{r}$.
4: Define $\mathcal{A} = \left\{ k \mid \min_{\alpha_i \in \mathcal{F}} |\underline{\hat{x}}_k - \alpha_i| \leq \eta,\ k \in \{1, \ldots, 2N\} \right\}$, $v_N = \text{card}(\overline{\mathcal{A}})$.
5: Compute $\underline{\tilde{\boldsymbol{x}}}_{\mathcal{A}}$ by $\underline{\tilde{x}}_k = \arg\min_{\alpha_i \in \mathcal{F}} |\underline{\hat{x}}_k - \alpha_i|$, $k \in \mathcal{A}$ and $\underline{\tilde{\boldsymbol{y}}} = \underline{\boldsymbol{y}} - \underline{\boldsymbol{H}}_{\mathcal{A}}\underline{\tilde{\boldsymbol{x}}}_{\mathcal{A}}$.
6: $\tilde{\boldsymbol{r}} = \arg\min_{\tilde{\boldsymbol{r}}} \|\underline{\tilde{\boldsymbol{y}}} - \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}\tilde{\boldsymbol{B}}_{\alpha}\tilde{\boldsymbol{r}}\|_2 \quad \text{subject to} \quad \tilde{\boldsymbol{B}}_1\tilde{\boldsymbol{r}} = \mathbf{1}_{v_N} \quad \text{and} \quad \tilde{\boldsymbol{r}} \geq 0.$
7: Compute $\underline{\tilde{\boldsymbol{x}}}_{\overline{\mathcal{A}}} = \tilde{\boldsymbol{B}}_{\alpha}\tilde{\boldsymbol{r}}$.
8: Output: $\underline{\tilde{\boldsymbol{x}}}$. =0

The shadow area constrained (SAC)- FAS detection procedure is detailed in Algorithm 2. The performance of the proposed iterative procedure highly depends on the choice of the parameter η, which needs optimization. We chose to use the error probability as an optimization criterion. The error probability is a monotonically increasing function of the variance of the components of the detector output. Then, we propose to optimize the parameter η so as to minimize the variance $\sigma^2_{\underline{\tilde{\boldsymbol{x}}}}$. Theorem 4.2.1 provides an approximation of $\sigma^2_{\underline{\tilde{\boldsymbol{x}}}}$.

Theorem 4.2.1 ***Variance of the iterative detector-output***
Let $\underline{\tilde{\boldsymbol{x}}}$ be the output of Algorithm 2 for a given $\eta \in \mathbb{R}^+$. Then, the variance of its components can be approximated by:

$$\sigma^2_{\underline{\tilde{\boldsymbol{x}}}}(\eta) = (1 - Z_\eta)\frac{2\,n\,\sigma^2_{\tilde{\zeta}}(\eta)}{2n - 2N(1 - Z_\eta) - 1} + Z_\eta Y_\eta. \tag{4.3}$$

where Z_η is the probability that $\hat{\underline{x}}_k$ is reliable and is equal to

$$\begin{aligned} Z_\eta &= \Pr(k \in \mathcal{A}) = \sum_{i=1}^{p} \Pr(|\hat{\underline{x}}_k - \alpha_i| < \eta) && (4.4)\\ &= \sum_{\ell=0}^{p-1} \frac{1}{p} \left((p-\ell)\mathrm{erfc}\left(\frac{\ell\Delta - \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) - (p-\ell-1)\mathrm{erfc}\left(\frac{\ell\Delta + \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right)\right) && (4.5)\\ &+ \frac{1}{2}\left(\mathrm{erfc}\left(\frac{\eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) - \mathrm{erfc}\left(\frac{-\eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right)\right). \end{aligned}$$

with $\Delta = \frac{\alpha_p - \alpha_1}{p-1}$. Y_η is the variance of the components of the vector $\tilde{\underline{\boldsymbol{x}}}_{\mathcal{A}}$ given by

$$\begin{aligned} Y_\eta &= \mathbb{E}[\tilde{\underline{x}}_k^2 \mid k \in \mathcal{A}] - \mathbb{E}[\tilde{\underline{x}}_k \mid k \in \mathcal{A}]^2 && (4.6)\\ &= \frac{\Delta^2}{Z_\eta} \sum_{\ell=1}^{p-1} \frac{\ell^2}{p} \left((p-\ell)\mathrm{erfc}\left(\frac{\ell\Delta - \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) - (p-\ell-1)\mathrm{erfc}\left(\frac{\ell\Delta + \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right)\right). \end{aligned}$$

and $\sigma^2_{\tilde{\zeta}}(\eta)$ is the variance of the components of the vector $\tilde{\zeta}$:

$$\sigma^2_{\tilde{\zeta}}(\eta) = \frac{1}{2n} Y_\eta + \sigma^2. \tag{4.7}$$

The proof of equation (4.5) and (4.6) is provided in the Appendix.

[Proof of Theorem 4.2.1] Let $\tilde{\underline{\boldsymbol{x}}}$ be the output of Algorithm 2 for a given $\eta \in \mathbb{R}^+$. Then, the variance of its components is given by:

$$\sigma^2_{\tilde{\underline{\boldsymbol{x}}}}(\eta) = \mathrm{var}(\tilde{\underline{x}}_k) = \mathrm{var}(\tilde{\underline{x}}_k \mid k \in \mathcal{A}) \Pr(k \in \mathcal{A}) + \mathrm{var}(\tilde{\underline{x}}_k \mid k \in \overline{\mathcal{A}}) \Pr(k \in \overline{\mathcal{A}})$$

$$\begin{aligned} Z_\eta &= \Pr(k \in \mathcal{A}) \quad \text{and} \quad 1 - Z_\eta = \Pr(k \in \overline{\mathcal{A}}) && (4.8)\\ Y_\eta &= \mathrm{var}(\tilde{\underline{x}}_k \mid k \in \mathcal{A}) \quad \text{for any} \quad k \in \mathcal{A}. && (4.9) \end{aligned}$$

To compute the variance $\mathrm{var}(\tilde{\underline{x}}_k \mid k \in \overline{\mathcal{A}})$, let us study the covariance matrix $\boldsymbol{\Sigma}_{\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}}$ by exploiting the fact that the number of elements of $\overline{\mathcal{A}}$ denoted by v_N is a random variable independent from $\tilde{\underline{\boldsymbol{x}}}$. Therefore, $\boldsymbol{\Sigma}_{\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}}$ is given by:

$$\boldsymbol{\Sigma}_{\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}} = \mathbb{E}\left[\mathbb{E}[(\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}} - \mathbb{E}[\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}])(\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}} - \mathbb{E}[\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}])^T | v_N = k]\right]. \tag{4.10}$$

By assuming that the vector $\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}$ can be estimated by

$$\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}} = (\underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}})^{-1} \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \tilde{\underline{\boldsymbol{y}}} = \underline{\boldsymbol{x}}_{\overline{\mathcal{A}}} + (\underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}})^{-1} \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \tilde{\boldsymbol{\zeta}}, \tag{4.11}$$

we can compute the covariance matrix as

$$\boldsymbol{\Sigma}_{\tilde{\underline{\boldsymbol{x}}}_{\overline{\mathcal{A}}}} = \sigma^2_{\tilde{\zeta}}(\eta)\, \mathbb{E}[(\underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}})^{-1} | v_N = k] = \sigma^2_{\tilde{\zeta}}(\eta) \frac{2n}{2n-k-1} \boldsymbol{I}_k, \tag{4.12}$$

where we have exploited that, given $v_N = k$, the matrix $(\underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}})^{-1}$ follows an inverse Wishart distribution and then $\mathbb{E}[(\underline{\boldsymbol{H}}_{\overline{\mathcal{A}}}^T \underline{\boldsymbol{H}}_{\overline{\mathcal{A}}})^{-1} | v_N = k] = \frac{2n}{2n-k-1} \boldsymbol{I}_k$ (see [40]). The distribution of v_N is provided by following Proposition 4.2.1.

Proposition 4.2.1 *The number of elements of the set $\overline{\mathcal{A}}$ follows the binomial distribution with parameters $2N$ and $(1 - Z_\eta)$:*

$$v_N = \mathrm{Card}(\overline{\mathcal{A}}) \sim \mathcal{B}(2N, 1 - Z_\eta).$$

From Proposition 4.2.1, we observe that the probability of event $v_N \geq 2n-1$ is not significant and these events will thus be neglected in the computation of $\boldsymbol{\Sigma}_{\underline{\tilde{\boldsymbol{x}}}_{\overline{\mathcal{A}}}}$. We then obtain

$$\boldsymbol{\Sigma}_{\underline{\tilde{\boldsymbol{x}}}_{\overline{\mathcal{A}}}} = \mathrm{var}(\underline{\tilde{x}}_k \mid k \in \overline{\mathcal{A}})\, \boldsymbol{I}_{v_N}, \tag{4.13}$$

with

$$\begin{aligned} \mathrm{var}(\underline{\tilde{x}}_k \mid k \in \overline{\mathcal{A}}) &= \sum_{k=0}^{2n-2} \Pr\left(v_N = k\right) \frac{2n\,\sigma^2_{\tilde{\zeta}}(\eta)}{2n-k-1} \\ &= \sum_{k=0}^{2n-2} \binom{2N}{k} Z_\eta^{2N-k} \left(1 - Z_\eta\right)^k \frac{2n\,\sigma^2_{\tilde{\zeta}}(\eta)}{2n-k-1}. \end{aligned}$$

Lemma 4.2.1 ***Variance approximation***
The variance of the components of the vector $\underline{\tilde{\boldsymbol{x}}}_{\overline{\mathcal{A}}}$ can be approximated for $n > N\left(1 - Z_\eta\right) + 1$ as

$$\mathrm{var}(\underline{\tilde{x}}_k \mid k \in \overline{\mathcal{A}}) \approx \frac{2n\,\sigma^2_{\tilde{\zeta}}(\eta)}{2n - 2N\left(1 - Z_\eta\right) - 1}. \tag{4.14}$$

In the simulation results section, the proposed iterative scheme will be mentioned as FAS and FAS-SAC for the first and second iterations respectively.

4.2.2 Simulation results

In this section, we evaluate the performance of the proposed FAS-SAC detection for a QAM constellation with different modulation orders. We also check the validity of the theoretical analysis and we optimize the parameters through simulations.

In Fig. 4.1, the variance of the detector output is plotted as a function of η for different SNR values, $N = 32, n = 32$ and 4-QAM. The parameter η should be chosen so as to get the minimum value of the variance.

In Fig. 4.2, we have plotted the BER after first (FAS) and second (FAS-SAC) iteration of proposed detection compared to RTS and LAS algorithms for $N = 64$, $n = 64$ and different M-QAM ($M = p^2 = 4, 16$, and 64). We observe that the proposed FAS-SAC detection improves the performance of the FAS algorithm at all SNR values and for all QAM modulations. For instance, FAS-SAC detection achieves a gain of around 2dB to 3dB at 10^{-3} BER.

We also show that the FAS-SAC better exploits the receive diversity than the LAS detector and it achieves a gain that gets higher as the BER decreases or M increases. For 4-QAM, the FAS-SAC outperforms the LAS by 2.2dB at 10^{-3} BER, the gain increases when M increases to achieve 7dB at 10^{-2} BER (64-QAM).

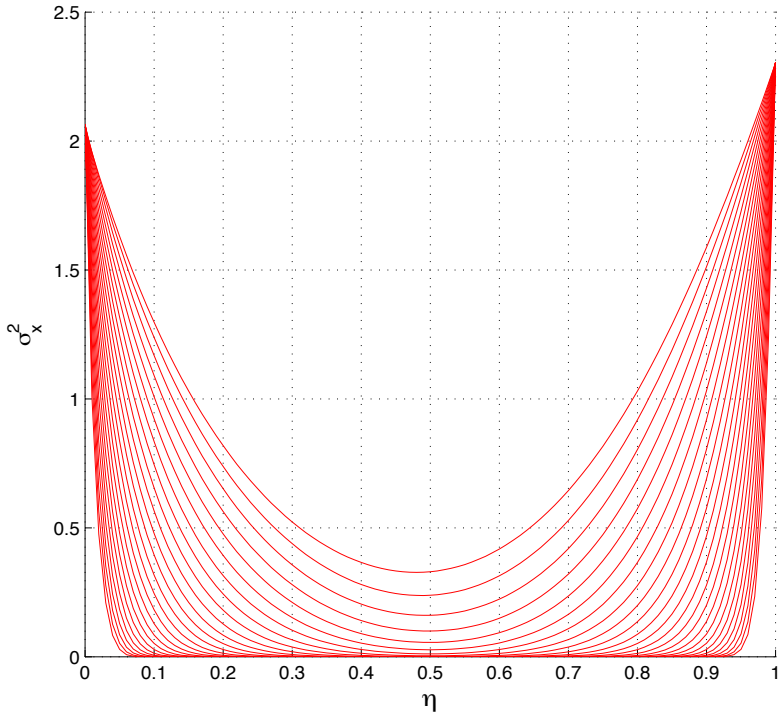

Figure 4.1: FAS output variance variation in function of the parameter η for $SNR =$ 15 to 30dB (up-to-down) and 16-QAM.

As for RTS algorithm, we observe that it outperforms the FAS by 1dB at 10^{-2} BER for 4-QAM. As the modulation order increases the FAS gets better than RTS from a given BER value (5.10^{-4} BER for 16-QAM, 4.10^{-3} BER for 64-QAM) with a flattering effect on the RTS performance curve. The FAS-SAC performs colse to the RTS for 4-QAM and gets beller than RTS below 3.10^{-3} BER for 16-QAM and 2.10^{-2} BER for 64-QAM.

In Fig. 4.3 and 4.4, we consider underdetermined systems with 4-QAM and 16-QAM respectively. We observe that proposed FAS-SAC algorithm performs remarkably even with underdetermined configurations. For instance, at BER 10^{-4}, the gains of FAS-SAC over FAS vary between 1 dB and 2 dB for 4-QAM and between 2 dB and 3.8 dB for 16-QAM.

4.2.3 Complexity Analysis

Table 4.1 summarizes the complexity order of proposed FAS-SAC algorithm compared to standard FAS algorithm. We observe that they have the same order of complexity.

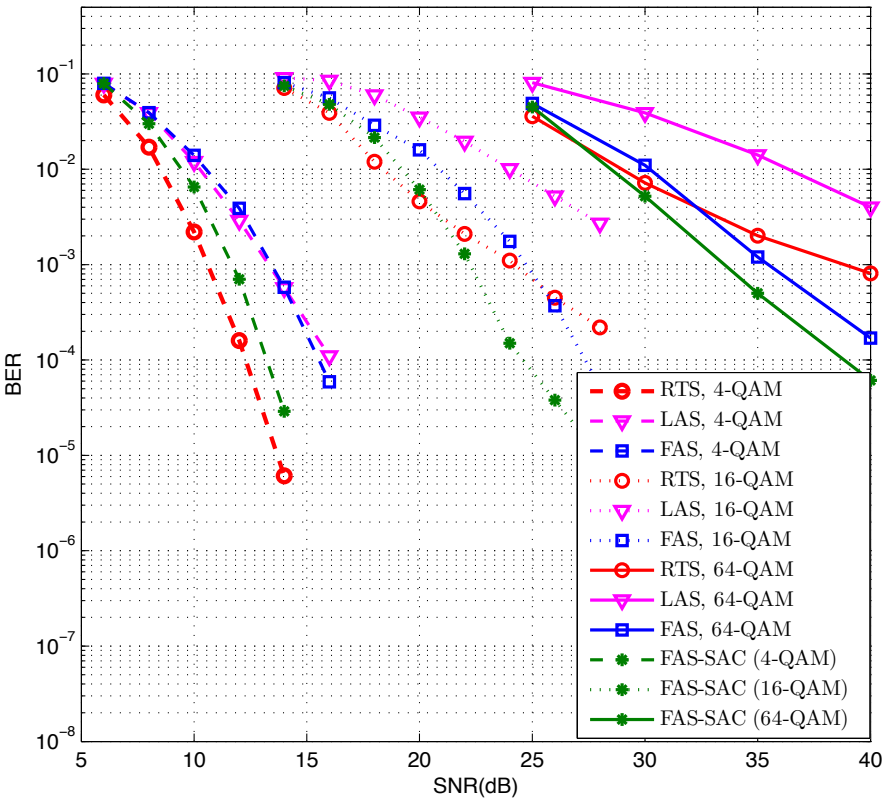

Figure 4.2: BER performance comparison of FAS, FAS-SAC detection and local search-based algorithms for $N = n = 32$ and 4-QAM, 16-QAM and 64-QAM

4.3 Proposed turbo detection scheme

In this section, we focus on FEC-coded large-scale MIMO systems. Our goal is the design of an iterative receiver consisting of a detector based on the FAS algorithm and a soft-input soft-output FEC decoder. The best iterative receiver of the state-of-the-art includes a soft-input soft-output maximum-likelihood detection. It provides the FEC decoder with log-likelihood ratios (LLR) whose computation involves the consideration of all possible transmitted sequences, which makes its practical use limited to low-order modulations and low-dimensional systems. In this section, we first propose to use the FAS detection to reduce the set involved in the computation of log-likelihood ratios. Although decreased, the computaton cost of the resulting receiver keeps high in the case of high-order modulations. We then design a second iterative receiver, whose detection uses an optimized regularization of the FAS criterion. Compared to the first proposed scheme, the complexity of the second one is significantly lower at the cost of a contained performance loss.

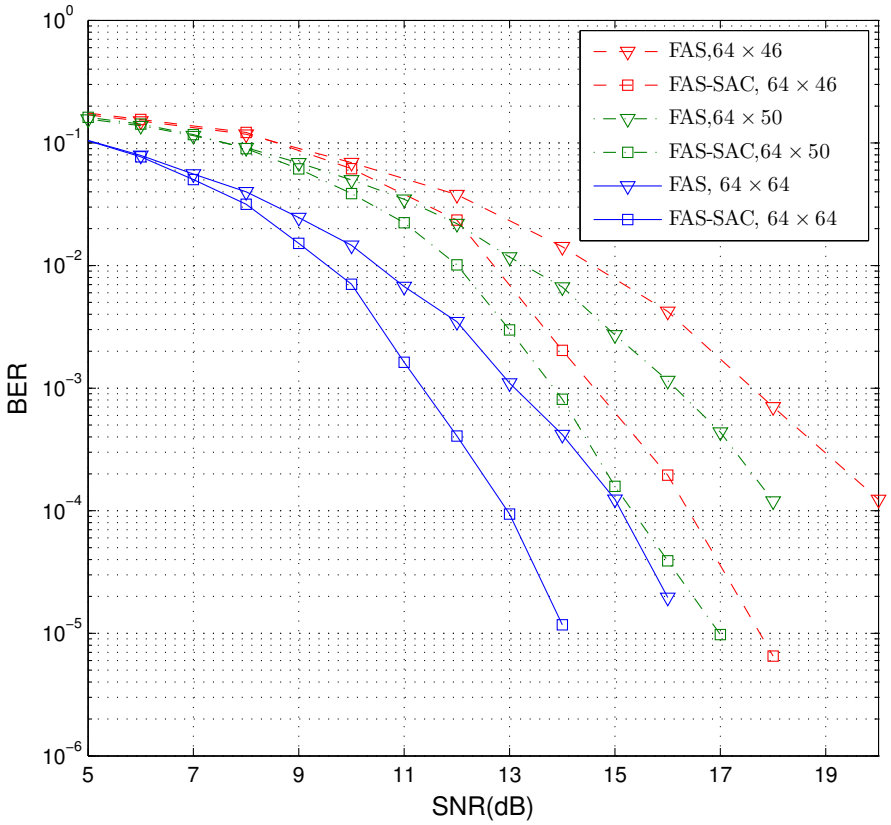

Figure 4.3: BER performance of FAS-SAC detection for $N = 64$, $n = 64, 50, 46$ and 4-QAM

4.3.1 Iterative receiver principle and notations

Let us first mention the assumptions regarding the transmitter. We consider that the binary stream is FEC-encoded, then randomly interleaved before being converted into QAM symbols and passed through a serial-to-parallel converter.
Let $m = \log_2(p)$ and let $\boldsymbol{c}$ be the coded and interleaved binary information sequence of length L. Let also ψ be the binary-to-symbol conversion defined by:

$$\psi : [c_{km}\ c_{km+1}\ \dots\ c_{(k+1)m-1}] \in \{0,1\}^m \mapsto \underline{x}_k \in \mathcal{F} \tag{4.15}$$

and $\boldsymbol{c}^{(j)} = \psi^{-1}(\alpha_j)$.

The receiver structure is depicted in Fig. 4.5. Λ_{in}^{dec} and Λ_{out}^{dec} stand for the soft FEC input and output LLR respectively. Both proposed iterative schemes differ from the detection box definition. We denote by Λ_{in}^{det} and Λ_{out}^{det} the detection input and output respectively. Λ_{in}^{det} is defined as the interleaving of the difference between Λ_{out}^{dec} and Λ_{in}^{dec} (extrinsic information).

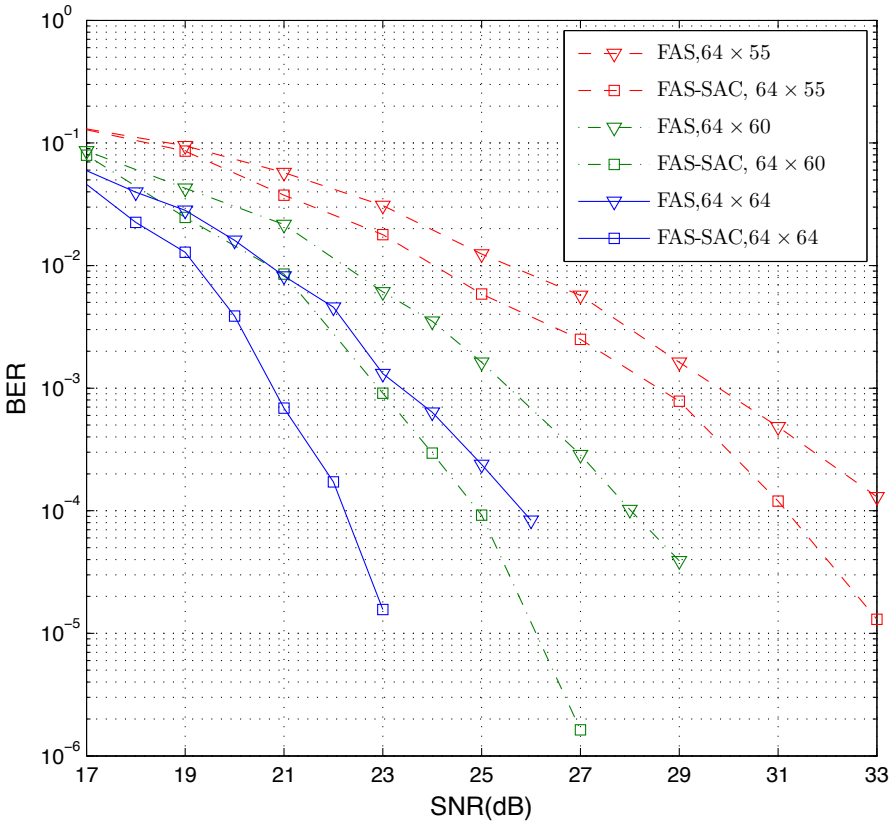

Figure 4.4: BER performance of FAS-SAC detection for $N = 64$ and $n = 64, 60$ and 50 and 16-QAM.

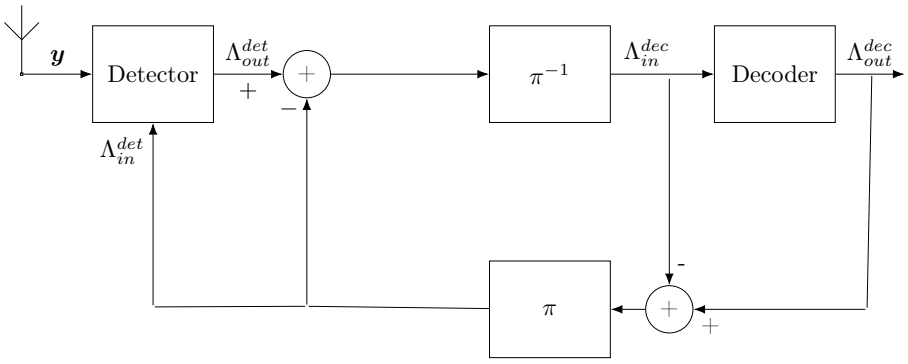

Figure 4.5: Iterative receiver scheme.

	Iteration number	*Computational cost per iteration*	*Total*
FAS	$\mathcal{O}(\sqrt{N})$	$\mathcal{O}(N^{2.5})$	$\mathcal{O}(N^3)$
FAS-SAC	$\mathcal{O}(\sqrt{N})$ $+\mathcal{O}(\sqrt{(1-Z_\eta)N})$	$\mathcal{O}(\sqrt{1-Z_\eta}N^{2.5})$	$\mathcal{O}((2-Z_\eta)N^3)$

Table 4.1: Computational cost with the interior point method.

4.3.2 FAS Maximum Likelihood like iterative receiver (FAS-ML)

Usual turbo-detection schemes are based on a ML detection followed by a decoder [49]. In such a scheme the detection output Λ_{out}^{det} is defined as follows:

$$\begin{aligned}
\Lambda_{out}^{det}(k) &= \log\left(\frac{\Pr(c_k=1|\underline{\boldsymbol{y}})}{\Pr(c_k=0|\underline{\boldsymbol{y}})}\right) && (4.16)\\
&= \log\left(\frac{\sum_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,+1}} f(\underline{\boldsymbol{y}}|\underline{\boldsymbol{x}})\Pr\left(\underline{\boldsymbol{x}}|\Lambda_{in}^{det}\right)}{\sum_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,-1}} f(\underline{\boldsymbol{y}}|\underline{\boldsymbol{x}})\Pr\left(\underline{\boldsymbol{x}}|\Lambda_{in}^{det}\right)}\right) && (4.17)\\
&= \log\left(\frac{\sum_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,+1}} \exp\left(-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right)\exp\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}\right)}{\sum_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,-1}} \exp\left(-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right)\exp\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}\right)}\right)\\
&\approx \max_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,+1}}\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right)\\
&\quad -\max_{\underline{\boldsymbol{x}}\in\mathcal{X}_{k,-1}}\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right) && (4.18)
\end{aligned}$$

where $\tilde{\boldsymbol{c}}=2\boldsymbol{c}-1$ and $\mathcal{X}_{k,\epsilon}$ corresponds to the set of sequences $\underline{\boldsymbol{x}}$ such that $c_k=\epsilon$.

The complexity of such a detection increases exponentially with M and N. Therefore we propose to reduce the complexity by substituting a limited-size subset $\Xi_{k,\epsilon}$ for $\mathcal{X}_{k,\epsilon}$. For that purpose, we first run the FAS detection once and make a hard decision on its output $\hat{\underline{\boldsymbol{x}}}$. We denote by $\tilde{\underline{\boldsymbol{x}}}_{out}^{det}$ this hard decision output. Then we define the subset Ξ such that it includes $\tilde{\underline{\boldsymbol{x}}}_{out}^{det}$ and sequences $\underline{\boldsymbol{x}}$ which differ by one element from $\tilde{\underline{\boldsymbol{x}}}_{out}^{det}$. To limit the size of Ξ, we only take neighbors of $\tilde{\underline{\boldsymbol{x}}}_{out}^{det}$. More precisely, if $\tilde{\underline{\boldsymbol{x}}}_{out}^{det}$ and $\underline{\boldsymbol{x}}$ differ from their i-th element, then $\underline{x}_i$ is an adjacent symbol of $\tilde{\underline{x}}_{out,i}^{det}$ in $\mathcal{F}$. After the initialization step during which the FAS detection is carried out, an iterative process is applied alternating from a ML-like detection and a FEC decoder. The ML-like detection computes Λ_{out}^{det} from Λ_{in}^{det} and $\underline{\boldsymbol{y}}$ as follows:

$$\begin{aligned}
\Lambda_{out}^{det}(k) &\approx \max_{\underline{\boldsymbol{x}}\in\Xi_{k,+1}}\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right)\\
&\quad -\max_{\underline{\boldsymbol{x}}\in\Xi_{k,-1}}\left(\frac{\tilde{\boldsymbol{c}}\left(\Lambda_{in}^{det}\right)^T}{2}-\frac{||\underline{\boldsymbol{y}}-\boldsymbol{H}\underline{\boldsymbol{x}}||^2}{2\sigma^2}\right) && (4.19)
\end{aligned}$$

In the remaining of the PhD dissertation, we refer to the resulting iterative receiver

as FAS-ML. In the case of uniform square constellations and except for $M = 4$, each symbol has at most four neighbors, and thus the complexity of FAS-ML only depends on the length of $\boldsymbol{c}$.

4.3.3 FAS Mean Absolute Error-based iterative receiver (FAS-MAE)

To further reduce the receiver complexity, we propose a second receiver whose detection is based on a regularization of the FAS criterion. The receiver structure is detailed in Fig. 4.6. Compared to [50], two major differences can be highlighted. First, the FEC output is directly exploited without any preprocessing in order to preserve the information. Secondly, the regularization parameter is optimized and an analytical expression is given.

4.3.3.1 New detection criterion design

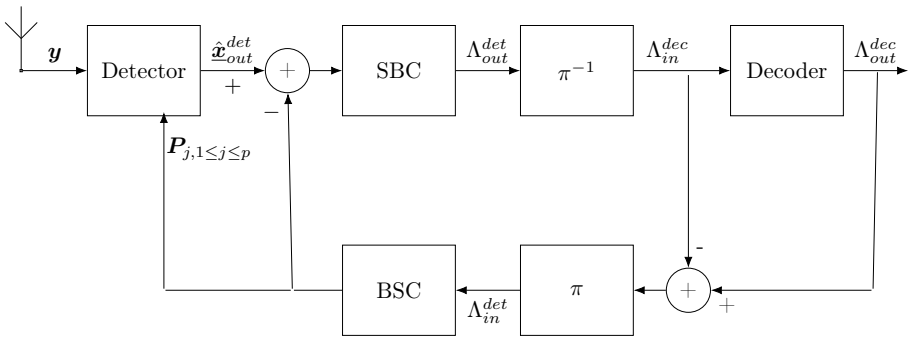

Figure 4.6: Iterative receiver scheme.

The first modification compared to [50] is the use of the Mean Absolute Error (MAE) computed from conditional probabilities $\Pr(\underline{x}_k = \alpha_j|\Lambda_{in}^{det})$. We denote this error by $\boldsymbol{\varepsilon}(\underline{\hat{\boldsymbol{x}}}, \underline{\boldsymbol{x}}|\Lambda_{in}^{det})$ and we define it by:

$$\boldsymbol{\varepsilon}(\underline{\hat{\boldsymbol{x}}}, \underline{\boldsymbol{x}}|\Lambda_{in}^{det}) = \sum_{j=1}^{p} \boldsymbol{P}_j^T \, |\boldsymbol{r} - \boldsymbol{d}_j|, \tag{4.20}$$

where $\boldsymbol{P}_j = \left[\Pr(\underline{x}_1 = \alpha_j|\Lambda_{in}^{det}) \quad \Pr(\underline{x}_2 = \alpha_j|\Lambda_{in}^{det}) \quad \dots \quad \Pr(\underline{x}_{2N} = \alpha_j|\Lambda_{in}^{det})\right]^T$ and $\boldsymbol{d}_j = \alpha_j \mathbf{1}_{2N}$.
Using Λ_{in}^{det} provided by the FEC decoder, we compute $\Pr\left(\underline{x}_k = \alpha_j|\Lambda_{in}^{det}\right)$ as follows:

$$\Pr\left(\underline{x}_k = \alpha_j|\Lambda_{in}^{det}\right) = \prod_{\substack{0\leq i\leq m-1 \\ c^{(j)}=\psi^{-1}(\alpha_j)}} \Pr\left(c_{km+i} = c_i^{(j)}|\Lambda_{in}^{det}\right),$$

with $\Pr\left(c_{km+i} = c_i^{(j)} | \Lambda_{in}^{det}\right) = \frac{\exp(u_{i,j} v_{k,i})}{\exp(v_{k,i}) + \exp(-v_{k,i})}$, $u_{i,j} = 2c_i^{(j)} - 1$ and $v_{k,i} = \frac{\Lambda_{in}^{det}(km+i)}{2}$.

The MAE is introduced as a regularization term in the FAS criterion to define the following optimization problem:

$$\underset{\boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}, \boldsymbol{r} \geq 0.}{\arg\min} ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}} \boldsymbol{B}_\alpha \boldsymbol{r}||_2 + \gamma \sum_{j=1}^{p} \boldsymbol{P}_j^T |\boldsymbol{r} - \boldsymbol{d}_j|, \tag{4.21}$$

where γ is a positive weight less than 1. The resulting iterative receiver is referred to as FAS-MAE in the remaining of the PhD report. On one hand, the regularization term $\boldsymbol{\varepsilon}(\underline{\hat{\boldsymbol{x}}}, \underline{\boldsymbol{x}} | \Lambda_{\text{in}}^{\text{det}})$ can be seen as a penalty, imposed to ensure that the detector output remains in the neighborhood of the decoder output. On the other hand, γ enables to regulate the contribution of the FEC extrinsic information and thereby to question the FEC decision if necessary. The idea is to ensure that the resulted vector $\boldsymbol{r}$ is sparse. We mention that imposing its sparsity takes into account the different probabilities delivered by the decoder.

4.3.3.2 Optimization of the regularization parameter

The performance of the proposed FAS-MAE detector highly depends on the choice of the regularization parameter. However, its optimization is difficult. It depends on many parameters among which the SNR value and the level of $\Pr(\underline{\boldsymbol{x}}_k = \alpha_j | \Lambda_{in}^{det})$ (either close to their bounds 0, 1 or not).

According to the proposed optimization criterion, the algorithm convergence is optimum when the cost function tends to 0, that is to say when the following condition is satisfied:

$$\frac{||\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}} \boldsymbol{B}_\alpha \boldsymbol{r}||_2}{\sum_{j=1}^{p} \boldsymbol{P}_j^T |\boldsymbol{B}_\alpha \boldsymbol{r} - \boldsymbol{d}_j|} \approx \gamma. \tag{4.22}$$

The analytical determination of γ from (4.22) is not possible as it requires the analytical distribution of the FEC output, which is not available. We propose two ways to optimize γ. The first one is empirical and uses pilot symbols. The second one gives an analytical expression for γ. In the simulations, the first one will be used as a benchmark for the second one and we will refer to it as FAS-MAE (genie).

The first optimization requires a pilot sequence $\underline{\hat{\boldsymbol{x}}}_{pilot}$. Pilot symbols are usually inserted within the data frame to help synchronization and parameter estimation. Their position as well as their value are perfectly known at the receiver. Assuming the transmission of the pilot sequence, we perform only one iteration (both detection and decoding) and we compute $\frac{||\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}} \boldsymbol{B}_\alpha \boldsymbol{r}||_2}{\sum_{j=1}^{p} \boldsymbol{P}_j^T |\boldsymbol{B}_\alpha \boldsymbol{r} - \boldsymbol{d}_j|}$ by considering the true values of $\boldsymbol{r}$ and the value of $\boldsymbol{P}_j$ delivered by the decoder. We then fix γ to the following ratio

$$\gamma_1 = \frac{\sigma_\zeta^2}{\sum_{j=1}^{p} \boldsymbol{P}_j^T |\underline{\hat{\boldsymbol{x}}}_{pilot} - \boldsymbol{d}_j|}. \tag{4.23}$$

Previous optimization method suffers from two drawbacks. First, it requires the use of pilots, yielding spectral efficiency loss and secondly, a detection step followed by a decoding step is carried out. The second method overcomes both of them by providing an analytical expression for γ. The problem criterion in (4.21) defines a ℓ_1-norm penalized least squares estimator similar to the one studied in [51]. Then, the second term of regularization in (4.21) can be seen as a weighted ℓ_1 term and we propose to fix γ as developed in [51]. It depends on the noise variance and on the system dimensions:

$$\gamma_2 = \sigma_\zeta \sqrt{\frac{\log N}{n}}. \tag{4.24}$$

4.3.3.3 Definition of the decoder input

In this part, we focus on the information exchange from the detector to the decoder. Contrary to FAS-ML, we will use the statistical distribution of the FAS detection established in [52].

Using the detector output $\underline{\hat{x}}_{\text{out}}^{\text{det}}$, the symbol to binary converter (SBC) computes the log likelihood ratio on the i-th bit associated to the k-th symbol, denoted by $\Lambda_{\text{out}}^{\text{det}}(km+i)$ and defined as:

$$\Lambda_{out}^{det}(km+i) \quad = \quad \log\left(\frac{\sum_{\alpha_j \in \mathcal{F}_{i,1}} f_{\underline{\hat{x}}_k|\underline{x}_k=\alpha_j}(\underline{\hat{x}}_{out,k}^{det}) \Pr(\underline{x}_k = \alpha_j | \Lambda_{in}^{det}))}{\sum_{\alpha_j \in \mathcal{F}_{i,0}} f_{\underline{\hat{x}}_k|\underline{x}_k=\alpha_j}(\underline{\hat{x}}_{out,k}^{det}) \Pr(\underline{x}_k = \alpha_j | \Lambda_{in}^{det}))}\right)$$

with $\mathcal{F}_{i,\epsilon} = \{\alpha \in \mathcal{F} | \boldsymbol{c} = \psi^{-1}(\alpha), c_i = \epsilon\}$.
Let us mention that an empirical study proved that the expression of $\sigma_{\underline{\hat{x}}}$ given by (3.16) keeps valid throughout the iterative process.

$f_{\underline{\hat{x}}_k|\underline{x}_k=\alpha_j}$ is given by Theorem. 3.4.2. In [50], we used a Gaussian approximation combined with the LogSumExp approximation [53] to avoid saturation precision problems of the floating point, especially for high SNR and after some iterations. Doing so, we degrade the information available for the symbol decisions which equal the alphabet bounds. In this chapter, we overcome the problem by proposing a new approximation that takes into account the hard decisions available at the FAS-MAE output and which we previously denoted $\underline{\tilde{x}}_{out,k}^{det}$. This LLR approximation is given

by:

$$\begin{aligned}
\Lambda_{out}^{det}(km+i) &\approx \max_{\alpha_j \in \mathcal{F}_{i,1}} \left(-\frac{(\hat{\underline{x}}_{out,k}^{det} - \alpha_j)^2}{2\sigma_{\underline{\hat{x}}}^2} + u_{i,j} v_{k,i} \right) \\
&\quad - \max_{\alpha_j \in \mathcal{F}_{i,0}} \left(-\frac{(\hat{\underline{x}}_{out,k}^{det} - \alpha_j)^2}{2\sigma_{\underline{\hat{x}}}^2}) + u_{i,j} v_{k,i} \right), \text{ if } \tilde{\underline{x}}_{out,k}^{det} \notin \{\alpha_1, \alpha_p\} \\
&\approx \log \sum_{\alpha_j \in \mathcal{F}_{i,1}} \left(\text{erfc} \left(\frac{\alpha_j - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \exp\left(u_{i,j} v_{k,i}\right) \right) \\
&\quad - \log \sum_{\alpha_j \in \mathcal{F}_{i,0}} \left(\text{erfc} \left(\frac{\alpha_j - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \exp\left(u_{i,j} v_{k,i}\right) \right), \text{ if } \tilde{\underline{x}}_{out,k}^{det} = \alpha_1 \\
&\approx \log \sum_{\alpha_j \in \mathcal{F}_{i,1}} \left(\text{erfc} \left(\frac{\alpha_p - \alpha_j}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \exp\left(u_{i,j} v_{k,i}\right) \right) \\
&\quad - \log \sum_{\alpha_j \in \mathcal{F}_{i,0}} \left(\text{erfc} \left(\frac{\alpha_p - \alpha_j}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \exp\left(u_{i,j} v_{k,i}\right) \right), \text{ if } \tilde{\underline{x}}_{out,k}^{det} = \alpha_p.
\end{aligned} \tag{4.25}$$

Performance were significantly improved thanks to this new approximation as will be shown in Section 4.3.4 dedicated to simulations.

4.3.4 Simulation results

In this section, we study the performance of the proposed FAS-ML and FAS-MAE iterative schemes. We also compare them to the Turbo-MMSE detector and to the iterative receiver introduced in [50].
The convolutional code (CC) polynomials in octal are $(13, 15)$ with a code rate equal to 0.5. The decoder uses the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm [54].

We will observe that as established in Chapter 3, FAS detection is perfectly adapted to underdetermined systems provided the recovery success condition is satisfied.

4.3.4.1 Comparison of FAS-MAE to FAS-SSE

FAS-MAE is an enhanced version of our proposed receiver detailed in [50], which will be referred to as FAS-SSE for soft symbol error in the simulations. The detection criterion taken into account in our previous work was

$$\underset{\boldsymbol{B}_1 \boldsymbol{r} = \mathbf{1}_{2N}, \boldsymbol{r} \geq 0.}{\arg\min} ||\underline{\boldsymbol{y}} - \underline{\boldsymbol{H}} \boldsymbol{B}_\alpha \boldsymbol{r}||_2 + \gamma ||\boldsymbol{B}_\alpha \boldsymbol{r} - \hat{\underline{\boldsymbol{x}}}_{in}^{det}||_2 \tag{4.26}$$

with soft symbol decision $\hat{\underline{\boldsymbol{x}}}_{in}^{det}$ computed as follows:

$$\hat{\underline{x}}_{in,k}^{det} = \sum_{\alpha_j \in \mathcal{F}} \alpha_j Pr\left(\underline{x}_k = \alpha_j | \Lambda_{in}^{det}\right), \tag{4.27}$$

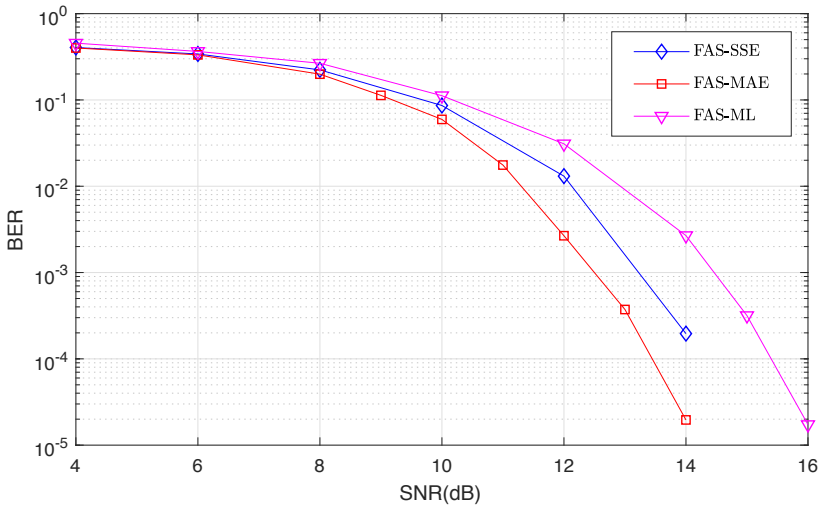

Figure 4.7: Comparison of FAS-based iterative receivers with $N = n = 64$ and coded 16-QAM.

and γ chosen empirically. Performance of FAS-MAE with γ_2 and FAS-SSE with empirically optimized γ are compared in Fig. 4.7 and Fig. 4.8 for $N = 64$, $n = 64, 50$ and 16-QAM. We observe the efficiency of both the new criterion and the optimization of γ as FAS-MAE outperforms FAS-SSE (gains of roughly 1.0 and 0.8 at BER=10^{-3} for $n = 50$ and $n = 64$ respectively). The gain slowly increases as SNR gets higher.

We now consider MIMO systems with $N = 64, L = 256$, 4-QAM modulation and $n = 64$, 50 and 40 in Fig. 4.9, 4.10 and 4.11 respectively. We have plotted the BER measured at the FEC decoder output after 6 iterations for FAS-ML, FAS-MAE and Turbo-MMSE receivers.

4.3.4.2 Optimization of the regularization parameter γ (Fig. 4.9, 4.10 and 4.11)

We observe that proposed FAS-MAE (genie) and FAS-MAE perform the same, which supports the choice of the analytical expression (4.23) used to fix the penalization parameter.

4.3.4.3 Comparison of FAS-MAE to FAS-ML

To compare FAS-MAE to FAS-ML, let us study the influence of the modulation order. We remind that in order to reduce the candidate subset, given a position in $\underline{\tilde{\boldsymbol{x}}}$, FAS-ML considers all candidates in the case of 4-QAM while it limits itself to adjacent neighbours in the case of higher order modulations. The consequence is that FAS-ML outperforms FAS-MAE in the case of 4-QAM while it achieves lower results in the case of 16-QAM. Whereas the gain of FAS-ML over FAS-MAE varies

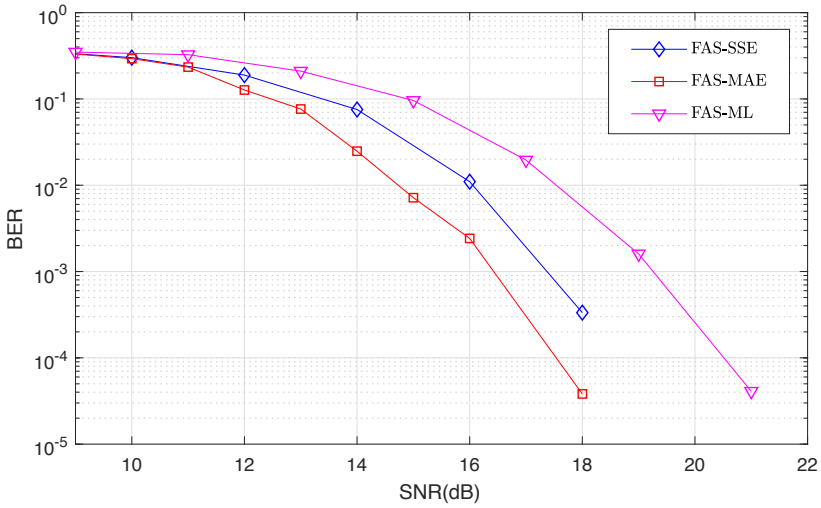

Figure 4.8: Comparison of FAS-based iterative receivers with $N = 64$ and $n = 50$ and coded 16-QAM.

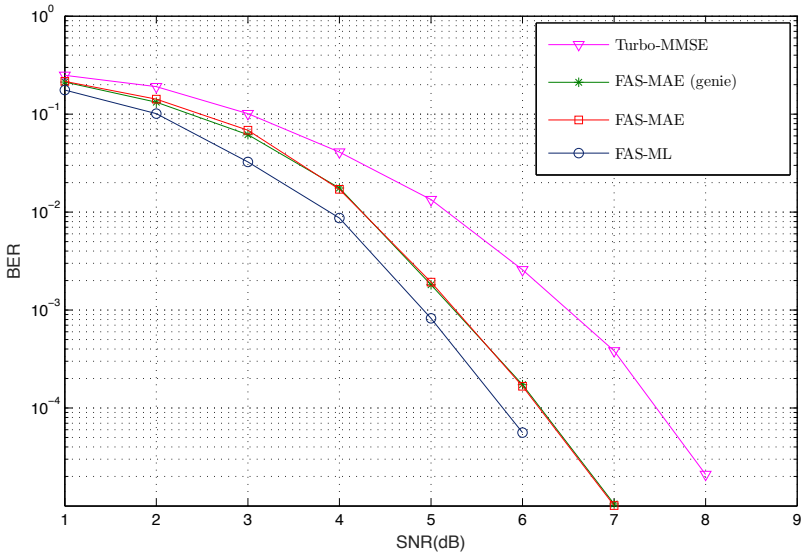

Figure 4.9: Comparison of FAS-MAE, FAS-ML and Turbo-MMSE with $N = 64$, $n = 64$ and coded 4-QAM.

between 0.2 and 0.5 dB at BER=10^{-4} depending on n for 4-QAM, we observe a degradation of FAS-ML compared to FAS-MAE of about 2 dB and 2.6 dB at BER=10^{-4} for $n = 64$ and $n = 50$ respectively.

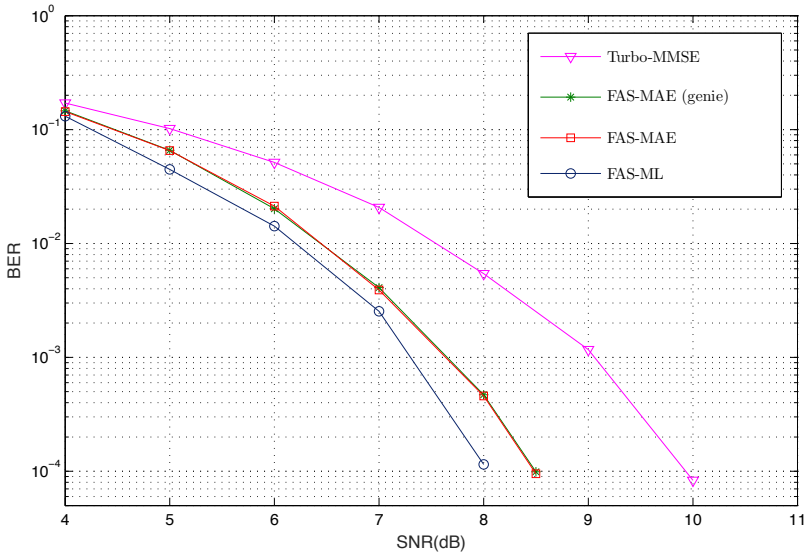

Figure 4.10: Comparison of FAS-MAE, FAS-ML and Turbo-MMSE with $N = 64$, $n = 50$ and coded 4-QAM.

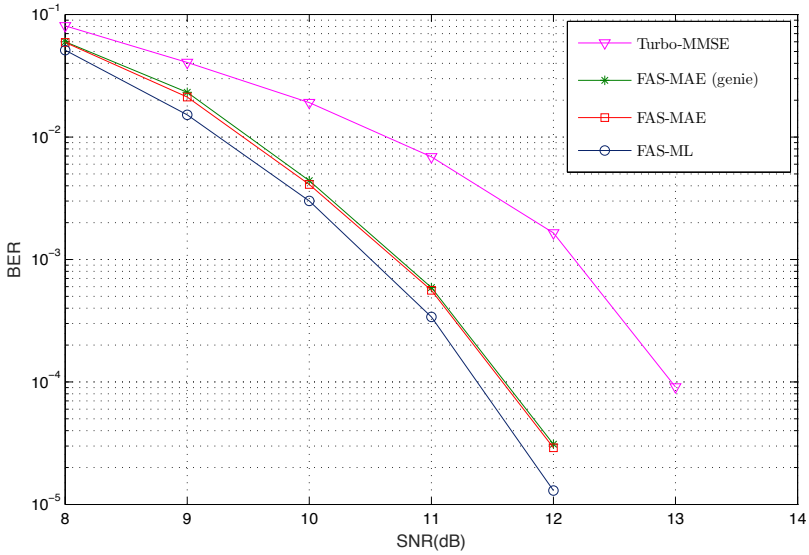

Figure 4.11: Comparison of FAS-MAE, FAS-ML and Turbo-MMSE with $N = 64$, $n = 40$ and coded 4-QAM.

4.3.4.4 Comparison of FAS-MAE and Turbo-MMSE (Fig. 4.9, 4.10 and 4.11)

In all cases, FAS-ML and FAS-MAE outperform the Turbo-MMSE detection. The gain is all the higher as the system is underdetermined. The gain of FAS-MAE over

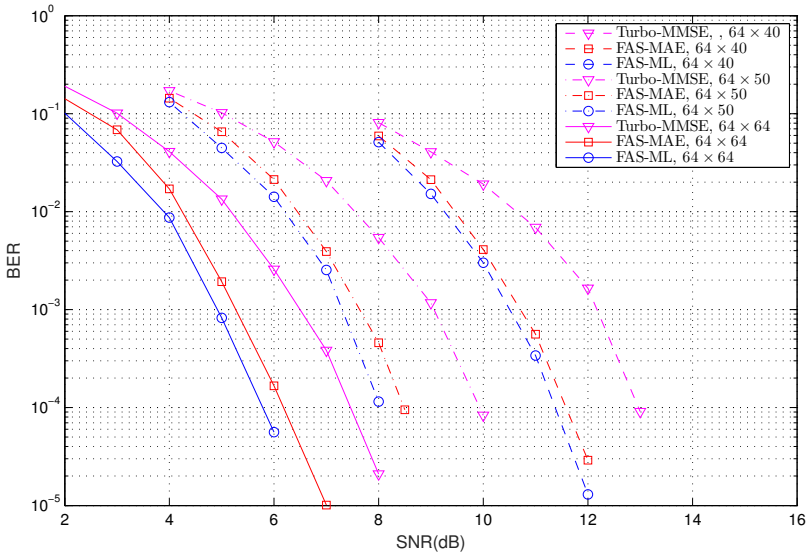

Figure 4.12: Comparison of FAS-MAE, FAS-ML and Turbo-MMSE with coded 4-QAM.

Turbo-MMSE equals about 1.25 dB for $n = 64$, 1.5 dB for $n = 50$ and 2dB for $n = 40$ at BER 10^{-4}.

Fig.4.12 gathers all configurations for the three receivers (FAS-MAE, FAS,ML, Turbo-MMSE). We observe that they achieve similar diversity orders and differ from coding gains.

4.4 Conclusion

This chapter focused on finite-alphabet iterative source recovery for large-scale MIMO systems either uncoded or coded. For uncoded case, we developed an iterative FAS algorithm which uses shadow area constraints with an optimized shadow area defining parameter. The simulation results showed that the proposed FAS-SAC algorithm significantly outperforms standard FAS, LAS and RTS algorithms in almost cases with the same order of computational complexity. Then, for FEC-encoded case, we introduced the FAS-ML receiver which reduces the complexity of ML detection by restricting the candidate subset from the FAS algorithm output. To further reduce the receiver complexity, we proposed FAS-MAE receiver whose detection is based on a regularization of the FAS criterion without any preprocessing of the FEC-decoder output and where its regularization parameter is analytically fixed. Simulations showed that both receivers outperform Turbo-MMSE in all cases and that FAS-MAE achieves better results (lower error rate and less complexity load) than FAS-ML for M-QAM with $M > 4$. Until now all the work done is based on a perfect knowledge of the channel matrix. Nevertheless, in the case of large-

scale MIMO systems, this scenario is difficult to have. Next chapter is dedicated to the impact evaluation of imperfect channel estimation and the design of efficient algorithms that deal with CSI estimation.

Chapter 5

Channel estimation in large-scale MIMO systems

Contents

5.1 Introduction

In previous chapters, the detection algorithms were presented under the assumption of perfect channel knowledge at the receiver. However, in practice, the channel gains are estimated at the receiver, either blindly/semi-blindly or through pilot training. In FDD systems, the estimated channel gains at the receiver are fed back to the transmitter for precoding. In TDD systems, where the channel reciprocity holds, the transmitter can estimate the channel and use it for precoding. Due to noise and

[0]This chapter was partially proposed for publication in *IEEE Transactions on Signal Processing*: Z. Hajji, and K. Amis, and A. Aïssa-El-Bey, "Channel estimation with finite-alphabet simplicity-based detection for large-scale MIMO systems" .

the finite number of pilot symbols, the channel estimates are not perfect. This has an influence on the achieved capacity of the MIMO channel and the performance of detection algorithms. This chapter addresses the effect of imperfect CSI on MIMO system preformance and proposes channel estimation algorithms. Simulations consider the uplink of large-scale multiuser TDD MIMO communications.
This chapter is organized as follows. Section 5.2 presents an overview of imperfect CSI impact and the problem of channel estimation. Section 5.3 describes the large-scale MIMO system model and usual channel estimation algorithms. The Cramer-Rao bound (CRB) is also investigated for semi-blind channel estimation. In Section 5.4, we propose semi-blind channel estimation algorithms fed by the output of FAS and FAS-SAC detection algorithms in the uncoded case. A theoretical study is done where CRBs and asymptotic MSEs are calculated. Section 5.5 deals with the channel estimation in the coded case combined with the FAS-MAE algorithm detailed in Chapter 4. We propose two ways of exploitation of FEC decoder output to update the channel estimation block. Finally, Section 5.6 concludes the chapter.

5.2 Overview of Imperfect CSI effects

The capacity of MIMO channels can be degraded when CSI is imperfect. This case is well studied in the literature of MIMO communications and some key results are listed hereinafter:

- Gaussian source distribution, which is the distribution that can achieve the MIMO capacity in the case of perfect channel estimate gets suboptimal when CSI is inaccurate [55], [56].
- When Gaussian distribution is considered, the mutual information saturates when increasing SNR with imperfect channel estimate but it still increases linearly with $\min(n, N)$ where n and N stand for receive antenna number and transmit antenna number respectively [57].
- The capacity gain of perfect CSI case over imperfect one decreases when SNR increases. This is due to the fact that the two cases share the same optimal input covariance matrix which is the identity matrix for high SNR. However, when the perfect CSI is exploited at the transmitter side this gain becomes significant because in imperfect CSI the transmitter is fed by erroneous channel estimates and lead to the saturation of the effective SNR [57].
- The effect of imperfect CSI on the achievable capacity was established, and a lower bound on capacity was defined as a function of the Cramer-Rao bound (CRB) [58], [59], [60].

A widely adopted technique in MIMO systems is channel estimation based on known training sequences. This approach is well investigated in several works. The problem of needed training sequences number is addressed in point-to-point frequency flat and selective channels [61], [62], multiuser MIMO channels [63]. In this approach,

the transmission is often divided into training phase and data phase. In the training phase, pilot sequences known at the transmitter and receiver sides are transmitted in order to calculate an estimate at the receiver. This estimate can be obtained using ML or MMSE criterion [64] and then it is used for detection of data. When a small number of pilot sequences is considered, the capacity loss due to pilots is less, however, we get an inaccurate estimate of the channel which degrades the performance of the system. On the other hand, when long pilot sequences are considered, the quality of the channel estimate can improve but we get less time for data transmission. This tradeoff is analyzed in [61]. The authors showed that when allocating more transmit power for pilot sequences their optimal number can be equal to the number of transmit antennas. However, when powers are equally allocated between pilots and data the optimal pilot number might be much larger than the number of transmit antennas.

The multiuser MIMO channel estimation is investigated in [63] and the question of how many training sequences is addressed. For a given coherence time and the number of BS antennas, the optimal number of pilot sequences and the optimal number of users to serve simultaneously are fixed by maximizing a lower bound on the sum-rate in the downlink.

5.3 Overview of channel estimation techniques

5.3.1 System model

Let us consider the large-scale MIMO system equipped with N antennas at the transmitter and n antennas at the receiver. Each transmitted frame consists of T_p pilot vectors and T_d data vectors ($T = T_p + T_d$).

Under the above assumptions, the received signal matrix can be modelled as:

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}. \tag{5.1}$$

$\boldsymbol{Y} = (\boldsymbol{Y}_p, \boldsymbol{Y}_d)$ is the received signal matrix. $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_d$ are the $n \times T_p$ pilot received matrix and the $n \times T_d$ data received matrix respectively. $\boldsymbol{X}$ stands for the transmitted signal matrix. This $N \times T$ complex matrix can be decomposed as $\boldsymbol{X} = (\boldsymbol{X}_p, \boldsymbol{X}_d)$. $\boldsymbol{X}_p$ and $\boldsymbol{X}_d$ are the $N \times T_p$ pilot transmitted matrix and the $N \times T_d$ data transmitted matrix respectively. Note that the channel use at time t, for $t = 1, \ldots, T$, corresponds to the received vector given by:

$$\boldsymbol{y}(t) = \boldsymbol{H}\boldsymbol{x}(t) + \boldsymbol{\zeta}(t). \tag{5.2}$$

The Maximum Likelihood (ML) estimate of $\boldsymbol{H}$ based on both training and data signals is given by

$$\hat{\boldsymbol{H}}_{ML} = \arg\max_{\boldsymbol{H}} \log p(\boldsymbol{Y}|\boldsymbol{H}) \tag{5.3}$$

5.3.2 ML estimators

In this section, we present the ML-based estimator which only uses the training sequences as well as a full data ML estimator which assumes perfect data estimation. The last one can serve as a lower bound on the performance of semi-blind estimation (which consists of a first step of initialization of the channel estimate thanks to pilots followed by estimate refinement with data decisions).

5.3.2.1 ML estimation based on training Pilot Sequences

The ML estimate of the channel matrix $\boldsymbol{H}$ based on pilot sequences $\boldsymbol{X}_p$ is given by:

$$\hat{\boldsymbol{H}}_{ML}^{training} = (\boldsymbol{Y}_p \boldsymbol{X}_p^H)(\boldsymbol{X}_p \boldsymbol{X}_p^H)^{-1}. \tag{5.4}$$

To minimize the MSE subject to the transmit power, the training sequences must be orthogonal, i.e $\boldsymbol{X}_p \boldsymbol{X}_p^H = T_p \boldsymbol{I}_N$. The corresponding mean square error (MSE) is then computed as

$$\mathbb{E}\left[||\boldsymbol{H} - \hat{\boldsymbol{H}}_{ML}^{training}||_2^2\right] = \frac{nN\sigma^2}{T_p} \tag{5.5}$$

The reliability of the channel estimate based on pilot sequences highly depends on the number of orthogonal sequences and the estimator requires more sequences to be more reliable.

5.3.2.2 ML estimation based on full data

The full data-based ML estimator is an estimator supposing that all data symbols are known at the receiver side and the performance of such estimator serves the lower bound on the performance of semi-blind estimation. In this case, all the symbols are assumed to be known at the BS. The channel estimate based on all data $\boldsymbol{X}$ denoted by $\hat{\boldsymbol{H}}_{ML}^{full}$ is computed as:

$$\hat{\boldsymbol{H}}_{ML}^{full} = (\boldsymbol{Y}\boldsymbol{X}^H)(\boldsymbol{X}\boldsymbol{X}^H)^{-1}. \tag{5.6}$$

Its corresponding mean square error (MSE) equals to:

$$\mathbb{E}\left[||\boldsymbol{H} - \hat{\boldsymbol{H}}_{ML}^{full}||_2^2\right] = n\sigma^2 \mathrm{tr}\left(\mathbb{E}\left[(\boldsymbol{X}\boldsymbol{X}^H)^{-1}\right]\right). \tag{5.7}$$

5.3.3 EM algorithm

As data symbols are not known, the ML problem cannot be analytically solved in practice. It is necessary to use iterative algorithms that converge to the solution of (5.3). Among them, the EM algorithm updates the channel estimate based on an old one in the following manner:

$$\hat{\boldsymbol{H}}_{i+1} = \arg\max_{\boldsymbol{H}} \mathbb{E}_{\Pr(\boldsymbol{X}_d|\boldsymbol{Y},\hat{\boldsymbol{H}}_i)} \left(\log \Pr(\boldsymbol{Y}, \boldsymbol{X}_d|\boldsymbol{H})\right). \tag{5.8}$$

As we can see, the algorithm involves an expectation step and a maximization one. The maximization step can be simplified and the updated estimate of the channel matrix can be written as

$$\hat{\boldsymbol{H}}_{i+1} = \left(\boldsymbol{Y}_p\boldsymbol{X}_p^H + \boldsymbol{Y}_d\,\mathbb{E}\left[\boldsymbol{X}_d|\boldsymbol{Y},\hat{\boldsymbol{H}}_i\right]^H\right) \times \left(\boldsymbol{X}_p\boldsymbol{X}_p^H + \mathbb{E}\left[\boldsymbol{X}_d\boldsymbol{X}_d^H|\boldsymbol{Y},\hat{\boldsymbol{H}}_i\right]^H\right)^{-1} \quad (5.9)$$

To compute the estimate (5.9), the expectation step (E-step) must be defined. Using a discrete random variables such as 4-QAM leads to complex E-step whose complexity grows exponentially with N. To overcome this problem, in [64], the authors propose to assume that the data symbols are Gaussian. Thus $\mathbb{E}\left[\boldsymbol{X}_d|\boldsymbol{Y},\hat{\boldsymbol{H}}_i\right]^H$ and $\mathbb{E}\left[\boldsymbol{X}_d\boldsymbol{X}_d^H|\boldsymbol{Y},\hat{\boldsymbol{H}}_i\right]^H$ can be computed from the conditional density of circularly symmetric Gaussian random vectors and we get the updated estimate as follows:

$$\hat{\boldsymbol{H}}_{i+1} = \left(\boldsymbol{Y}_p\boldsymbol{X}_p^H + \sum_{t=T_p+1}^{T}\boldsymbol{y}(t)\left(\hat{\boldsymbol{x}}(t)\right)^H\right) \times \left(\boldsymbol{X}_p\boldsymbol{X}_p^H + \sum_{t=T_p+1}^{T}\left(\hat{\boldsymbol{x}}(t)\left(\hat{\boldsymbol{x}}(t)\right)^H + \boldsymbol{\Sigma}\right)\right)^{-1} \quad (5.10)$$

where

$$\hat{\boldsymbol{x}}(t) = \left(\hat{\boldsymbol{H}}_i^H\hat{\boldsymbol{H}}_i + \sigma^2\boldsymbol{I}_N\right)^{-1}\hat{\boldsymbol{H}}_i^H\boldsymbol{y}(t), \quad (5.11)$$

and

$$\boldsymbol{\Sigma} = \sigma^2\left(\hat{\boldsymbol{H}}_i^H\hat{\boldsymbol{H}}_i + \sigma^2\boldsymbol{I}_N\right)^{-1} \quad (5.12)$$

5.3.4 Cramer-Rao bound of semi-blind channel estimation

In order to calculate the CRB of semi-blind channel estimation, we define for $t = T_p+1,\ldots,T$, the matrix $\boldsymbol{\Omega}_t$ as $\boldsymbol{\Omega}_t = [\underline{x}_1(t)\underline{\boldsymbol{H}},\ldots,\underline{x}_{2N}(t)\underline{\boldsymbol{H}}]$. Let us define also $\underline{\boldsymbol{x}}^i$ as the i-th line of the matrix $\underline{\boldsymbol{X}}$ and the matrix $\boldsymbol{\Pi}$ as follows:

$$\boldsymbol{\Pi} = \begin{pmatrix} \left(\frac{2}{\sigma^2}(\underline{\boldsymbol{x}}^1)(\underline{\boldsymbol{x}}^1)^T\right)\boldsymbol{I}_{2N} & \cdots & \left(\frac{2}{\sigma^2}(\underline{\boldsymbol{x}}^1)(\underline{\boldsymbol{x}}^{2N})^T\right)\boldsymbol{I}_{2N} \\ \vdots & \ddots & \vdots \\ \left(\frac{2}{\sigma^2}(\underline{\boldsymbol{x}}^{2N})(\underline{\boldsymbol{x}}^1)^T\right)\boldsymbol{I}_{2N} & \cdots & \left(\frac{2}{\sigma^2}(\underline{\boldsymbol{x}}^{2N})(\underline{\boldsymbol{x}}^{2N})^T\right)\boldsymbol{I}_{2N} \end{pmatrix} \quad (5.13)$$

The following theorem gives the CRB of any unbiased semi-blind channel estimation based on least squares algorithm.

Theorem 5.3.1 *The deterministic Cramer-Rao bound of the covariance matrix of any unbiased semi-blind estimate of $\boldsymbol{H}$ is given by:*

$$CRB(\boldsymbol{H}) = \left(\boldsymbol{\Pi} - \sum_{t=T_p+1}^{T}\boldsymbol{\Omega}_t^T\left(\underline{\boldsymbol{H}}^T\underline{\boldsymbol{H}}\right)\boldsymbol{\Omega}_t\right)^{-1} \quad (5.14)$$

[Proof of Theorem 5.3.1] *The log likelihood function of the received signal is given by:*

$$\mathcal{L} = q - \sum_{t=1}^{T} \frac{2}{\sigma^2} \left\| \underline{\boldsymbol{y}}(t) - \underline{\boldsymbol{H}}\underline{\boldsymbol{x}}(t) \right\|^2, \tag{5.15}$$

where q is real and constant. The CRB of both channel coefficients and unknown data symbols is computed as:

$$CRB(\boldsymbol{X}_d, \boldsymbol{H}) = \mathbb{E}(\mathcal{I}\mathcal{I}^T)^{-1}, \tag{5.16}$$

where

$$\mathcal{I} = \partial\mathcal{L}/\partial \left[\underline{\boldsymbol{x}}^T(T_p+1), ..., \underline{\boldsymbol{x}}^T(T), \underline{\boldsymbol{h}}_1^T, ..., \underline{\boldsymbol{h}}_{2N}^T \right] \tag{5.17}$$

Following similar steps as in the proof of [65], for $t = T_p+1, \ldots, T$ and $k = 1, \ldots, 2N$, we can show that

$$\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{x}}(t)} = \frac{2}{\sigma^2} \underline{\boldsymbol{H}}^T \underline{\boldsymbol{\zeta}}(t) \tag{5.18}$$

$$\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{\boldsymbol{\zeta}}(t) \underline{x}_k(t) \tag{5.19}$$

Using $\mathbb{E}\left[\underline{\boldsymbol{\zeta}}(t) \left(\underline{\boldsymbol{\zeta}}(t) \right)^H \right] = \frac{\sigma^2}{2} \boldsymbol{I}_{2n} \delta(t-t')$, for $t, t' = T_p+1, \ldots, T$ and $k, i = 1, \ldots, 2N$. We then get

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{x}}(t)} \left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{x}}(t')} \right)^T \right) = \frac{2}{\sigma^2} \left(\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}} \right) \delta(t-t') \tag{5.20}$$

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{x}}(t)} \left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \right)^T \right) = \frac{2}{\sigma^2} \underline{x}_k(t) \underline{\boldsymbol{H}}^T \tag{5.21}$$

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{h}}_i} \left(\frac{\partial\mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \right)^T \right) = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{x}_i(t) \underline{x}_k(t) \tag{5.22}$$

Substituting (5.20), (5.21) and (5.22) in (5.16), we get

$$CRB(\boldsymbol{H}) = \begin{pmatrix} \frac{2}{\sigma^2} \left(\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}} \right) & \cdots & 0 & \boldsymbol{\Omega}_{(T_p+1)} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{2}{\sigma^2} \left(\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}} \right) & \boldsymbol{\Omega}_{(T)} \\ \boldsymbol{\Omega}_{(T_p+1)}^H & \cdots & \boldsymbol{\Omega}_{(T)}^H) & \boldsymbol{\Pi} \end{pmatrix}^{-1} \tag{5.23}$$

By results on the inverse of a block matrix, the CRB of the channel matrix $\boldsymbol{H}$

$$CRB(\boldsymbol{H}) = \left(\boldsymbol{\Pi} - \sum_{t=T_p+1}^{T} \boldsymbol{\Omega}_t^T \left(\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}} \right) \boldsymbol{\Omega}_t \right)^{-1} \tag{5.24}$$

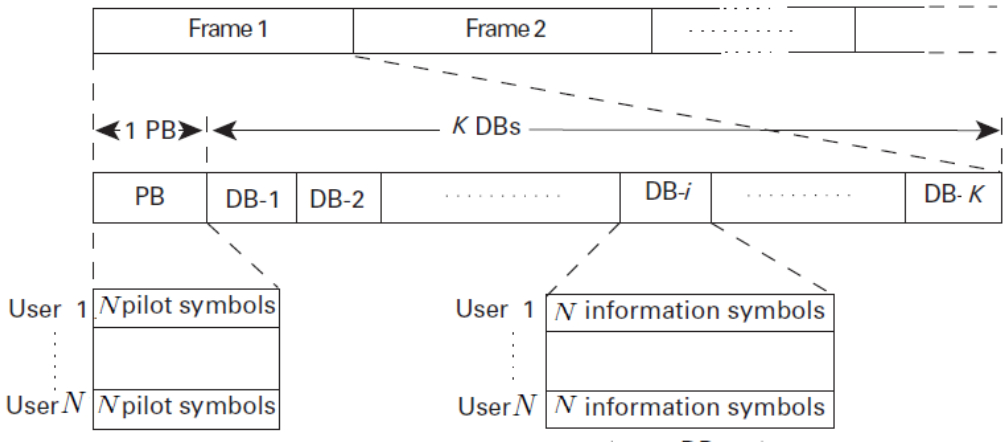

Figure 5.1: Frame structure

Remark 5.3.1 *When the symbols are all correctly detected, we get the* $\sum_{t=T_p+1}^{T} \boldsymbol{\Omega}_t^T \left(\underline{\boldsymbol{H}}^T \underline{\boldsymbol{H}}\right) \boldsymbol{\Omega}_t = 0$ *and the CRB is computed as* $CRB(\boldsymbol{H}) = \boldsymbol{\Pi}^{-1}$ *which is equal to the covariance matrix of the full data and we get the MSE of the semi-blind estimation equal to* $tr(\boldsymbol{\Pi}^{-1}) = \mathbb{E}(||\boldsymbol{H} - \hat{\boldsymbol{H}}_{ML}^{full}||_2^2)$ *as defined in (5.7).*

5.4 Semi-blind uplink channel estimation for large-scale MIMO systems

In this section, we develop joint semi-blind channel estimation and detection schemes based on FAS algorithms proposed and described in previous chapters.

5.4.1 Proposed Semi-blind uplink channel estimation algorithms

The frame structure is depicted in Fig. 5.1. A slow fading channel is assumed, where the channel is assumed to be constant over one frame duration. Each frame consists of a pilot block for the purpose of initial channel estimation, followed by K data blocks. The pilot block consists of N pilot symbol vectors of length N. Each one is allocated to a given user. Each data block consists of N information symbol vectors, each of length N (one data symbol vector per user). Compared to the system model of 5.3, we get $T_p = N$ and $T_d = KN$. Data blocks are detected using FAS and FAS-SAC detection algorithms using an initial estimate provided by a pilot sequences-based channel estimation. The detected data blocks are then iteratively used to refine the channel estimates thanks to least squares based channel estimation algorithm described below.

5.4.1.1 Initial channel estimate during pilot phase

Let $\boldsymbol{x}_p^u$ denote the transmitted pilot symbol vector from user u. Let $\boldsymbol{X}_p = [\boldsymbol{x}_p^1, \boldsymbol{x}_p^2, \ldots, \boldsymbol{x}_p^N]$ denote the $N \times N$ pilot matrix formed by the pilot symbol vec-

tors transmitted by all users in the pilot transmission phase. The received signal matrix at BS, $\boldsymbol{Y}_p$, is given by:

$$\boldsymbol{Y}_p = \boldsymbol{H}\boldsymbol{X}_p + \boldsymbol{Z}_p, \tag{5.25}$$

where $\boldsymbol{Z}_p$ is the noise matrix at the BS. The following pilot sequence is used:

$$\boldsymbol{x}_p^u = [\boldsymbol{0}_{u\times 1}, \lambda, \boldsymbol{0}_{(N-u-1)\times 1}], \tag{5.26}$$

where $\lambda = \sqrt{NE_s}$ and E_s is the average symbol energy. Using the scaled identity nature of $\boldsymbol{x}_p$, an initial estimate $\hat{\boldsymbol{H}}_0$ is obtained as:

$$\begin{aligned} \hat{\boldsymbol{H}}_0 &= \boldsymbol{Y}_p/\lambda \\ &= \boldsymbol{H} + \frac{1}{\lambda}\boldsymbol{Z}_p \end{aligned} \tag{5.27}$$

5.4.1.2 Data detection using initial channel estimate

During data transmission phase, the received signal matrix at BS, $\boldsymbol{Y}_d$, is given by:

$$\boldsymbol{Y}_d = \boldsymbol{H}\boldsymbol{X}_d + \boldsymbol{Z}_d, \tag{5.28}$$

where $\boldsymbol{X}_d$ is the concatenation of different data blocks. The received vector for the channel use at time t, for $t = T_p + 1, \ldots, T$ is

$$\boldsymbol{y}(t) = \boldsymbol{H}\boldsymbol{x}(t) + \boldsymbol{\zeta}(t), \tag{5.29}$$

The initial channel estimate $\hat{\boldsymbol{H}}_0$ obtained from (5.27) is used to detect the transmitted data vectors using FAS and FAS-SAC algorithms proposed in previous chapters to get $\hat{\boldsymbol{X}}_d$ an estimate for the transmitted data matrix $\boldsymbol{X}_d$.
One knows that $\hat{\boldsymbol{H}}_0 = \boldsymbol{H} + \frac{1}{\lambda}\boldsymbol{Z}_p$. This initial CSI error is used to calculate the statistics needed to study the performance of the proposed schemes in the uncoded cases and to interface the proposed detectors with FEC decoder and channel estimation block in the coded case. We then get, at channel use t, $\boldsymbol{y}(t) - \hat{\boldsymbol{H}}_0\boldsymbol{x}(t)$, the updated noise vector at first iteration with zero mean and variance $(1 + \frac{N}{T_p})\sigma^2 = 2\sigma^2$.

5.4.1.3 Channel estimation refinement based on FAS soft-decision output

Let $\hat{\boldsymbol{X}}_{FAS} = [\hat{\boldsymbol{x}}(T_p+1), \ldots, \hat{\boldsymbol{x}}(T)]$; Where $\hat{\boldsymbol{x}}(t)$, for $t = T_p + 1, ..., T$, is the detected complex-valued FAS soft-decision output at time t based on the channel estimate at i-th iteration. To update the channel estimate at the $(i+1)$-th iteration, we propose to compute the following criterion:

$$\hat{\boldsymbol{H}}_{i+1} = \left(\boldsymbol{Y}_p\boldsymbol{X}_p^H + \boldsymbol{Y}_d\hat{\boldsymbol{X}}_{FAS}^H\right)\left(\boldsymbol{X}_p\boldsymbol{X}_p^H + \hat{\boldsymbol{X}}_{FAS}\hat{\boldsymbol{X}}_{FAS}^H\right)^{-1}. \tag{5.30}$$

In order to compute the CRB of the proposed channel estimation algorithm based on FAS detection, let us remember that at time t, the elements of a detected real-valued vector $\underline{\hat{\boldsymbol{x}}}(t)$ by FAS algorithm can be classified into two sets (see Definition

3.4.1). The first set is the set of reliable elements which are equal exactly to the transmitted symbols. This set was referred to as Λ_t. Its cardinality follows the binomial distribution with parameters $2N$ and $\frac{1}{2}$ assuming 4-QAM modulation. The second set referred to as $\bar{\Lambda}_t$ contains the rest of elements disturbed by noise. Its cardinality follows the same distribution as Λ_t.

Let us now define for $t = T_p + 1, \ldots, T$, the matrix $\boldsymbol{G}_t^{FAS}$ as the matrix $\underline{\underline{\boldsymbol{H}}}$ annulling the entries of all columns of indices in Λ_t, the matrix $\boldsymbol{A}_t^{FAS} = \left(\boldsymbol{G}_t^{FAS}\right)^H \boldsymbol{G}_t^{FAS}$ and the matrix $\boldsymbol{\Omega}_t^{FAS}$ as $\boldsymbol{\Omega}_t^{FAS} = \left[\underline{x}_1(t)\boldsymbol{G}_t^{FAS}, \ldots, \underline{x}_{2N}(t)\boldsymbol{G}_t^{FAS}\right]$. Let us also introduce $\boldsymbol{C}_t^{FAS} = \left(\underline{\underline{\boldsymbol{H}}}_{\Lambda_t}^T \underline{\underline{\boldsymbol{H}}}_{\bar{\Lambda}_t}\right)^{-1}$. We define the matrix $\mathfrak{C}_t^{FAS}$ such that $\mathfrak{C}_t^{FAS}\left(\bar{\Lambda}_t(i), \bar{\Lambda}_t(j)\right) = \boldsymbol{C}_t^{FAS}(i,j)$ for $(i,j) \in \{1, \ldots, |\bar{\Lambda}_t|\}^2$ and the other entries are equal to zero.
The following theorem defines the CRB of the proposed channel estimation/detection scheme.

Theorem 5.4.1 *The deterministic CRB of the channel estimation based on FAS detection algorithm at second iteration is defined by*

$$CRB(\boldsymbol{H}) = \left(\boldsymbol{\Pi} - \sum_{t=T_p+1}^{T} (\boldsymbol{\Omega}_t^{FAS})^T \mathfrak{C}_t(\boldsymbol{\Omega}_t^{FAS})\right)^{-1} \tag{5.31}$$

[Proof of Theorem 5.4.1] *Following similar steps as in the proof of 5.3.4, for $t = T_p + 1, \ldots, T$ and $k = 1, \ldots, 2N$, we can show that*
for $k \in \bar{\Lambda}_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} = \frac{2}{\sigma^2} \underline{\boldsymbol{h}}_k^T \underline{\boldsymbol{\zeta}}(t) \tag{5.32}$$

for $k \in \Lambda_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} = 0 \tag{5.33}$$

for $k \in \{1, \ldots, 2N\}$

$$\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{\boldsymbol{\zeta}}(t) \underline{x}_k(t) \tag{5.34}$$

We then get
for $(k, k') \in (\bar{\Lambda}_t \times \bar{\Lambda}_{t'})$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{x}_{k'}(t')}\right)^T\right) = \frac{2}{\sigma^2} \left(\underline{\boldsymbol{h}}_k^T \underline{\boldsymbol{h}}_{k'}\right) \delta(t - t') \tag{5.35}$$

for $(k, k') \in (\Lambda_t \times \Lambda_{t'})$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{x}_{k'}(t')}\right)^T\right) = 0 \tag{5.36}$$

so

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{x}}(t)}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{x}}(t')}\right)^T\right)=\frac{2}{\sigma^2}\left(\boldsymbol{A}_t^{FAS}\right)\delta(t-t') \tag{5.37}$$

for $k\in\bar{\Lambda}_t$

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial\underline{x}_k(t)}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{h}}_k}\right)^T\right)=\frac{2}{\sigma^2}\left(\underline{\boldsymbol{h}}_k^T\underline{x}_k(t)\right) \tag{5.38}$$

for $k\in\Lambda_t$

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial\underline{x}_k(t)}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{h}}_k}\right)^T\right)=0 \tag{5.39}$$

so

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{x}}(t)}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{h}}_k}\right)^T\right)=\frac{2}{\sigma^2}\underline{x}_k(t)(\boldsymbol{G}_t^{FAS})^T \tag{5.40}$$

for $k\in\{1,\ldots,2N\}$

$$\mathbb{E}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{h}}_k}\left(\frac{\partial\mathcal{L}}{\partial\underline{\boldsymbol{h}}_i}\right)^T\right)=\frac{2}{\sigma^2}\sum_{t=1}^{T}\underline{x}_k(t)\underline{x}_i(t) \tag{5.41}$$

We then get

$$CRB(\boldsymbol{H})=\begin{pmatrix}\boldsymbol{A}_{(T_p+1)}^{FAS} & \ldots & 0 & \boldsymbol{\Omega}_{(T_p+1)}^{FAS}\\ \vdots & \ddots & \vdots & \vdots\\ 0 & \ldots & \boldsymbol{A}_{(T)}^{FAS} & \boldsymbol{\Omega}_{(T)}^{FAS}\\ (\boldsymbol{\Omega}_{(T_p+1)}^{FAS})^H & \ldots & (\boldsymbol{\Omega}_{(T)}^{FAS})^H & \boldsymbol{\Pi}\end{pmatrix}^{-1} \tag{5.42}$$

Finally,

$$CRB(\boldsymbol{H})=\left(\boldsymbol{\Pi}-\sum_{t=T_p+1}^{T}(\boldsymbol{\Omega}_t^{FAS})^H\mathfrak{C}_t^{FAS}(\boldsymbol{\Omega}_t^{FAS})\right)^{-1} \tag{5.43}$$

5.4.1.4 Channel estimation refinement based on FAS-SAC soft-decision output

Let $\hat{\boldsymbol{X}}_{FS}=[\hat{\boldsymbol{x}}(T_p+1),\ldots,\hat{\boldsymbol{x}}(T)]$, where for $t=T_p+1,...,T$, $\hat{\boldsymbol{x}}(t)$ is the detected complex-valued FAS-SAC soft-decision output at time t based on channel estimate at i-th iteration. We propose to update the channel estimate at the $(i+1)$-th iteration by:

$$\hat{\boldsymbol{H}}_i=\left(\boldsymbol{Y}_p\boldsymbol{X}_p^H+\boldsymbol{Y}_d\hat{\boldsymbol{X}}_{FS}^H\right)\left(\boldsymbol{X}_p\boldsymbol{X}_p^H+\hat{\boldsymbol{X}}_{FS}\hat{\boldsymbol{X}}_{FS}^H\right)^{-1}. \tag{5.44}$$

In order to determine the CRB of the channel estimation algorithm based on FAS-SAC algorithm, we consider the iterative FAS-SAC detection with two iterations. As detailed in Section 4.2, denoting by $\hat{\underline{\boldsymbol{x}}}(t)$ the output of the second iteration, its elements belong to different sets. The set $\mathcal{A}_t$ contains the more reliable elements that can be divided into correct decision symbols and erroneous decision symbols. The sets of correct decisions and erroneous decisions are referred to as $(\mathcal{A}_c)_t$ and $(\mathcal{A}_e)_t$ respectively and we get $\mathcal{A}_t = (\mathcal{A}_c)_t \cup (\mathcal{A}_e)_t$. The complementary of $\mathcal{A}_t$ is the set $\bar{\mathcal{A}}_t$ which contains the symbols that are re-estimated in the second iteration. The updated noise vector in the second iteration is $\tilde{\boldsymbol{\zeta}}(t)$ with variance $\sigma^2_{\tilde{\zeta}}(\eta)$ calculated in 4.7.

Let us then define for $t = T_p + 1, \ldots, T$, the matrix $\boldsymbol{G}_t^{FS}$ as the matrix $\underline{\boldsymbol{H}}$ annulling the entries of all columns of indices in $\mathcal{A}_t$, the matrix $\boldsymbol{A}_t^{FS} = \left(\boldsymbol{G}_t^{FS}\right)^H \boldsymbol{G}_t^{FS}$ and the matrix $\boldsymbol{\Omega}_t^{FS} = \left[\underline{x}_1(t)\boldsymbol{G}_t^{FS}, \ldots, \underline{x}_{2N}(t)\boldsymbol{G}_t^{FS}\right]$.

Let us also introduce $\boldsymbol{C}_t^{FS} = \left(\underline{\boldsymbol{H}}_{\bar{\mathcal{A}}_t}^T \underline{\boldsymbol{H}}_{\bar{\mathcal{A}}_t}\right)^{-1}$. We define the matrix $\mathfrak{C}_t^{FS}$ such that $\mathfrak{C}_t^{FS}\left(\bar{\mathcal{A}}_t(i), \bar{\mathcal{A}}_t(j)\right) = \boldsymbol{C}_t^{FS}(i,j)$ for $(i,j) \in \{1, \ldots, |\bar{\mathcal{A}}_t|\}^2$ and the other entries are equal to zero, for $t = T_p + 1, \ldots, T$.
Let $\tilde{\boldsymbol{X}}$ be the symbol matrix such that the elements of the t-th column with indices in $(\mathcal{A}_e)_t$ are fixed to those of the resulted matrix $\hat{\boldsymbol{X}}_{FS}$ and the rest of elements are fixed to the elements of the transmitted symbol matrix $\boldsymbol{X}$. Let then $\boldsymbol{\Pi}^{FS}$ the matrix defined by:

$$\boldsymbol{\Pi}^{FS} = \begin{pmatrix} \left(\frac{2}{\sigma^2}(\tilde{\underline{\boldsymbol{x}}}^1)(\tilde{\underline{\boldsymbol{x}}}^1)^T\right)\boldsymbol{I}_{2N} & \ldots & \left(\frac{2}{\sigma^2}(\tilde{\underline{\boldsymbol{x}}}^1)(\tilde{\underline{\boldsymbol{x}}}^{2N})^T\right)\boldsymbol{I}_{2N} \\ \vdots & \ddots & \vdots \\ \left(\frac{2}{\sigma^2}(\tilde{\underline{\boldsymbol{x}}}^{2N})(\tilde{\underline{\boldsymbol{x}}}^1)^T\right)\boldsymbol{I}_{2N} & \ldots & \left(\frac{2}{\sigma^2}(\tilde{\underline{\boldsymbol{x}}}^{2N})(\tilde{\underline{\boldsymbol{x}}}^{2N})^T\right)\boldsymbol{I}_{2N} \end{pmatrix} \tag{5.45}$$

The following theorem gives the deterministic CRB of the proposed channel estimation scheme.

Theorem 5.4.2 *The deterministic CRB of the channel estimation based on FAS-SAC detection algorithm is given by*

$$CRB(\boldsymbol{H}) = \left(\boldsymbol{\Pi}^{FS} - \sum_{t=T_p+1}^{T} (\boldsymbol{\Omega}_t^{FS})^T \mathfrak{C}_t^{FS} (\boldsymbol{\Omega}_t^{FS})\right)^{-1} \tag{5.46}$$

[Proof of Theorem 5.4.2] *Following similar steps as in the proof of 5.3.4, for $t = T_p + 1, \ldots, T$ and $k = 1, \ldots, 2N$, we can show that*
for $k \in \bar{\mathcal{A}}_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \underline{\boldsymbol{h}}_k^T \tilde{\underline{\boldsymbol{\zeta}}}(t) \tag{5.47}$$

for $k \in \mathcal{A}_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} = 0 \tag{5.48}$$

for $k \in \bar{\mathcal{A}}_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \sum_{t=1}^{T} \tilde{\underline{\boldsymbol{\zeta}}}(t) \underline{x}_k(t) \tag{5.49}$$

for $k \in (\mathcal{A}_c)_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{\boldsymbol{\zeta}}(t) \underline{x}_k(t) \tag{5.50}$$

for $k \in (\mathcal{A}_e)_t$

$$\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{\boldsymbol{\zeta}}(t) (\underline{x}_k(t) + \delta \underline{x}_k(t)) \tag{5.51}$$

Note that $\delta \underline{x}_k(t)$ *takes values in* $\{-2, 2\}$ *when 4-QAM modulation is assumed. We get then*
for $(k, k') \in (\bar{\mathcal{A}}_t \times \bar{\mathcal{A}}_{t'})$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{x}_{k'}(t')} \right)^T \right) = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \left(\underline{\boldsymbol{h}}_k^T \underline{\boldsymbol{h}}_{k'} \right) \delta(t - t') \tag{5.52}$$

for $(k, k') \in (\mathcal{A}_t \times \mathcal{A}_{t'})$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{x}_{k'}(t')} \right)^T \right) = 0 \tag{5.53}$$

for $k \in \bar{\mathcal{A}}_t$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \right)^T \right) = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \left(\underline{x}_k(t) \underline{\boldsymbol{h}}_k^T \right) \tag{5.54}$$

for $k \in \mathcal{A}_t$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{x}_k(t)} \left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \right)^T \right) = 0 \tag{5.55}$$

for $(k, i) \in (\bar{\mathcal{A}}_t)^2$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_i} \right)^T \right) = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \sum_{t=1}^{T} \underline{x}_k^T(t) \underline{x}_i(t) \tag{5.56}$$

for $(k, i) \in (\mathcal{A}_c)_t^2$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k} \left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_i} \right)^T \right) = \frac{2}{\sigma^2} \sum_{t=1}^{T} \underline{x}_k^T(t) \underline{x}_i(t) \tag{5.57}$$

for $(k,i) \in ((\mathcal{A}_c)_t \times (\mathcal{A}_e)_t)$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_i}\right)^T\right) = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \sum_{t=1}^{T} \underline{x}_k^T(t)\left(\underline{x}_i(t) + \delta\underline{x}_i(t)\right) \tag{5.58}$$

for $(k,i) \in ((\mathcal{A}_e)_t \times (\mathcal{A}_c)_t)$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_i}\right)^T\right) = \frac{2}{\sigma^2_{\tilde{\zeta}}(\eta)} \sum_{t=1}^{T} \left(\underline{x}_k(t) + \delta\underline{x}_k(t)\right)^T \underline{x}_i(t) \tag{5.59}$$

for $(k,i) \in (\mathcal{A}_e)_t^2$

$$\mathbb{E}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_k}\left(\frac{\partial \mathcal{L}}{\partial \underline{\boldsymbol{h}}_i}\right)^T\right) = \frac{2}{\sigma^2} \sum_{t=1}^{T} \left(\underline{x}_k(t) + \delta\underline{x}_k(t)\right)^T \left(\underline{x}_i(t) + \delta\underline{x}_i(t)\right) \tag{5.60}$$

We get then

$$CRB(\boldsymbol{H}) = \begin{pmatrix} \boldsymbol{A}^{FS}_{(T_p+1)} & \dots & 0 & \boldsymbol{\Omega}^{FS}_{(T_p+1)} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \boldsymbol{A}^{FS}_{(T)} & \boldsymbol{\Omega}^{FS}_{(T)} \\ (\boldsymbol{\Omega}^{FS}_{(T_p+1)})^H & \dots & (\boldsymbol{\Omega}^{FS}_{(T)})^H & \boldsymbol{\Pi}^{FS} \end{pmatrix}^{-1} \tag{5.61}$$

Finally,

$$CRB(\boldsymbol{H}) = \left(\boldsymbol{\Pi}^{FS} - \sum_{t=T_p+1}^{T} (\boldsymbol{\Omega}_t^{FS})^H \mathfrak{C}_t^{FS} (\boldsymbol{\Omega}_t^{FS})\right)^{-1} \tag{5.62}$$

5.4.1.5 Channel estimation refinement based on FAS/FAS-SAC hard decision output

Let $\tilde{\boldsymbol{X}}_{FAS} = [\tilde{\boldsymbol{x}}(T_p+1), \dots, \tilde{\boldsymbol{x}}(T)]$, where $\tilde{\boldsymbol{x}}(t)$ is the hard decision of the detected complex-valued FAS output at time t based on the channel estimate at the i-th iteration. To update the channel estimate at the $(i+1)$-th iteration, we propose to compute the following criterion:

$$\hat{\boldsymbol{H}}_i = \left(\boldsymbol{Y}_p \boldsymbol{X}_p^H + \boldsymbol{Y}_d \tilde{\boldsymbol{X}}_{FAS}^H\right)\left(\boldsymbol{X}_p \boldsymbol{X}_p^H + \tilde{\boldsymbol{X}}_{FAS} \tilde{\boldsymbol{X}}_{FAS}^H\right)^{-1}. \tag{5.63}$$

Let $\tilde{\boldsymbol{X}}_{FAS-SAC} = [\tilde{\boldsymbol{x}}(T_p+1), \dots, \tilde{\boldsymbol{x}}(T)]$, where $\tilde{\boldsymbol{x}}(t)$ is the hard decision of the detected complex-valued FAS-SAC output at time t based on the channel estimate at the i-th iteration. To update the channel estimate at the $(i+1)$-th iteration, we propose to compute the following criterion:

$$\hat{\boldsymbol{H}}_i = \left(\boldsymbol{Y}_p \boldsymbol{X}_p^H + \boldsymbol{Y}_d \tilde{\boldsymbol{X}}_{FAS-SAC}^H\right)\left(\boldsymbol{X}_p \boldsymbol{X}_p^H + \tilde{\boldsymbol{X}}_{FAS-SAC} \tilde{\boldsymbol{X}}_{FAS-SAC}^H\right)^{-1}. \tag{5.64}$$

The following theorem gives the asymptotic MSE of the proposed channel estimation schemes based on hard decisions of proposed detection algorithm outputs. Let us consider 4-QAM complex-valued symbols yielding BPSK real-valued symbols in the real-equivalent system. The estimated data matrix $\underline{\tilde{\boldsymbol{X}}}$ based on hard decisions simply writes $\underline{\tilde{\boldsymbol{X}}} = \underline{\boldsymbol{X}} + \underline{\Delta \boldsymbol{X}}$ where $\underline{\Delta \boldsymbol{X}}$ is error matrix with entries in the set $\{-2, 0, 2\}$.

Theorem 5.4.3 *Let us consider 4-QAM complex-valued alphabet. The asymptotic MSE of the channel estimation combined with hard output detection equals:*

$$\begin{aligned} MSE &= \mathbb{E}(||\boldsymbol{H} - \hat{\boldsymbol{H}}||^2) \\ &= \frac{2N\sigma^2}{T} + \frac{\sigma^2}{T^2}\mathrm{tr}\left(\mathbb{E}\left[((\underline{\Delta \boldsymbol{X}})^H \underline{\boldsymbol{X}\boldsymbol{H}}^H \underline{\boldsymbol{H}\boldsymbol{X}}^H (\underline{\Delta \boldsymbol{X}})\right]\right) \end{aligned} \tag{5.65}$$

[Proof of Theorem 5.4.3] *The channel estimate can be written as follows:*

$$\begin{aligned} \underline{\hat{\boldsymbol{H}}} &= \left(\underline{\boldsymbol{Y}\tilde{\boldsymbol{X}}}^H\right)\left(\underline{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}}^H\right)^{-1} \\ &= \frac{1}{T}(\underline{\boldsymbol{H}\boldsymbol{X}} + \underline{\boldsymbol{Z}})(\underline{\boldsymbol{X}} + (\underline{\Delta \boldsymbol{X}}))^H \\ &= \frac{1}{T}(\underline{\boldsymbol{H}\boldsymbol{X}\boldsymbol{X}}^H + \underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H + \underline{\boldsymbol{Z}\boldsymbol{X}}^H + \underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H) \end{aligned} \tag{5.66}$$

Assuming that the frame size is very large (i.e $2N << T$), the covariance matrix of the frame block can be approximated by $\underline{\boldsymbol{X}}^H \underline{\boldsymbol{X}} \approx T\boldsymbol{I}_{2N}$. The channel error can then be written as:

$$\begin{aligned} \underline{\Delta \boldsymbol{H}} &= \underline{\hat{\boldsymbol{H}}} - \underline{\boldsymbol{H}} \\ &= \frac{1}{T}(\underline{\boldsymbol{Z}\boldsymbol{X}}^H + \underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H + \underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H) \end{aligned} \tag{5.67}$$

Assuming that the channel noise is independent of the transmitted symbols and symbols errors we get:

$$\mathbb{E}\left[\underline{\Delta \boldsymbol{H}}\right] = \frac{1}{T}(\mathbb{E}\left[\underline{\boldsymbol{H}}\right]\mathbb{E}\left[\underline{\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H\right]) \tag{5.68}$$

As the channel is a zero-mean random matrix, the estimation bias is always zero. The MSE of the channel estimation reads:

$$\begin{aligned} &\mathbb{E}(||\boldsymbol{H} - \hat{\boldsymbol{H}}||^2) = \mathbb{E}\left[\left(\mathrm{tr}\left(\underline{\boldsymbol{H}} - \underline{\hat{\boldsymbol{H}}}\right)(\underline{\boldsymbol{H}} - \underline{\hat{\boldsymbol{H}}})^H\right)\right] \\ = \ &\frac{1}{T^2}\mathbb{E}\left[\mathrm{tr}\left(\left[\underline{\boldsymbol{Z}\boldsymbol{X}}^H + \underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H + \underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H\right]\left[\underline{\boldsymbol{Z}\boldsymbol{X}}^H + \underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H + \underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H\right]^H\right)\right] \end{aligned} \tag{5.69}$$

Let us now calculate the error terms of (5.69). The first error term equals:

$$\begin{aligned} \mathbb{E}\left[\mathrm{tr}\left(\left(\underline{\boldsymbol{Z}\boldsymbol{X}}^H\right)\left(\underline{\boldsymbol{Z}\boldsymbol{X}}^H\right)^H\right)\right] &= T\sigma^2\mathrm{tr}\left(\boldsymbol{I}_{2N}\right) \\ &= 2NT\sigma^2 \end{aligned} \tag{5.70}$$

The second term is calculated as:

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\left(\underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H\right)\left(\underline{\boldsymbol{Z}\boldsymbol{X}}^H\right)^H\right)\right] &= \sigma^2 \mathrm{tr}\left(\mathbb{E}\left[(\underline{\Delta \boldsymbol{X}})^H \underline{\boldsymbol{X}}\right]\right) \\ &= -4\sigma^2 NF, \end{aligned} \tag{5.71}$$

where F is the number of errors in the detected symbols block $\hat{\boldsymbol{X}}$.
The third term which considers the noise, the data block and the hard decisions errors can be calculated as follows:

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\left(\underline{\boldsymbol{Z}\boldsymbol{X}}^H\right)\left(\underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H\right)^H\right)\right] &= \sigma^2 \mathrm{tr}\left(\mathbb{E}\left[\underline{\boldsymbol{X}}^H(\underline{\Delta \boldsymbol{X}})\right]\right) \\ &= -4\sigma^2 NF \end{aligned} \tag{5.72}$$

Another term is computed as:

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\left(\underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H\right)\left(\underline{\boldsymbol{Z}}(\underline{\Delta \boldsymbol{X}})^H\right)^H\right)\right] &= \sigma^2 \mathrm{tr}\left(\mathbb{E}\left[(\underline{\Delta \boldsymbol{X}})^H(\underline{\Delta \boldsymbol{X}})\right]\right) \\ &= 8\sigma^2 NF \end{aligned} \tag{5.73}$$

We also get,

$$\mathrm{tr}\left(\mathbb{E}\left[(\underline{\Delta \boldsymbol{X}})^H \underline{\boldsymbol{X}}\right]\right) + \mathrm{tr}\left(\mathbb{E}\left[\underline{\boldsymbol{X}}^H(\underline{\Delta \boldsymbol{X}})\right]\right) + \mathrm{tr}\left(\mathbb{E}\left[(\underline{\Delta \boldsymbol{X}})^H(\underline{\Delta \boldsymbol{X}})\right]\right) = 0. \tag{5.74}$$

Last term in (5.69) is computed as:

$$\mathbb{E}\left[\mathrm{tr}\left(\left(\underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H\right)\left(\underline{\boldsymbol{H}\boldsymbol{X}}(\underline{\Delta \boldsymbol{X}})^H\right)^H\right)\right] = \sigma^2 \mathrm{tr}\left(\mathbb{E}\left[(\underline{\Delta \boldsymbol{X}})^H \underline{\boldsymbol{X}\boldsymbol{H}}^H \underline{\boldsymbol{H}\boldsymbol{X}}^H(\underline{\Delta \boldsymbol{X}})\right]\right).$$

We finally get (5.65).

5.4.2 Simulation results

5.4.2.1 Comparison with EM algorithm

In Fig. 5.2, we compare the MSE of the proposed iterative channel estimation algorithm based on soft decisions FAS-output under one iteration given in (5.30) to the ML estimators and the EM algorithm with two iterations described in Sections 5.3.2 and 5.3.3 respectively for an overdetermined system with $N = 8$ and $n = 64$. We show that the same performance can be achieved by the proposed algorithm in just one iteration.
In Fig. 5.3, we consider a determined system with $N = n = 64$. It is shown that the EM algorithm exhibits no improvement compared to ML training-based estimation in such configuration. However, the proposed algorithm based on soft decision FAS output presents a gain over the ML training-based one beyond SNR=11dB. This gain is of about 2dB at 10^{-3} MSE. The second proposed algorithm based on soft decision FAS-SAC output outperforms EM algorithm and soft decision FAS-output algorithm over the whole SNR range. It achieves the same MSE as the full data based estimation beyond SNR=19dB.

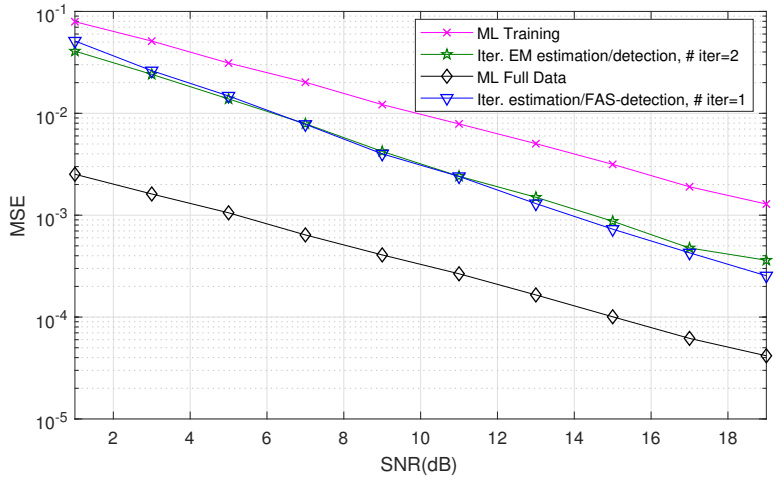

Figure 5.2: MSE versus SNR with uncoded 4-QAM, $N = 8$, $n = 64$, $T_p = 16$ and $T = 512$, (soft decision FAS output-based scheme).

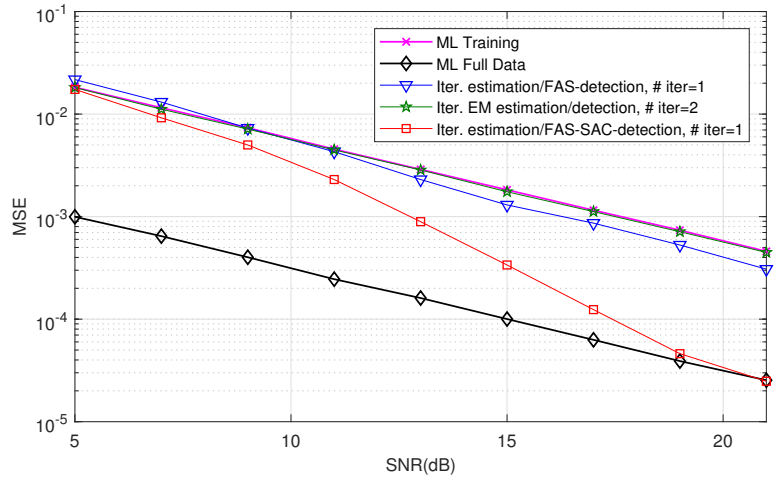

Figure 5.3: MSE versus SNR with uncoded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$, (soft decision FAS and FAS-SAC output-based schemes).

5.4.2.2 Comparison with theoretical bounds

In Fig. 5.4, the overdetermined case with $N = 8$ and $n = 64$ is considered and we show that the performance of the proposed algorithm based on soft decision FAS-output is close to its CRB. The same holds in Fig. 5.5 where the determined system is considered and the soft decision FAS-SAC-output based channel estimation algorithm is compared to its CRB.

Fig. 5.6 shows the validity of the asymptotic MSEs defined in (5.65) for both proposed channel estimation algorithms based on hard-decision detection output (FAS and FAS-SAC respectively).

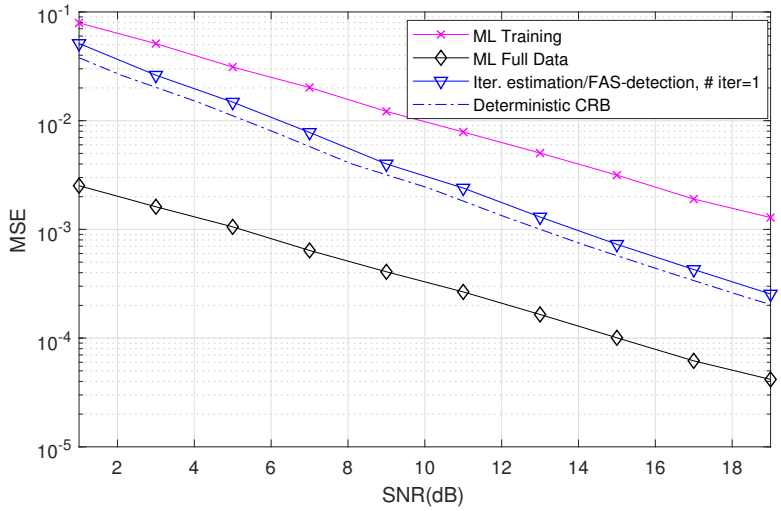

Figure 5.4: MSE versus SNR with uncoded 4-QAM, $n = 64$, $N = 8$, $T_p = 16$ and $T = 512$ (overdetermined system, soft decision FAS output-based scheme).

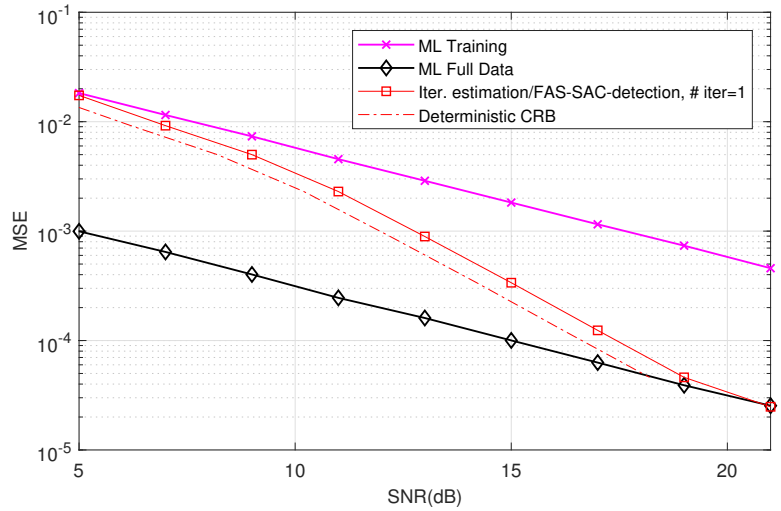

Figure 5.5: MSE versus SNR with uncoded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$ (determined system, soft decision FAS-SAC output-based scheme).

5.4.2.3 Superposition of MSE and BER of hard decision output-based algorithms

Fig. 5.7 and Fig. 5.8 represent the MSE and BER performance of the iterative channel estimation/detection schemes based on FAS and FAS-SAC hard decisions outputs respectively for the determined system with $N = n = 64$. It can be seen that the MSE performance can be improved for an increased number of iterations between channel estimation and detection for both schemes. It can also be seen that

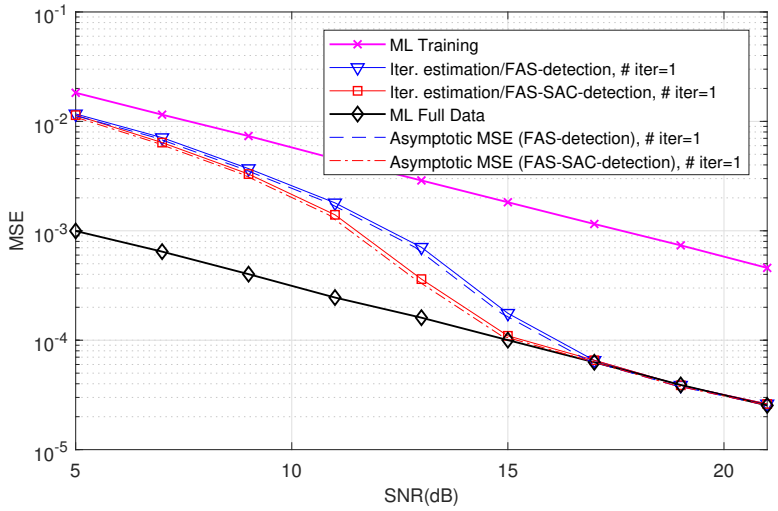

Figure 5.6: MSE versus SNR with uncoded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$ (hard decision FAS and FAS-SAC outputs-based schemes).

	Computational cost
ML training	$\mathcal{O}(nNT_p)$
ML full data	$\mathcal{O}(nNT)$
EM	$\mathcal{O}(\max((T_d + NQ)n^2), ((T + N)nNQ))$
Proposed algorithms (FAS-based)	$\mathcal{O}(T_d Q N^3)$
Proposed algorithms (FAS-SAC-based)	$\mathcal{O}(T_d(2 - Z_\eta)QN^3)$

Table 5.1: Computational cost with the interior point (iteration number: Q).

with just two iterations of the channel estimation/detection procedure, we can get a BER close to perfect channel detection BER. We show that the FAS and FAS-SAC based-iterative schemes achieve 10^{-3} BER within 0.5dB and 0.9dB of the perfect channel knowledge respectively.

In Fig. 5.9 and Fig. 5.10, we show that the proposed channel estimation/detection schemes are efficient in underdetermined systems where $N = 64$ and $n = 50$. It can be confirmed as the determined system that the MSE performance can be improved increasing the number of iterations. It is the same for the BER where we get a gain of about 2.7dB and 2.8dB over the initial channel estimation detection for FAS and FAS-SAC hard decisions-based channel estimation/detection schemes respectively at the second iteration.

5.4.3 Complexity analysis

We now compare the computational complexity of proposed semi-blind channel estimation with the ML estimators and the EM algorithms detailed in Section 5.3. Calculation of ML training-based channel estimation consists of matrix multipli-

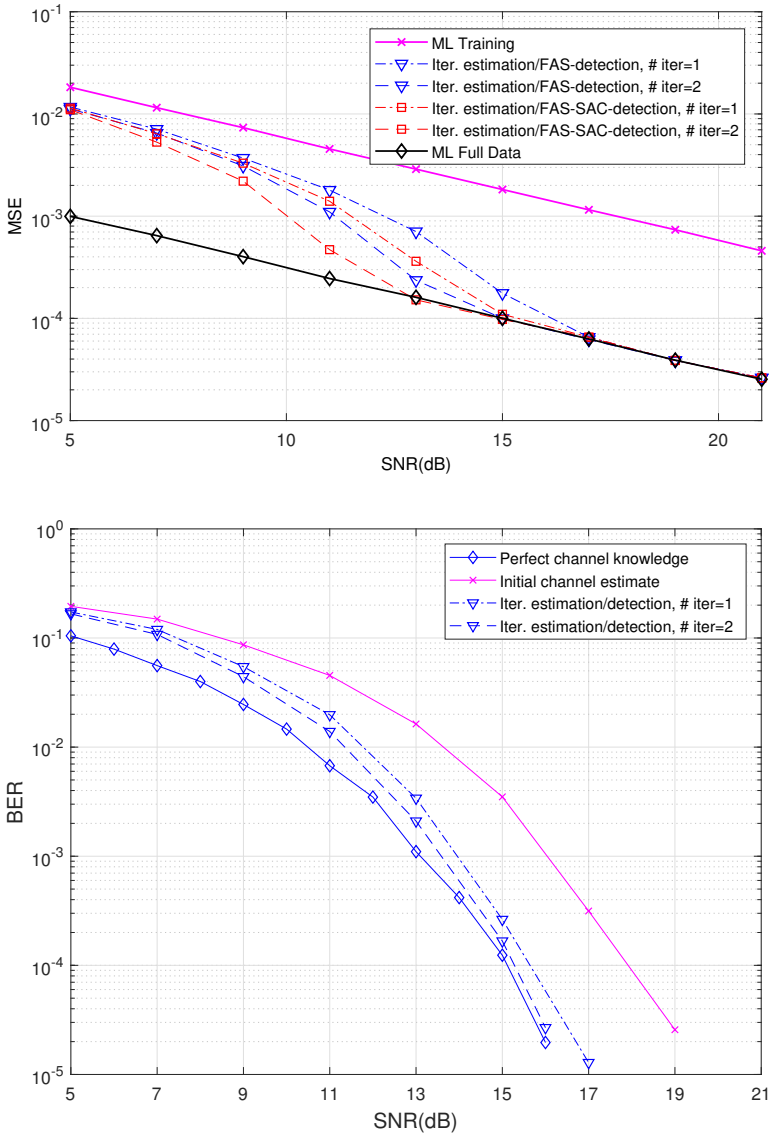

Figure 5.7: MSE versus SNR (hard decision FAS and FAS-SAC output-based schemes), BER versus SNR (hard decision FAS output-based schemes) with uncoded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$.

cations with dominant factor of $\mathcal{O}(nNT_p)$ and a matrix inversion with complexity $\mathcal{O}(N^3)$. Therefore, the whole complexity order is $\mathcal{O}(nNT_p)$. Similarly we can show that the complexity of full data based channel estimation is about $\mathcal{O}(nNT)$. Let denote Q the number of iterations taken into account in the iterative algorithms. We get that the EM channel estimation algorithm has a complexity of order $\mathcal{O}(\max((T_d + NQ)n^2), ((T + N)nNQ))$. The complexity orders of proposed algorithms based on FAS soft decision output and FAS hard decision output are the

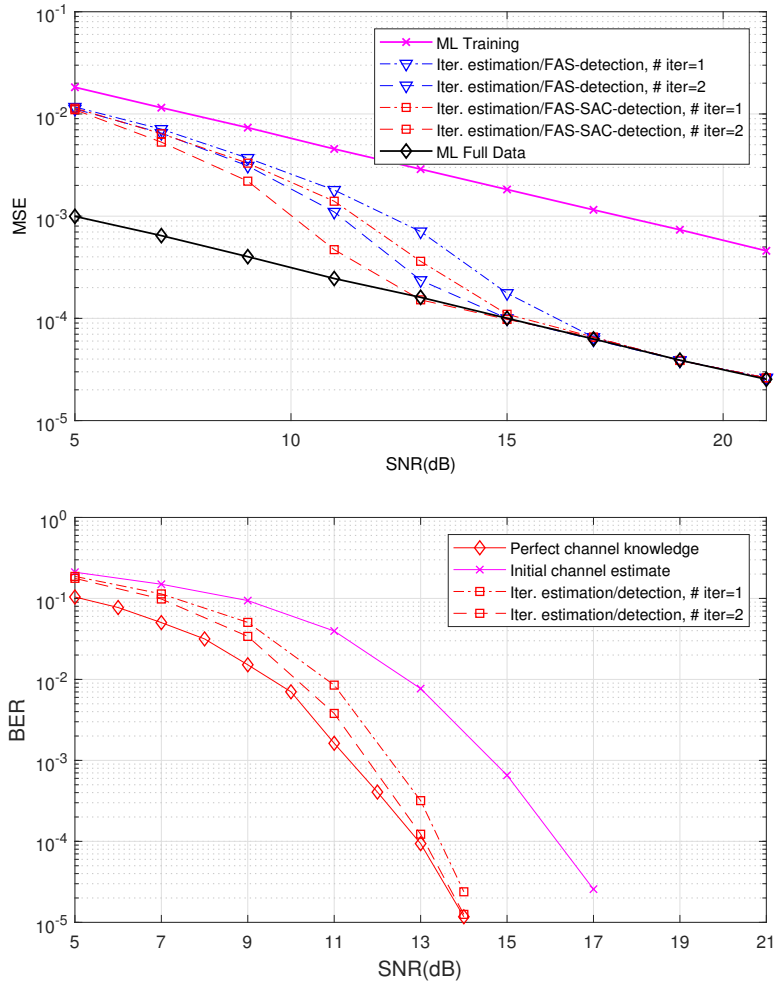

Figure 5.8: MSE versus SNR (hard decision FAS and FAS-SAC output-based schemes), BER versus SNR (hard decision FAS-SAC output-based scheme) with uncoded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$.

same and equal $\mathcal{O}(T_d Q N^3)$. The channel estimation based on FAS SAC with both soft and hard decisions outputs represents a complexity order of $\mathcal{O}(T_d(2 - Z_\eta)QN^3)$. Finally, it can be mentioned that all the proposed iterative algorithms represent the same order of complexity as EM algorithm. The computational complexities of the different algorithms are reported in Table 5.1.

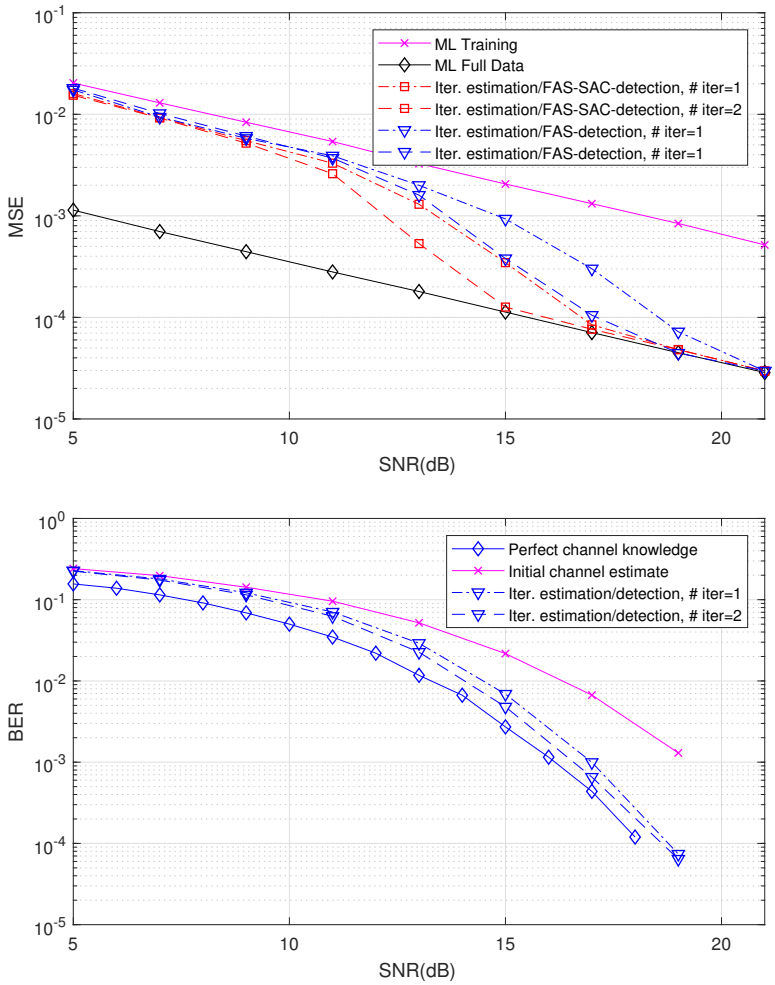

Figure 5.9: MSE versus SNR (hard decision FAS and FAS-SAC output-based schemes), BER versus SNR (hard decision FAS output-based scheme) with uncoded 4-QAM, $n = 64$, $N = 50$, $T_p = 64$ and $T = 1280$.

5.5 Channel estimation for large-scale FEC-coded MIMO systems

5.5.1 Channel estimation algorithm combined with FAS-MAE

Contrary to previous section which considered uncoded systems, we propose to take into account the FEC constraint in uplink multiuser large-scale MIMO systems (Fig. 5.11) and to feed the channel estimation with the FEC decoder output. We propose to combine the coded iterative receiver FAS-MAE described in previous chapter with a channel estimation block as shown in Fig. 5.12.

As the initial channel estimate for the first iteration is not perfect, the variance of real-valued FAS-detected vector taken into account when the knowledge of the

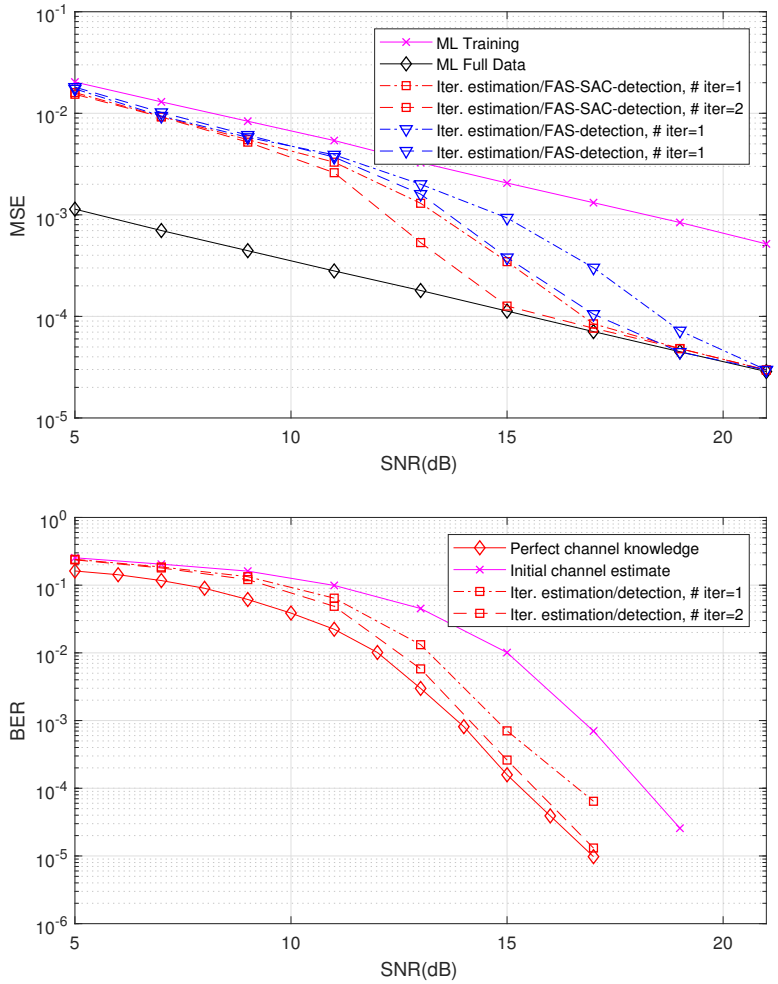

Figure 5.10: MSE versus SNR (hard decision FAS and FAS-SAC output-based schemes), BER versus SNR (hard decision FAS-SAC output-based scheme) with uncoded 4-QAM, $n = 64$, $N = 50$, $T_p = 64$ and $T = 1280$.

channel is perfect is changed here as follows:

$$\sigma^2_{\underline{\hat{x}}} = \sum_{k=0}^{2n-2} \binom{2N}{k} \left(\frac{1}{p}\right)^{2N-k} \left(\frac{p-1}{p}\right)^{k} \frac{2n(2\sigma^2)}{2n-k-1}. \tag{5.75}$$

To update the channel estimation, we propose two approaches that exploit the probability vectors $\boldsymbol{P}_j$ delivered by the decoder: the first based on hard-decisions and the second on soft-decisions.

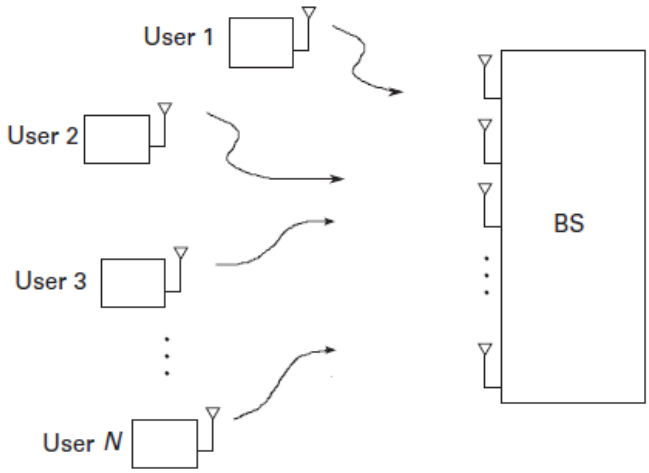

Figure 5.11: Uplink multiuser MIMO system

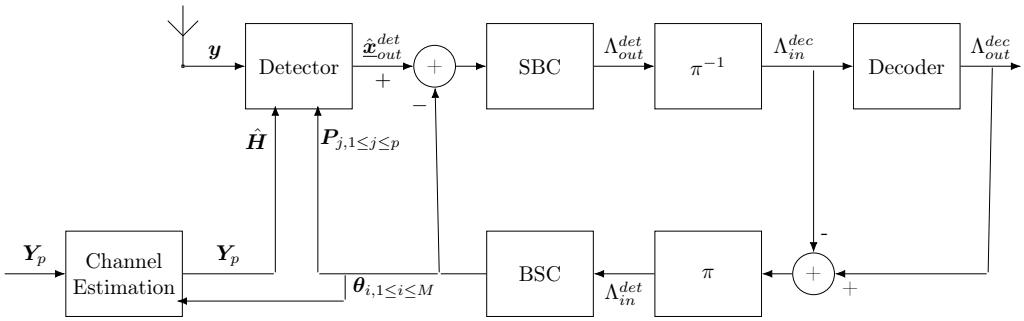

Figure 5.12: Turbo joint channel estimation and FAS-MAE detection scheme

5.5.1.1 Hard decision-based estimation

Let $\tilde{\boldsymbol{X}}_d$ be the $n \times T_p$ hard-decision matrix computed from BSC output. Let $\theta_{i,k} = Pr\left(x_k = \beta_i | \Lambda_{in}^{det}\right)$ computed from $\boldsymbol{P}_j$. The hard decision on x_k is defined by

$$\tilde{X}_{d,k} = \beta_{i^*} \text{ with } i^* = \arg \max_{1\leq i\leq M} \theta_{i,k}. \tag{5.76}$$

We then propose to update the channel estimation by

$$\hat{\boldsymbol{H}}_i = \left(\boldsymbol{Y}_p \boldsymbol{X}_p^H + \boldsymbol{Y}_d \tilde{\boldsymbol{X}}_d^H\right)\left(\boldsymbol{X}_p \boldsymbol{X}_p^H + \tilde{\boldsymbol{X}}_d \tilde{\boldsymbol{X}}_d^H\right)^{-1}. \tag{5.77}$$

5.5.1.2 Soft decision-based estimation

So as to preserve the information delivered by the FEC decoder, we propose to use $\boldsymbol{\Theta}(t) = (\theta_{i,k}(t))_{1\leq i\leq M, 1\leq k\leq N}$ the probabilities matrix at time t to compute soft-

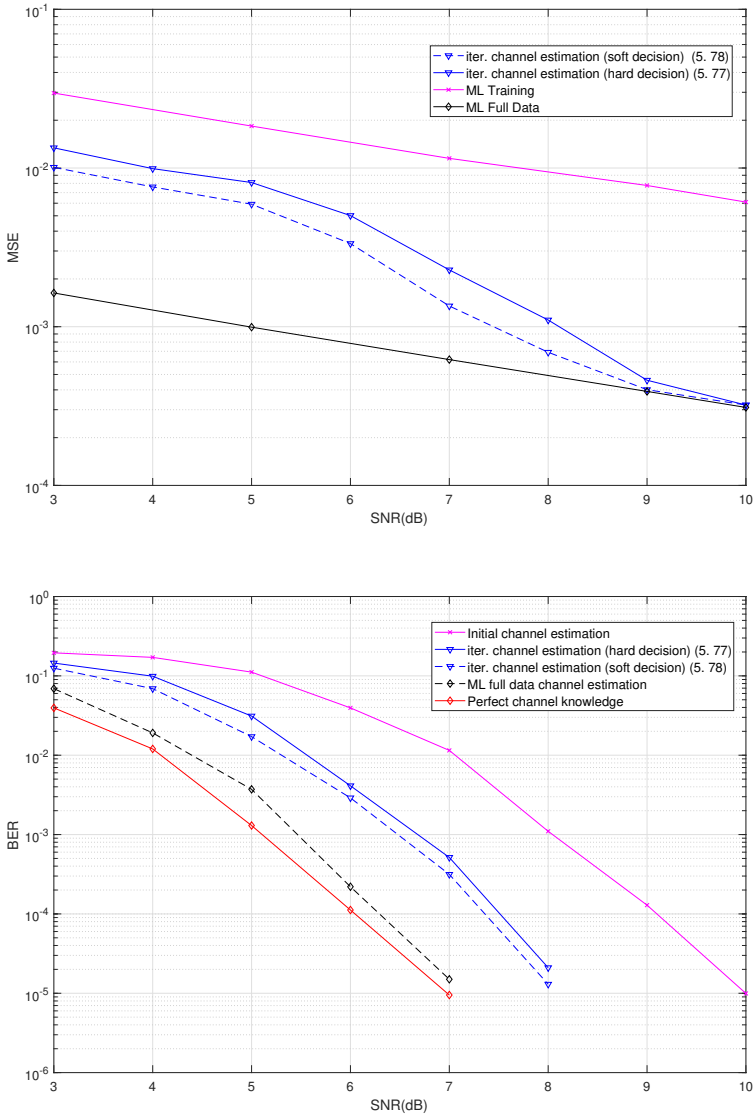

Figure 5.13: MSE versus SNR with coded 4-QAM, $n = N = 64$, $T_p = 64$ and $T = 1280$.

decisions and update the FAS-MAE based channel estimation as follows:

$$\hat{\boldsymbol{H}}_i = \left(\boldsymbol{Y}_p \boldsymbol{X}_p^H + \sum_{t=T_p+1}^{T} \boldsymbol{y}(t) \boldsymbol{\beta}^H \boldsymbol{\Theta}(t) \right) \times \left(\boldsymbol{X}_p \boldsymbol{X}_p^H + \sum_{t=T_p+1}^{T} \boldsymbol{\Theta}^T(t) \boldsymbol{\beta}\, \boldsymbol{\beta}^H \boldsymbol{\Theta}(t) \right)^{-1} \quad (5.78)$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_M]^T$ is the modulation vector.

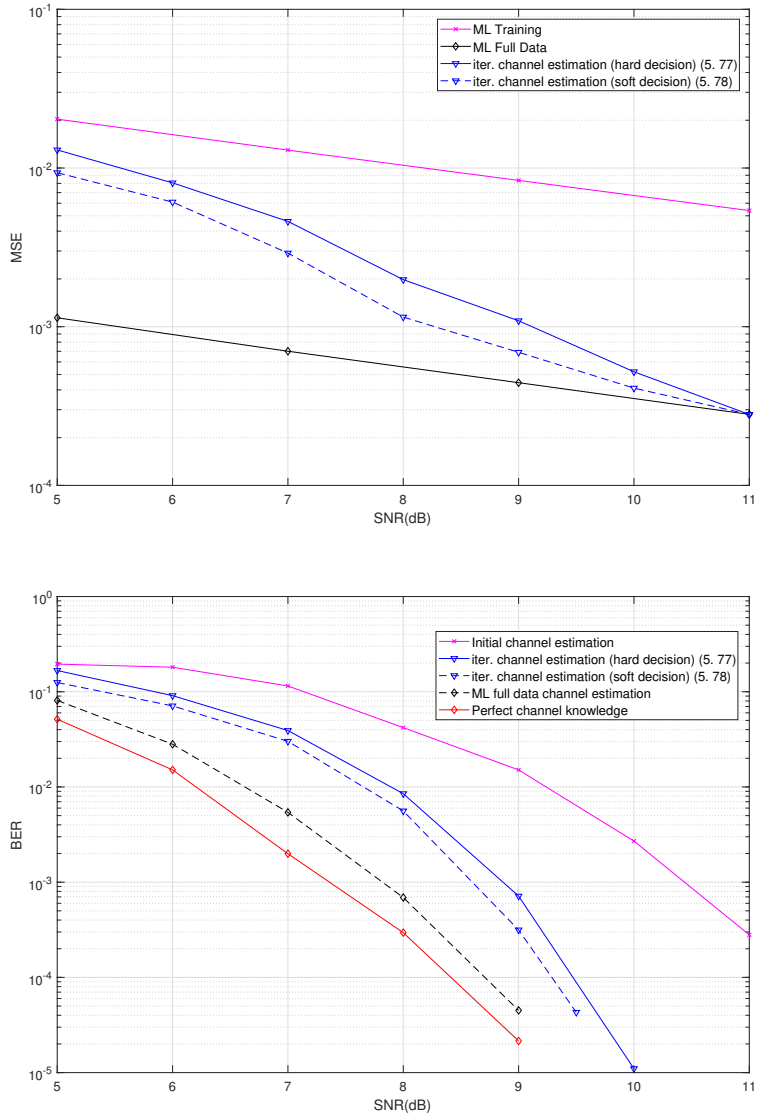

Figure 5.14: MSE versus SNR with coded 4-QAM, $n = 64$, $N = 50$, $T_p = 64$ and $T = 1280$.

5.5.2 Simulation results

We consider coded systems with 4-QAM and convolutional code (CC) whose polynomials in octal are $(13, 15)$ (code rate equal to 0.5). A frame consists of 1216 short codewords of length equal to 256 coded bits, $T_p = 64$, which makes $T = 1280$.

In Fig. 5.13 ($n = N = 64$) and Fig. 5.14 ($n = 64$, $n = 50$), our purpose is to evaluate the channel estimate accuracy achieved by the turbo receiver (soft decision-based, after four iterations) as compared to the ML training-based estimation and to the lower ML bound referred to as "ML full-data". Let us remind that ML full-data

assumes that the whole frame (data and pilot) is known at the receiver and used as training sequence. We observe that the proposed scheme achieves the same MSE as the ML full-data beyond $SNR = 9$dB and $SNR = 11$dB for the determined and the underdetermined systems respectively.

The proposed semi-blind estimation outperforms the ML training-based estimation, with a gain of about $3dB$-$4dB$ at $MSE = 10^{-2}$. These observations support the efficiency of the proposed estimation.

Then, we compare the two proposed strategies to update the EM channel estimation (hard-decision based and soft-decision based) and we study their impact on the turbo receiver error rate performance. We have also plotted the performance with perfect channel state information, ML full-data estimation and ML training-based estimation.

In Fig. 5.13, $n = N = 64$. Compared to the perfect CSI knowledge lower bound, at $BER = 10^{-3}$, we observe a loss of 0.3dB for ML-full data, 1.75dB for soft decision-based channel estimation, 1.9dB for hard decision-based channel estimation and 3dB for ML-training. The difference between soft and hard decisions-based versions is about 0.15dB for the whole SNR values. In Fig. 5.14, we consider the underdetermined case with $N = 64$ and $n = 50$. We get roughly the same losses compared to the perfect CSI knowledge as in the determined case. We can thus deduce that the use of detected data to refine the channel estimation is efficient as it enables to improve the receiver performance by 1.6-1.9dB depending on the approach (either hard decision or soft decision-based).

5.6 Conclusion

In this chapter we have addressed the problem of imperfect CSI and we have proposed semi-blind channel estimation in large-scale MIMO in both uncoded and FEC-coded systems with finite alphabets assuming limited pilot sequence length. We have proposed channel estimation schemes based on soft and hard decisions output of FAS and FAS-SAC algorithms. We have shown that taking into account a number of pilot sequences equal to the number of users is sufficient. Theoretical studies for both algorithms are established and we have determined the CRBs when soft decisions are considered and the asymptotic MSEs when hard decisions are used. Simulation results showed their validity. Then, we proposed a turbo FAS-based detection receiver which combines estimation, detection and FEC decoding and we defined two ways of updating the FAS-based channel estimation from FEC decoder output. Simulations showed the efficiency of the proposed scheme which performs close to the ML full-data lower bound, with a superiority of the ones based on hard decisions in the uncoded case which is not the case when coded systems are considered. The whole work is done assuming a slow fading channel. Future work will focus how the proposed schemes can be extended to frequency-selective channels.

Chapter 6

Conclusions & Perspectives

Contents

6.1 Conclusions

This thesis is motivated by the new opportunities in large-scale MIMO systems and the different challenges to be addressed to make them operational. For example, algorithms and techniques which are known to work well with a small number of antennas may not support the passage to high dimensions. Therefore, new and alternative approaches are needed. Also, in addition to increased rate and diversity gains, large dimensions bring other advantages (e.g., channel hardening, which can be exploited to achieve low-complexity signal processing) which do not come with smaller-dimension systems. Bringing out such large MIMO opportunities, issues, and solutions is our key objective. The PhD manuscript can be summarized as follows:

In Chapter 1, we have first introduced the MIMO communication in the context of both point-to- point as well as multiuser scenarios. We have shown the advantages of such systems over the SISO communications. Different performance parameters such as spectral efficiency and error probability have been presented. It has been mentioned that the more the number of antennas the best the system. Large-scale MIMO systems have been introduced as well as some advantages. The channel hardening effect that happens in large dimensional MIMO channels and can simplify the implementation of large-scale MIMO systems significantly, has been described. Then, we have presented some challenges when such sytems are deployed such as the necessity of low complexity signal processing algorithms to deal with very high number of antennas. Channel estimation has also been evoked with a special focus on pilot contamination channel in multi-cell communications.

In Chapter 2, we have presented the state-of-the-art of detection algorithms in MIMO systems and large-scale MIMO systems. We have detailed the different algorithms and we have showed their weak and strong points. We have mentioned that the classical algorithms such as ML-based algorithms and linear detection algorithms are inadequate for large-scale systems, on account of either poor performance

or high complexity. Then, we have presented local search-based algorithms that aim to improve the performance of low-complexity algorithms like linear detection algorithms while preserving the same order of complexity. These algorithms referred to as RTS and LAS have been used as benchmark in Chapters 3 and 4.

In Chapter 3, dedicated to large-scale MIMO detection, we have considered Compressive sensing (CS) techniques to propose new detection algorithms with low complexity. We have considered finite alphabet signals and we have exploited their simplicity property to propose an efficient algorithm. Then recovery and detection algorithms have been designed for both noiseless and noisy MIMO systems. The noiseless case has been theoretically investigated to obtain the necessary conditions of successful recovery. The proposed recovery scheme has been then extended to the large-scale MIMO systems and a theoretical study of the detection output statistics has been successfully carried out. We have showed that the proposed algorithm has the same order of complexity as the low-complexity algorithms of the state-of-art (MMSE, LAS, RTS) and outperform them (significant gains over LAS and MMSE).

In Chapter 4, our purpose was to integrate the proposed algorithm in Chapter 3 in an iterative procedure for both uncoded and coded cases in order to improve its performance. First, we have considered the uncoded case and we have proposed an iterative successive interference scheme based on the shadow area principle, with parameters theoretically fixed based on the original proposed detection scheme output statistics. Simulation results show that the proposed scheme improves significantly the performance of the original algorithm and preserves the same order of complexity. Then, we have proposed to integrate the original scheme in a turbo-like iterative receiver. Two turbo detection schemes have been proposed. The first scheme is an ML-like iterative receiver with search space limited to neighbors in order to reduce its complexity. To further reduce the receiver complexity, we have proposed a second scheme which uses the output statistics investigated in previous chapter to feed the decoder with reliable LLRs. A criterion reformulated as a regularized constrained least squares problem has been designed at the detector side to keep symbols near to the decoder output. The regularization term can be seen as the mean of absolute error in function of the probabilities delivered by the decoder.

In Chapter 5, we have considered the case of imperfect channel state information and we have introduced semi-blind iterative least squares channel estimation algorithms fed with the proposed detection algorithms outputs. We have shown that taking into account a number of pilot sequences equal to the number of users is sufficient to obtain a reliable CSI estimate. The EM algorithm works efficiently only in overdetermined systems. The proposed iterative channel estimation/ detection schemes are efficient in both determined or underdetermined systems and we have demonstrated that two iterations are sufficient to perform close to the lower bound (perfect CSI). The coded case has also been treated and an iterative turbo-like channel estimation and detection schemes have been proposed whose performance is promising to expect new pilot sequences design.

6.2 Perspectives

Short-term perspectives

Perspectives that aim to improve the performance of the proposed algorithms in both uncoded and coded cases assuming a non frequency-selective channel are summarized as follows:

First, the proposed detection schemes are based on a criterion which can be resolved by optimization algorithms such as simplex or interior point methods. Our aim is to propose less complex algorithms that resolve the proposed criterion exploiting the channel hardening phenomenon in large-scale MIMO systems detailed in Chapter 1.

Second, as mentioned in Chapter 4, the calculated LLRs in 4.25 that correspond to the output-vector elements equal to the bounds of the finite alphabet, take the same and the highest value compared to the other output-vector elements. However, when low and medium SNR values are considered, some bounded elements are erroneous and lead the system to consider an erroneous symbol as reliable. This fact can degrade the performance of the iterative receiver. Therefore, other LLR calculation strategies should be proposed to deliver more reliable soft decisions to the decoder.

Third, the proposed algorithms could be interfaced with the local search algorithms (LAS and RTS) and Lattice reduction methods to further improve detection performance.

Fourth, in Chapter 4, the theoretical analysis is based on 4-QAM modulation. The extension to higher order modulations should be investigated.

Mid-term perspectives

The main works on the whole PhD are done assuming a non frequency-selective channel. In more practical scenarios, channels can be frequency-selective, causing ISI. Future works could deal with the extension of proposed detection and channel estimation algorithms to frequency-selective channels.

In multi-cell large-scale MIMO systems, usually, the length of the training sequences is not sufficient enough to separate the channels of multi-cell users which results in pilot contamination effect [66]. It is shown that the efficiency of channel estimation in one cell becomes corrupted by the channel between that base station and the users in other cells in an undesirable manner. Chapter 5 studied semi-blind estimation in single cell scenarios with large-scale systems and thus neglected inter-cell interference. Future work could investigate either pilot sequences design combined with proposed receivers or blind estimation techniques based on sparse signal processing. To manage interference at the terminal side, we could study whether it is possible to exploit the simplicity property as well as compressive sensing to propose efficient beamforming techniques.

List of Publications

Journal Papers

Published

- Z. Hajji, A. Aïssa-El-Bey, and K. Amis, "Simplicity-based recovery of finite-alphabet signals for large-scale MIMO systems" *Digital Signal Processing,vol. 80, pp. 70-82, September 2018.* (Chap.3)

Submitted

- Z. Hajji, K. Amis, and A. Aïssa-El-Bey, "Iterative receivers for large-scale MIMO systems with finite-alphabet simplicity-based detection" *submitted to IEEE Transactions on Communications.* (Chap.4).
- Z. Hajji, K. Amis, and A. Aïssa-El-Bey, "Channel estimation with finite-alphabet simplicity-based detection for large-scale MIMO systems" *submitted to IEEE Transactions on Signal Processing.* (Chap.5).

Conference Papers

Published

- Z. Hajji, K. Amis, A. Aïssa-El-Bey, F. Abdelkefi, "Low-Complexity Half-Sparse Decomposition based Detection for massive MIMO Transmission" *Int. Conf. on Commun. and Networking (ComNet), Nov. 2015.*
- Z. Hajji, K. Amis, A. Aïssa-El-Bey, "Turbo Detection Based On Sparse Decomposition For Massive MIMO Transmission", *IEEE Int. Symp. On Turbo Codes and Iterative Inform. Process. (ISTC), Sept. 2016.*
- Z. Hajji, K. Amis, and A. Aïssa-El-Bey, "Turbo detection based on signalsimplicity and compressed sensing for massive MIMO transmission," *IEEE Wireless Communications and Networking Conference (WCNC), Barcelona,Spain, April 2018.*
- Z. Hajji, K. Amis, A. Aïssa-El-Bey, "Joint channel estimation and simplicity-based detection for large-scale MIMO FEC-coded systems", *IEEE Int. Symp. On Turbo Codes and Iterative Inform. Process. (ISTC), Dec. 2018.*

Chapter 7

Appendix

7.1 Generic random matrix

We assume that the components of $\boldsymbol{H}$ are independent, complex circularly-symmetric Gaussian random variables with zero mean and unit variance. We aim to prove that $\underline{\boldsymbol{H}}$ is a generic random matrix. The assumption on the distribution of the components of $\boldsymbol{H}$ ensures that the columns of $\underline{\boldsymbol{H}}$ are symmetrically distributed about the origin. We thus have to prove that $\underline{\boldsymbol{H}}$ is completely general with probability 1, that is to say whatever ℓ, any $\ell \times \ell$ submatrix of $\underline{\boldsymbol{H}}$ has full-rank. This result is given by the following Theorem A.1.

Theorem A.1 *:* Given a complex-valued matrix $\boldsymbol{H}$ and its real-valued transform $\underline{\boldsymbol{H}}$, if $\boldsymbol{H}$ is a generic random matrix with independent circularly-symmetric Gaussian-distributed components, then $\underline{\boldsymbol{H}}$ is completely general with probability 1. Let us prove the theorem by induction. Let $\boldsymbol{H}$ be a complex-valued matrix with independent circularly-symmetric Gaussian-distributed components. Let us define the property $\mathcal{P}_\ell$ by "any $\ell \times \ell$ submatrix of $\underline{\boldsymbol{H}}$ has full-rank with probability 1". Let $\ell = 1$. Then, any 1×1 submatrix of $\underline{\boldsymbol{H}}$ is a real-valued Gaussian variable, that is to say a continuous random variable and the probability that it equals zero is null. $\mathcal{P}_1$ is true. Let us suppose that $\mathcal{P}_\ell$ is true and let us prove that $\mathcal{P}_{\ell+1}$ is also true. Let S a $(\ell+1) \times (\ell+1)$ submatrix of $\underline{\boldsymbol{H}}$. Then, as $\mathcal{P}_\ell$ is true, all minors of $\boldsymbol{S}$ have non-zero determinant with probability 1. Let us compute the determinant of $\boldsymbol{S}$ according to a given row (or column). It corresponds to the linear combination of minors, where due to the independence of the components of $\underline{\boldsymbol{H}}$ (the components of $\boldsymbol{H}$ are circularly-symmetric Gaussian and independent) the weights and the minors form a family of random variables that are mutually independent, continuously distributed and different from zero with probability equal to one. Thus, the determinant of $\boldsymbol{S}$ is a continuous random variable different from zero with probability equal to one.

7.2 Proof of Proposition 4.2.1

The number of non binding constraints of $(P_{SI,2})$ can be seen as the sum of the inactive constraint number and the non binding active constraint number. The probability, that a constraint is inactive, is denoted by p_{in} and corresponds to the probability that $\underline{x}_i \notin \{\alpha_1, \alpha_p\}$, that is to say $p_{in} = \frac{p-2}{p}$. The probability that a constraint is non binding and active is denoted by p_{nba} and corresponds to the probability that either $\underline{x}_i = \alpha_1$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\boldsymbol{x}} - \underline{\boldsymbol{y}})\}_i < 0$, or $\underline{x}_i = \alpha_p$ and $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\boldsymbol{x}} -$

$\underline{\boldsymbol{y}})\}_i > 0\}$. As a constraint cannot be active and inactive, the probability, that a constraint is non binding, equals $p_{nb} = p_{in} + p_{nba}$.

It remains to find the value of $p_{nba} = \Pr(i \in \Omega \backslash \Lambda)$. To that purpose, we focus on the sign of $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i = \{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}(\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{x}}) - \boldsymbol{\zeta})\}_i$, $i \in \Lambda$. As the elements in the set $\mathcal{F} = \{\alpha_1, \alpha_2, .., \alpha_p\}$ are equiprobable and the real-valued matrix channel $\underline{\boldsymbol{H}}$ as well as the noise are Gaussian, we can affirm that the estimated vector $\hat{\underline{\boldsymbol{x}}}$ is a symmetrically erroneous version of the original vector $\underline{\boldsymbol{x}}$. Then, the sign of $\{(\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{x}})\}_i$, $i \in \Lambda$ takes on equiprobable values. Consequently, exploiting the same hypothesis for the channel matrix and the noise we deduce that the sign of $\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i$ can be negative or positive with probability $1/2$. Then, p_{nba} can be decomposed as:

$$\begin{aligned} p_{nba} &= \Pr(\underline{x}_i = \alpha_1) \Pr(\{\tilde{\underline{\boldsymbol{H}}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \leq 0 | \underline{x}_i = \alpha_1) \\ &+ \Pr(\underline{x}_i = \alpha_p) \Pr(\{\underline{\boldsymbol{H}}^T(\underline{\boldsymbol{H}}\hat{\underline{\boldsymbol{x}}} - \underline{\boldsymbol{y}})\}_i \geq 0 | \underline{x}_i = \alpha_p) \\ &= \frac{1}{p} \times \frac{1}{2} + \frac{1}{p} \times \frac{1}{2} = \frac{1}{p}. \end{aligned} \tag{7.1}$$

Consequently, $p_{nb} = p_{in} + p_{nba} = \frac{p-2}{p} + \frac{1}{p} = \frac{p-1}{p}$, Hence $\text{card}(\bar{\Lambda}) \sim \mathcal{B}(2N, \frac{p-1}{p})$.

7.3 Symbol error probability upper-bound

[Proof of Theorem 3.4.3] Let us denote by $\tilde{\underline{x}}_k$ the hard decision taken on $\underline{x}_k$ from the detection output $\hat{\underline{x}}_k$. Then, the symbol error probability is defined by

$$P_s = \Pr(\underline{x}_k \neq \tilde{\underline{x}}_k). \tag{7.2}$$

Considering the assumptions (alphabet and equiprobability), P_s reads

$$P_s = \frac{1}{p} \sum_{i=1}^{p} \sum_{\substack{q=1 \\ i \neq q}}^{p} \Pr(\tilde{\underline{x}}_k = \alpha_q | \underline{x}_k = \alpha_i). \tag{7.3}$$

This probability can be computed by considering a maximum-likelihood decision rule applied on $\hat{\underline{x}}_k$:

$$\begin{aligned} P_s &= \frac{1}{p} \sum_{i=1}^{p} \sum_{\substack{q=1 \\ i \neq q}}^{p} \Pr\left(\bigcap_{\substack{j=1 \\ j \neq q}}^{p} \left((\hat{\underline{x}}_k - \alpha_q)^2 \leq (\hat{\underline{x}}_k - \alpha_j)^2 \right) | \underline{x}_k = \alpha_i \right) \\ &\leq \frac{1}{p} \sum_{i=1}^{p} \sum_{\substack{q=1 \\ i \neq q}}^{p} \Pr\left((\hat{\underline{x}}_k - \alpha_q)^2 \leq (\hat{\underline{x}}_k - \alpha_i)^2 | \underline{x}_k = \alpha_i \right) \\ &\leq \frac{1}{p} \sum_{i=1}^{p} \sum_{\substack{q=1 \\ i \neq q}}^{p} \Pr\left(\hat{\underline{x}}_k \leq \frac{\alpha_i + \alpha_q}{2} | \underline{x}_k = \alpha_i \right). \end{aligned} \tag{7.4}$$

Using Theorem 3.4.2, we can write

$$\begin{aligned}\Pr\left(\underline{\hat{x}}_k \leq \frac{\alpha_i+\alpha_q}{2}|\underline{x}_k=\alpha_i\right) &= \frac{1}{2}\text{erfc}\left(\frac{\alpha_i-\alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\ &+ \frac{1}{2}\text{erfc}\left(\frac{\alpha_p-\alpha_i}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \int_{\alpha_1}^{\frac{\alpha_i+\alpha_q}{2}} \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}}\exp\left(-\frac{(x-\alpha_i)^2}{2\sigma^2_{\underline{\hat{x}}}}\right)dx.\end{aligned} \tag{7.5}$$

After computation, we get

$$\Pr\left(\underline{\hat{x}}_k \leq \frac{\alpha_i+\alpha_q}{2}|\underline{x}_k=\alpha_i\right) = \frac{1}{2}\text{erfc}\left(\frac{\alpha_p-\alpha_i}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \frac{1}{2}\text{erfc}\left(\frac{\alpha_i-\alpha_q}{2\sqrt{2}\sigma_{\underline{\hat{x}}}}\right). \tag{7.6}$$

Therefore, P_s can be upper-bounded by

$$P_s \leq \frac{1}{2p}\sum_{i=1}^{p}\sum_{\substack{q=1\\ i\neq q}}^{p}\text{erfc}\left(\frac{\alpha_i-\alpha_q}{2\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \frac{p-1}{2p}\sum_{i=1}^{p}\text{erfc}\left(\frac{\alpha_p-\alpha_i}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right). \tag{7.7}$$

At high SNR, due to the decreasing rate of erfc, the terms depending on differences between adjacent symbols are predominant and the following approximation is asymptotically tight

$$P_s \approx \frac{p-1}{p}\text{erfc}\left(\frac{\alpha_2-\alpha_1}{2\sqrt{2}\sigma_{\underline{\hat{x}}}}\right). \tag{7.8}$$

7.4 Proof of equations (4.5) and (4.6) of Theorem 4.2.1

In this appendix, we aim to prove the expression of Z_η and of Y_η given in (4.5) and (4.6) respectively.

7.4.1 Proof the expression of Z_η given by equation (4.5)

We first remind us about the distribution of the output detector vector which reads:

$$\begin{aligned}f_{\underline{\hat{x}}_k}(x) = \frac{1}{p}\sum_{\ell=1}^{p}\Bigg(&\frac{1}{2}\text{erfc}\left(\frac{\alpha_\ell-\alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\delta_{\alpha_1}(x) + \frac{1}{2}\text{erfc}\left(\frac{\alpha_p-\alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\delta_{\alpha_p}(x) \\ &+ \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}}\exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma^2_{\underline{\hat{x}}}}\right)1_{[\alpha_1,\alpha_p]}(x)\Bigg),\end{aligned} \tag{7.9}$$

with $\sigma^2_{\underline{\hat{x}}} = \frac{2n\sigma^2}{2n-2N(\frac{p-1}{p})-1}$. Let us denote by Z_η the probability that a source be decided after the first iteration, that is to say the probability that $\underline{\hat{x}}_k$ be reliable.

We assume that η is small enough. Then

$$
\begin{aligned}
Z_\eta &= \Pr((\cup_{i=1}^p |\underline{\hat{x}}_k - \alpha_i| < \eta)) = \sum_{i=1}^p \Pr(|\underline{\hat{x}}_k - \alpha_i| < \eta) = \sum_{i=1}^p \int_{|x-\alpha_i|<\eta} f_{\underline{\hat{x}}_k}(x)dx \qquad (7.10)\\
&= \sum_{i=1}^p \int_{|x-\alpha_i|<\eta} \left(\frac{1}{2p} \sum_{\ell=1}^p \text{erfc}\left(\frac{\alpha_\ell - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \delta_{\alpha_1}(x) + \frac{1}{2p} \sum_{\ell=1}^p \text{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \delta_{\alpha_p}(x) \right) dx \\
&+ \sum_{\ell=1}^p \int_{|x-\alpha_i|<\eta} \frac{1}{p\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) 1_{[\alpha_1,\alpha_p]}(x)dx \\
&= \frac{1}{2p} \sum_{\ell=1}^p \text{erfc}\left(\frac{\alpha_\ell - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) + \frac{1}{2p} \sum_{\ell=1}^p \text{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) \\
&+ \sum_{i=1}^p \sum_{\ell=1}^p \int_{|x-\alpha_i|<\eta} \frac{1}{p\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) 1_{[\alpha_1,\alpha_p]}(x)dx.
\end{aligned}
$$

For $i = 1$,

$$
\begin{aligned}
\int_{|x-\alpha_1|<\eta} \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) 1_{[\alpha_1,\alpha_p]}(x)dx &= \int_{\alpha_1}^{\alpha_1+\eta} \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) dx \\
= \int_{\frac{\alpha_1-\alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}}}^{\frac{c_1-c_\ell+\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}} \frac{1}{\sqrt{\pi}} \exp(-t^2)dt &= -\frac{1}{2}\text{erfc}\left(\frac{\alpha_1 - \alpha_\ell + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) + \frac{1}{2}\text{erfc}\left(\frac{\alpha_1 - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right).
\end{aligned}
$$

For $i = p$,

$$
\int_{|x-\alpha_p|<\eta} \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) 1_{[\alpha_1,\alpha_p]}(x)dx = -\frac{1}{2}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) + \frac{1}{2}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right).
$$

For $i \in [|1, p|]$,

$$
\int_{|x-\alpha_i|<\eta} \frac{1}{\sqrt{2\pi}\sigma_{\underline{\hat{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\underline{\hat{x}}}^2} \right) 1_{[\alpha_1,\alpha_p]}(x)dx = -\frac{1}{2}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right) + \frac{1}{2}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}} \right).
$$

Using the equality $\text{erfc}\,(-x) = 2 - \text{erfc}\,(x)$, we obtain

$$\begin{aligned}
Z_\eta &= \Pr\left((\cup_{i=1}^{p}|\underline{\hat{x}}_k - \alpha_i| < \eta)\right) \\
&= \frac{1}{2p}\sum_{\ell=1}^{p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \frac{1}{2p}\sum_{\ell=1}^{p}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&+ \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_1 - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_1}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&- \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&- \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&+ \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_i - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_i + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&= \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_1 - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_p - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&- \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_i - \alpha_\ell - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&+ \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_i - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{\alpha_\ell - \alpha_i + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \quad (7.11)
\end{aligned}$$

As we can define $\alpha_\ell - \alpha_i = (\ell - i)\Delta$, we have,

$$\begin{aligned}
Z_\eta &= \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{(\ell-1)\Delta - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{\ell=1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{(p-\ell)\Delta - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&- \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{(i-\ell)\Delta + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \sum_{i=2}^{p-1}\sum_{\ell=1}^{i}\frac{1}{2p}\text{erfc}\left(\frac{(i-\ell)\Delta - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \\
&+ \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{(\ell-i)\Delta - \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \sum_{i=2}^{p-1}\sum_{\ell=i+1}^{p}\frac{1}{2p}\text{erfc}\left(\frac{(\ell-i)\Delta + \eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \quad (7.12)
\end{aligned}$$

that is to say

$$
\begin{aligned}
Z_\eta &= \sum_{\ell=1}^{p-1} \frac{p-\ell}{p} \mathrm{erfc}\left(\frac{\ell\Delta-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}})\right) - \frac{p-2}{2p}\mathrm{erfc}\left(\frac{\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) \qquad (7.13)\\
&- \sum_{\ell=1}^{p-2} \frac{p-1-\ell}{p}\mathrm{erfc}\left(\frac{\ell\Delta+\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \frac{1}{2}\mathrm{erfc}\left(\frac{-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right).\\
&= \sum_{\ell=0}^{p-1}\left(\frac{p-\ell}{p}\mathrm{erfc}\left(\frac{\ell\Delta-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \frac{p-1-\ell}{p}\mathrm{erfc}\left(\frac{\ell\Delta+\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\right) \qquad (7.14)\\
&+ \left(\frac{p-1}{p} - \frac{p-2}{2p}\right)\mathrm{erfc}\left(\frac{\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) + \left(\frac{1}{2}-1\right)\mathrm{erfc}\left(\frac{-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right).
\end{aligned}
$$

Finally, after simplifications, we obtain

$$
\begin{aligned}
Z_\eta &= \sum_{\ell=0}^{p-1}\frac{1}{p}\left((p-\ell)\mathrm{erfc}\left(\frac{\ell\Delta-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - (p-\ell-1)\mathrm{erfc}\left(\frac{\ell\Delta+\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\right) \qquad (7.15)\\
&+ \frac{1}{2}\left(\mathrm{erfc}\left(\frac{\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right) - \mathrm{erfc}\left(\frac{-\eta}{\sqrt{2}\sigma_{\underline{\hat{x}}}}\right)\right).
\end{aligned}
$$

7.4.2 Proof the expression of Y_η given by equation (4.6)

Let us now compute the variance of the elements of $\mathcal{A}$ denoted by Y_η and defined as:

$$
\begin{aligned}
Y_\eta &= \mathrm{var}(\tilde{\underline{x}}_k \mid k \in \mathcal{A}) \qquad (7.16)\\
&= \mathbb{E}\left[\tilde{\underline{x}}_k^2 | k \in \mathcal{A}\right] - \mathbb{E}\left[\tilde{\underline{x}}_k | k \in \mathcal{A}\right]^2.
\end{aligned}
$$

Focusing on the first term of Eq. (7.16) we get:

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\underline{x}}_k | k \in \mathcal{A}\right] &= \frac{1}{\Pr(k \in \mathcal{A})}\sum_{\ell=1}^{p}\alpha_\ell \Pr(|\hat{\underline{x}}_k - \alpha_\ell| < \eta) \qquad (7.17)\\
&= \frac{1}{Z_\eta}\sum_{\ell=1}^{p}\alpha_\ell \Pr(|\hat{\underline{x}}_k - \alpha_\ell| < \eta)\\
&= \frac{1}{Z_\eta}\sum_{\ell=1}^{p}\alpha_\ell \int_{|x-\alpha_\ell|<\eta} f_{\hat{\underline{x}}_k}(x)dx
\end{aligned}
$$

As the distribution of $\underline{\hat{\boldsymbol{x}}}$ is an even function and the real constellation $\mathcal{F} = \{\alpha_1, \alpha_2, .., \alpha_p\}$ is symmetric with respect to the origin, we get $\mathbb{E}\left[\tilde{\underline{\boldsymbol{x}}}_k | k \in \mathcal{A}\right] = 0$. The second term of Eq. (7.16) is computed as:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\underline{x}}_k^2 | k \in \mathcal{A}\right] &= \frac{1}{Z_\eta} \sum_{m=1}^{p} \alpha_m^2 \Pr(|\hat{\underline{x}}_k - \alpha_m| < \eta) && (7.18)\\
&= \frac{1}{Z_\eta} \sum_{m=1}^{p} \alpha_m^2 \int_{|x-\alpha_m|<\eta} f_{\hat{\underline{x}}_k}(x)dx && (7.19)\\
&= \frac{1}{Z_\eta} \sum_{m=1}^{p} \frac{\alpha_m^2}{2p} \sum_{\ell=1}^{p} \mathrm{erfc}\left(\frac{\alpha_\ell - \alpha_1}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) \int_{|x-\alpha_m|<\eta} \delta_{\alpha_1}(x)dx \\
&\quad + \frac{1}{Z_\eta} \sum_{m=1}^{p} \frac{\alpha_m^2}{2p} \sum_{\ell=1}^{p} \mathrm{erfc}\left(\frac{\alpha_p - \alpha_\ell}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) \int_{|x-\alpha_m|<\eta} \delta_{\alpha_p}(x)dx \\
&\quad + \frac{1}{Z_\eta} \sum_{m=1}^{p} \frac{\alpha_m^2}{p} \sum_{\ell=1}^{p} \int_{|x-\alpha_m|<\eta} \frac{1}{\sqrt{2\pi}\sigma_{\hat{\underline{x}}}} \exp\left(-\frac{(x-\alpha_\ell)^2}{2\sigma_{\hat{\underline{x}}}^2}\right) 1_{[\alpha_1,\alpha_p]}(x)dx
\end{aligned}$$

Following the same approach as for Z_η, we finally get:

$$\begin{aligned}
Y_\eta &= \mathbb{E}\left[\tilde{\underline{x}}_k^2 | k \in \mathcal{A}\right] && (7.20)\\
&= \frac{\Delta^2}{Z_\eta} \sum_{\ell=1}^{p-1} \frac{\ell^2}{p} \left((p-\ell)\mathrm{erfc}\left(\frac{\ell\Delta - \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right) - (p-\ell-1)\mathrm{erfc}\left(\frac{\ell\Delta + \eta}{\sqrt{2}\sigma_{\hat{\underline{x}}}}\right)\right).
\end{aligned}$$

Bibliography

[1] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-blast: an architecture for realizing very high data rates over the rich-scattering wireless channel," in 1998 URSI International Symposium on Signals, Systems, and Electronics. Conference Proceedings (Cat. No.98EX167), Oct 1998, pp. 295–300. (Cited on pages 6, 8 and 24.)

[2] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," Wireless Personal Communications, pp. 311–335, Mar 1998. [Online]. Available: https://doi.org/10.1023/A:1008889222784 (Cited on page 7.)

[3] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," IEEE Transactions on Information Theory, vol. 49, no. 5, pp. 1073–1096, May 2003. (Cited on page 8.)

[4] H. Jafarkhani, Space-Time Coding: Theory and Practice, 1st ed. New York, NY, USA: Cambridge University Press, 2010. (Cited on page 9.)

[5] I. E. Telatar, "Capacity of multi-antenna gaussian channels," EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS, vol. 10, pp. 585–595, 1999. (Cited on page 9.)

[6] C. B. Peel, "On "dirty-paper coding"," IEEE Signal Processing Magazine, vol. 20, no. 3, May 2003. (Cited on page 9.)

[7] F. Gotze and A. Tikhomirov, "Rate of convergence in probability to the marchenko-pastur law," Bernoulli, vol. 10, no. 3, pp. 503–548, 06 2004. (Cited on page 13.)

[8] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," IEEE Transactions on Information Theory, vol. 50, no. 9, pp. 1893–1909, Sept 2004. (Cited on pages 13 and 18.)

[9] M. Chang and W. Chang, "Maximum-likelihood detection for mimo systems based on differential metrics," IEEE Transactions on Signal Processing, vol. 65, no. 14, pp. 3718–3732, July 2017. (Cited on page 18.)

[10] 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. (Cited on pages 18 and 22.)

[11] A. Elghariani and M. Zoltowski, "Successive interference cancellation for large-scale mimo ofdm," in 2015 IEEE International Conference on

Electro/Information Technology (EIT), May 2015, pp. 657–661. (Cited on pages 18 and 24.)

[12] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE, vol. 1, Nov 2002, pp. 424–428 vol.1. (Cited on pages 18, 25 and 46.)

[13] D. Pham, K. R. Pattipati, P. K. Willett, and J. Luo, "A generalized probabilistic data association detector for multiple antenna systems," in 2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577), vol. 6, June 2004, pp. 3519–3522 Vol.6. (Cited on page 18.)

[14] P. Merz and B. Freisleben, "Greedy and local search heuristics for unconstrained binary quadratic programming," Journal of Heuristics, vol. 8, no. 2, pp. 197–213, Mar 2002. [Online]. Available: https://doi.org/10.1023/A:1017912624016 (Cited on page 18.)

[15] T. Cui and C. Tellambura, "An efficient generalized sphere decoder for rank-deficient MIMO systems," IEEE Communications Letters, vol. 9, no. 5, pp. 423–425, 2005. (Cited on pages 20 and 39.)

[16] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 1639–1642, July 1999. (Cited on page 22.)

[17] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," IEEE Transactions on Information Theory, vol. 49, no. 10, pp. 2389–2402, Oct 2003. (Cited on page 22.)

[18] M. Ju, J. Qian, Y. Li, G. Tan, and X. Li, "Comparison of multiuser mimo systems with mf, zf and mmse receivers," in 2013 IEEE Third International Conference on Information Science and Technology (ICIST), March 2013, pp. 1260–1263. (Cited on page 22.)

[19] C. P. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," Math. Program., vol. 66, no. 2, pp. 181–199, Sep. 1994. [Online]. Available: http://dx.doi.org/10.1007/BF01581144 (Cited on page 25.)

[20] Y. Sun, "A family of likelihood ascent search multiuser detectors: an upper bound of bit error rate and a lower bound of asymptotic multiuser efficiency," IEEE Transactions on Communications, vol. 57, no. 6, pp. 1743–1752, June 2009. (Cited on page 26.)

[21] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity near-ML performance achieving algorithm for large mimo detection," in 2008 IEEE

International Symposium on Information Theory, July 2008, pp. 2012–2016. (Cited on page 26.)

[22] N. Srinidhi, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-ML decoding of large non-orthogonal stbcs using reactive tabu search," in 2009 IEEE International Symposium on Information Theory, June 2009, pp. 1993–1997. (Cited on page 29.)

[23] H. Zhao, H. Long, and W. Wang, "Tabu search detection for mimo systems," in 2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Sept 2007, pp. 1–5. (Cited on page 29.)

[24] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," Journal of the American Mathematical Society, vol. 22, no. 1, pp. 1–53, January 2009. (Cited on pages 35 and 40.)

[25] O. L. Mangasarian and B. Recht, "Probability of unique integer solution to a system of linear equations," European Journal of Operational Research, vol. 214, no. 1, pp. 27–30, 2011. (Cited on page 36.)

[26] J. G. Wendel, "A problem in geometric probability," Mathematica Scandinavica, vol. 11, pp. 109–112, 1962. (Cited on page 36.)

[27] A. Aïssa-El-Bey, D. Pastor, S. Aziz Sbaï, and Y. Fadlallah, "Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems," IEEE Transactions on Information Theory, vol. 61, no. 4, pp. 2008–2018, April 2015. (Cited on pages 37 and 41.)

[28] Y. Fadlallah, A. Aïssa-El-Bey, K. Amis, D. Pastor, and R. Pyndiah, "New iterative detector of MIMO transmission using sparse decomposition," IEEE Transactions on Vehicular Technology, vol. 64, no. 8, pp. 3458–3464, August 2015. (Cited on page 37.)

[29] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 33–61, 1998. (Cited on page 37.)

[30] Z. Hajji, K. Amis Cavalec, A. Aïssa-El-Bey, and F. Abdelkefi, "Low-complexity half-sparse decomposition-based detection for massive MIMO transmission," in 5th International Conference on Communications and Networking (ComNet), November 2015, pp. 1–6. (Cited on pages 37 and 38.)

[31] X. Fan, J. Song, D. P. Palomar, and O. C. Au, "Universal binary semidefinite relaxation for ml signal detection," IEEE Transactions on Communications, vol. 61, no. 11, pp. 4565–4576, November 2013. (Cited on page 37.)

[32] Y. Fadlallah, A. Aïssa-El-Bey, K. Amis, D. Pastor, and R. Pyndiah, "New decoding strategy for underdetermined MIMO transmission using sparse decomposition," in Proceedings of the 21st European Signal Processing Conference (EUSIPCO), September 2013, pp. 1–5. (Cited on page 38.)

[33] H. Karloff, The Simplex Algorithm. Boston, MA: Birkhäuser Boston, 1991, ch. 2, pp. 23–47. (Cited on page 40.)

[34] Y. Nesterov and A. Nemirovskii, Interior-Point Polynomial Algorithms in Convex Programming, Path-Following Interior-Point Methods. Society for Industrial and Applied Mathematics, 1994, ch. 3, pp. 57–99. (Cited on page 40.)

[35] D. L. Donoho and J. Tanner, "Counting the faces of randomly-projected hypercubes and orthants, with applications," Discrete & Computational Geometry, vol. 43, no. 3, pp. 522–541, 2010. (Cited on page 40.)

[36] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, March 2014. (Cited on page 41.)

[37] ——, "Graph implementations for nonsmooth convex programs," in Recent Advances in Learning and Control, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html. (Cited on page 41.)

[38] E. D. Andersen, C. Roos, and T. Terlaky, "On implementing a primal-dual interior-point method for conic quadratic optimization," Mathematical Programming, vol. 95, no. 2, pp. 249–277, Feb 2003. (Cited on page 41.)

[39] P. B. Stark and R. L. Parker, "Bounded-variable least-squares: an algorithm and applications," Computational Statistics, vol. 10, no. 2, pp. 129–141, 1995. (Cited on page 44.)

[40] K. V. Mardia, J. T. Kent, and J. M. Bibby, Multivariate analysis. Academic press, 1980. (Cited on pages 45 and 58.)

[41] K. Cho and D. Yoon, "On the general ber expression of one- and two-dimensional amplitude modulations," IEEE Transactions on Communications, vol. 50, no. 7, pp. 1074–1080, Jul 2002. (Cited on page 46.)

[42] I. Gurobi Optimization, "Gurobi optimizer reference manual," 2015. [Online]. Available: http://www.gurobi.com (Cited on page 47.)

[43] P. Li, R. C. de Lamare, and R. Fa, "Multiple feedback successive interference cancellation with shadow area constraints for MIMO systems," in Wireless Communication Systems (ISWCS), 2010 7th International Symposium on, Sept 2010, pp. 96–101. (Cited on page 56.)

[44] M. Chiani, "Introducing erasures in decision-feedback equalization to reduce error propagation," IEEE Transactions on Communications, vol. 45, no. 7, pp. 757–760, Jul 1997. (Cited on page 56.)

[45] M. Reuter, J. C. Allen, J. R. Zeidler, and R. C. North, "Mitigating error propagation effects in a decision feedback equalizer," IEEE Transactions on Communications, vol. 49, no. 11, pp. 2028–2041, Nov 2001. (Cited on page 56.)

[46] P. Li, R. C. de Lamare, and R. Fa, "Multiple feedback successive interference cancellation detection for multiuser MIMO systems," IEEE Transactions on Wireless Communications, vol. 10, no. 8, pp. 2434–2439, August 2011. (Cited on page 56.)

[47] P. Li and R. C. D. Lamare, "Adaptive decision-feedback detection with constellation constraints for MIMO systems," IEEE Transactions on Vehicular Technology, vol. 61, no. 2, pp. 853–859, Feb 2012. (Cited on page 56.)

[48] A. Elghariani and M. Zoltowski, "Low complexity detection algorithms in large-scale MIMO systems," IEEE Transactions on Wireless Communications, vol. 15, no. 3, pp. 1689–1702, March 2016. (Cited on page 56.)

[49] 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. (Cited on page 64.)

[50] Z. Hajji, K. Amis, and A. Aïssa-El-Bey, "Turbo detection based on signal simplicity and compressed sensing for massive MIMO transmission," in IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, April 2018. (Cited on pages 65, 67 and 68.)

[51] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," The Annals of Statistics, vol. 37, no. 4, pp. 1705–1732, August 2009. (Cited on page 67.)

[52] Z. Hajji, A. Aïssa-El-Bey, and K. Amis, "Simplicity-based recovery of finite-alphabet signals for large-scale MIMO systems," Digital Signal Processing, vol. 80, pp. 70–82, September 2018. (Cited on page 67.)

[53] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," IEEE Transactions on Communications, vol. 51, no. 3, pp. 389–399, March 2003. (Cited on page 67.)

[54] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," IEEE Transactions on Information Theory, vol. 20, no. 2, pp. 284–287, March 1974. (Cited on page 68.)

[55] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," IEEE Transactions on Information Theory, vol. 46, no. 3, pp. 933–946, May 2000. (Cited on page 76.)

[56] A. Lapidoth and S. Shamai, "Fading channels: how perfect need "perfect side information" be?" IEEE Transactions on Information Theory, vol. 48, no. 5, pp. 1118–1134, May 2002. (Cited on page 76.)

[57] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error," IEEE Transactions on Information Theory, vol. 52, no. 5, pp. 2203–2214, May 2006. (Cited on page 76.)

[58] L. Berriche, K. Abed-Meraim, and J. Belfiore, "Investigation of the channel estimation error on mimo system performance," in 2005 13th European Signal Processing Conference, Sept 2005, pp. 1–4. (Cited on page 76.)

[59] M. Dong and L. Tong, "Optimal design and placement of pilot symbols for channel estimation," IEEE Transactions on Signal Processing, vol. 50, no. 12, pp. 3055–3069, Dec 2002. (Cited on page 76.)

[60] L. Berriche, K. Abed-Meraim, and J. C. Belfiore, "Cramer-Rao bounds for MIMO channel estimation," in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, May 2004, pp. iv–iv. (Cited on page 76.)

[61] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" IEEE Transactions on Information Theory, vol. 49, no. 4, pp. 951–963, April 2003. (Cited on pages 76 and 77.)

[62] X. Ma, L. Yang, and G. B. Giannakis, "Optimal training for MIMO frequency-selective fading channels," IEEE Transactions on Wireless Communications, vol. 4, no. 2, pp. 453–466, March 2005. (Cited on page 76.)

[63] T. L. Marzetta, "How much training is required for multiuser MIMO?" in 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, Oct 2006, pp. 359–363. (Cited on pages 76 and 77.)

[64] E. Nayebi and B. D. Rao, "Semi-blind channel estimation for multiuser massive MIMO systems," IEEE Transactions on Signal Processing, vol. 66, no. 2, pp. 540–553, Jan 2018. (Cited on pages 77 and 79.)

[65] P. Stoica and A. Nehorai, "Music, maximum likelihood, and cramer-rao bound," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 5, pp. 720–741, May 1989. (Cited on page 80.)

[66] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell tdd systems," IEEE Transactions on Wireless Communications, vol. 10, no. 8, pp. 2640–2651, August 2011. (Cited on page 103.)

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Gestion des interférences dans les systèmes large-scale MIMO pour la 5G

Mots clés : Large-scale MIMO, Compressive sensing, Alphabet fini

Résumé : La thèse s'inscrit dans la perspective de l'explosion du trafic de données générée par l'augmentation du nombre d'utilisateurs ainsi que la croissance du débit qui doivent être prises en compte dans la définition des futures générations de communications radio-cellulaires. Une solution est la technologie « large-scale MIMO » (systèmes MIMO de grande dimension) qui pose plusieurs défis. La conception des nouveaux algorithmes de détection de faible complexité est indispensable vu que les algorithmes classiques ne sont plus adaptés à cette configuration à cause de leurs mauvaises performances de détection ou de leur complexité trop élevée fonction du nombre d'antennes. Une première contribution de la thèse est un algorithme basé sur la technique de l'acquisition comprimée en exploitant les propriétés des signaux à alphabet fini. Appliqué à des systèmes MIMO de grande dimension, déterminés et sous-déterminés, cet algorithme réalise des performances (qualité de détection, complexité) prometteuses et supérieures comparé aux algorithmes de l'état de l'art. Une étude théorique approfondie a été menée pour déterminer les conditions optimales de fonctionnement et la distribution statistique des sorties. Une seconde contribution est l'intégration de l'algorithme original dans un récepteur itératif en différenciant les cas codé (code correcteur d'erreurs présent) et non codé. Un autre défi pour tenir les promesses des systèmes large-scale MIMO (efficacité spectrale élevée) est l'estimation de canal. Une troisième contribution de la thèse est la proposition d'algorithmes d'estimation semi-aveugles qui fonctionnent avec une taille minimale des séquences d'apprentissage (égale au nombre d'utilisateurs) et atteignent des performances très proches de la borne théorique.

Title : Interference management in large-scale MIMO systems for 5G

Keywords : Large-scale MIMO, Compressive sensing, Finit-alphabet

Abstract : The thesis is part of the prospect of the explosion of data traffic generated by the increase of the number of users as well as the growth of the bit rate which must be taken into account in the definition of future generations of radio-cellular communications. A solution is the large-scale MIMO technology (MIMO systems of large size) which poses several challenges. The design of the new low complexity detection algorithms is indispensable since the conventional algorithms are no longer adapted to this configuration because of their poor detection performance or their too high complexity depending on the number of antennas. A first contribution of the thesis is an algorithm based on the technique of compressed sensing by exploiting the properties of the signals with finite alphabet. Applied to large-scale, determined and under-determined MIMO systems, this algorithm achieves promising and superior performance (quality of detection, complexity) compared to state-of-the-art algorithms. A thorough theoretical study was conducted to determine the optimal operating conditions and the statistical distribution of outputs. A second contribution is the integration of the original algorithm into an iterative receiver by differentiating the coded and uncoded cases. Another challenge to keeping the promise of large- scale MIMO systems (high spectral efficiency) is channel estimation. A third contribution of the thesis is the proposal of semi-blind channel estimation algorithms that work with a minimum size of pilot sequences (equal to the number of users) and reach performances very close to the theoretical bound.