Activity models and Bayesian estimation algorithms for wireless grant-free random access

Lélio Chetot

N° d'ordre NNT : 2022LYSEI062

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

Institut National des Sciences Appliquées de Lyon

École Doctorale 160
Électronique, Électrotechnique et Automatique

Spécialité / Discipline de doctorat :
Traitement du signal et de l'image

Soutenue publiquement le 07/07/2022, par :

Lélio CHETOT

Activity Models and Bayesian Estimation Algorithms for Wireless Grant-Free Random Access

---

Devant le jury composé de :

CIBLAT Philippe	Professeur des Universités	Télécom Paris	Président
DOUILLARD Catherine	Professeure des Universités	IMT Atlantique	Rapporteuse
VUKOBRATOVIĆ Dejan	Professeur des Universités	Université de Novi Sad	Rapporteur
ROUMY Aline	Directrice de Recherche	INRIA Rennes	Examinatrice
STEFANOVIĆ Čedomir	Professeur des Universités	Université d'Aalborg	Examinateur
GORCE Jean-Marie	Professeur des Universités	INSA Lyon	Directeur
EGAN Malcolm	Chargé de Recherche	INRIA Lyon	Co-directeur

Sigle	École doctorale	Nom et coordonnées du responsable
CHIMIE	Chimie de Lyon *Web:* https://www.edchimie-lyon.fr *Sec:* Renée El Melhem *Bât:* Blaise Pascal, 3ème étage *Tél:* 04 72 43 80 46 *eMail:* secretariat@edchimie-lyon.fr	M. Stéphane DANIELE C2P2-CPE LYON-UMR 5265 Bâtiment F308 bP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne *Tél:* 04 72 44 53 60 *eMail:* directeur@edchimie-lyon.fr
E.E.A.	Électronique, Électrotechnique, Automatique *Web:* https://edeea.universite-lyon.fr *Sec:* Stéphanie Cauvin *Bât:* Direction INSA Lyon *Tél:* 04 72 43 71 70 *eMail:* secretariat.edeea@insa-lyon.fr	M. Philippe Delachartre Laboratoire Creatis Bâtiment Blaise Pascal 7 avenue Jean Capelle 69621 Villeurbanne CEDEX *Tél:* 04 72 43 88 63 *eMail:* philippe.delachartre@insa-lyon.fr
E2M2	Évolution, Écosystème, Microbiologie, Modélisation *Web:* http://e2m2.universite-lyon.fr *Sec:* Bénédicte Lanza *Bât:* Atrium, UCB Lyon 1 *Tél:* 04 72 44 83 62 *eMail:* secretariat.e2m2@univ-lyon1.fr	Mme Sandrine Charles Université Claude Bernard Lyon 1 UFR Biosciences 43 boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX *Tél:* / *eMail:* sandrine.charles@univ-lyon1.fr
EDISS	Interdisciplinaire Sciences-Santé *Web:* http://ediss.universite-lyon.fr *Sec:* Bénédicte Lanza *Bât:* Atrium, UCB Lyon 1 *Tél:* 04 72 44 83 62 *eMail:* secretariat.ediss@univ-lyon1.fr	Mme Sylvie Ricard-Blum ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX *Tél:* 04 72 44 82 32 *eMail:* sylvie.ricard-blum@univ-lyon1.fr
INFOMATHS	Informatique et Mathématiques *Web:* http://edinfomaths.universite-lyon.fr *Sec:* Renée El Melhem *Bât:* Bâtiment Blaise Pascal, 3e étage *Tél:* 04 72 43 80 46 *eMail:* infomaths@univ-lyon1.fr	Mme Sylvie Ricard-Blum Université Claude Bernard Lyon 1 Bâtiment Nautibus 43 boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX *Tél:* 04 72 44 83 69 *eMail:* hamamache.kheddouci@univ-lyon1.fr
Matériaux	Matériaux de Lyon *Web:* http://edinfomaths.universite-lyon.fr *Sec:* Yann De Ordenana *Bât:* / *Tél:* 04 72 18 62 44 *eMail:* yann.de-ordenana@ec-lyon.fr	M. Stéphane Benayoun École Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX *Tél:* 04 72 18 64 37 *eMail:* stephane.benayoun@ec-lyon.fr
MEGA	Mécanique, Énergétique, Génie Civil, Acoustique *Web:* http://edmega.universite-lyon.fr *Sec:* Stéphanie Cauvin *Bât:* Direction INSA Lyon *Tél:* 04 72 43 71 70 *eMail:* mega@insa-lyon.fr	M. Jocelyn Bonjour INSA Lyon Laboratoire CETHIL – Bâtiment Sadi-Carnot 9 rue de la Physique 69621 Villeurbanne CEDEX *Tél:* / *eMail:* jocelyn.bonjour@insa-lyon.fr
ScSo	Sciences Sociales: Histoire, Géo., Aménagement, Urbanisme, Archéo., Science po., Socio., Anthropo. *Web:* https://edsciencessociales.universite-lyon.fr *Sec:* Mélina Faveton & J.Y. Toussaint (INSA) *Bât:* / *Tél:* 04 78 69 72 79 *eMail:* melina.faveton@univ-lyon2.fr	M. Christian Montes Université Lumière Lyon 2 / 86 Rue Pasteur 69365 Lyon CEDEX 07 *Tél:* / *eMail:* christian.montes@univ-lyon2.fr

ABBREVIATIONS

ABBREVIATIONS

2G second generation 3, 4, 9, 153, 155
3G third generation 3, 4, 153
3GPP 3rd Generation Partnership Project 3, 4, 153
4G fourth generation xix, 1, 3, 4, 8, 9, 16, 153, 155
- **4G-LTE** 4G long-term evolution 3, 6, 8, 154

5G fifth generation xix, 1, 3–9, 16, 17, 19, 33, 153–156, 158
- **5G-NR** 5G new radio 1, 16, 17, 19, 123, 158, 175
- **B5G** beyond 5G 4, 5, 8, 154, 155
- **eMBB** enhanced mobile broadband 1, 5–8, 11, 154–156
- **mMTC** massive machine-type communication 1, 5–7, 11, 16, 17, 19, 23, 26, 30, 33, 123, 154–158, 162, 175
- **uRLLC** ultra reliable and low-latency communication 1, 5–7, 11, 16, 17, 19, 23, 26, 30, 33, 123, 154–158, 162, 175

6G sixth generation xix, 4, 5, 7, 8, 154, 155

AMP approximate message passing 1, 35, 49, 51, 52, 54, 56, 57, 167
- **BiGAMP** bilinear GAMP 128
- **BiGVAMP** bilinear GVAMP 128, 177
- **GAMP** generalized AMP 2, 52–57, 60–62, 72, 74, 75, 77, 78, 81–84, 90, 91, 105, 107–111, 114, 123, 124, 128, 132, 140, 141, 145, 146, 152, 167, 168, 171, 174–176
- **GS-HGAMP** group-sparse HGAMP 2, 61, 62, 75, 77, 78, 81–84, 90, 91, 110, 111, 113, 114
- **GVAMP** generalized VAMP 57
- **HGAMP** hybrid GAMP 2, 61, 62, 70, 73, 75, 78, 81–84, 90, 91, 103, 105–108, 110–114, 123, 124, 126, 128, 141, 146, 168, 170–177
- **VAMP** vector AMP 57, 128

AP access point 1, 9, 11, 15–19, 21, 27, 33, 62, 63, 79–82, 125, 155, 157–159, 161, 162, 168, 169, 171, 176

AR augmented reality 6, 7

AUDaCE active user detection and channel estimation 1, 2, 11, 12, 32, 33, 35, 39, 69, 70, 72, 78, 80, 84, 89–91, 102, 113, 114, 123–126, 161, 162, 168, 170–173, 175, 176

BP belief propagation 35, 39, 45, 46, 48, 49, 51, 56, 60, 74, 77, 105, 107, 108, 114, 129, 167, 168, 174
- **EP** expectation propagation 45, 49, 53, 57, 128, 132, 133
- **LBP** loopy BP 46, 49, 51, 61, 70–72, 82, 91, 103–105, 113, 123, 124, 128, 175, 176

BPDN basis pursuit denoising 37, 39, 51, 52, 163

cdf cumulative distribution function viii, 93–95, 173

CS compressed sensing 1, 2, 10–12, 19, 28, 32, 33, 35–39, 49, 51, 54, 55, 123, 156, 157, 162, 163, 167, 171

DL downlink 8, 11, 80

DrAUDaCE data recovery, active user detection and channel estimation 2, 126, 127, 176

EM expectation maximization 57

ETSI European Telecommunications Standards Institute 3, 153

FAR false alarm rate 30, 83, 110, 114

GF grant-free 2, 18, 126

GHetA group-heterogeneous activity 2, 12, 90, 91, 98, 99, 102, 105, 108, 110–114, 123, 124, 126, 128, 146, 172–177

GHomA group-homogeneous activity 2, 11, 61, 62, 64, 69, 70, 73, 75, 78–84, 89–92, 98, 101, 103, 105–107, 110, 111, 113, 114, 123, 124, 141, 170–176

i.i.d. identically and independently distributed 63, 64, 69, 125

IEEE Institute of Electrical and Electronics Engineers 3, 153

IHTA iterative hard thresholding algorithm 38, 163

IoE internet of everything 7

IoT internet of things 1, 6–10, 18, 21, 26, 27, 90, 154, 155, 158
- **IIoT** industrial IoT 6, 7, 79, 89, 154, 159

ISTA iterative soft thresholding algorithm 38, 163

ITU International Telecommunication Union 3, 153

KPI key performance indicator 4, 6–8, 11

LASSO least absolute shrinkage and selection operator 37, 39, 51, 52, 60, 163

LDS low density spreading 10

LPWAN low-power wide area network 6, 9

MA multiple access 9–11, 155, 156
- **CDMA** code division MA 9
- **FDMA** frequency division MA 9

NOMA non-orthogonal multiple access 1, 9, 10, 19, 33, 35, 60, 123, 156, 158
- **C-NOMA** code-domain NOMA 10
- **CS-NOMA** CS-based NOMA 10, 19, 27, 158, 160, 162, 175
- **P-NOMA** power-domain NOMA 10

OFDMA orthogonal frequency division MA 9
OMA orthogonal MA 9, 155, 156
TDMA time division MA 9

MAP maximum a posteriori 29, 32, 39, 42–44, 46–48, 51, 52, 54, 161, 165–167
MDR missed detection rate 30, 83, 110, 114
ML maximum likelihood 43
MMSE minimum mean squared error 30–32, 43, 44, 46–48, 52–54, 70, 103, 132, 133, 161, 165–167
MP matching pursuit 37, 163
- **CoSaMP** compressive sampling MP 38, 163
- **FBMP** fast bayesian MP 38, 163
- **MMP** multipath MP 38, 163
- **OMP** orthogonal MP 37, 60, 163
- **StOMP** stagewise OMP 38, 163

NMSE normalized mean squared error 31, 62, 82–84, 110, 112, 113

OFDM orthogonal frequency division multiplexing 2, 3, 19, 20, 114, 125–127, 153, 158, 159, 176

pdf probability density function viii, 51, 54–56, 64–66, 69, 93, 94, 97, 104, 108, 126, 128, 134, 135, 144, 149, 161, 167, 170, 173, 174
pmf probability mass function viii, 22, 23, 25, 26, 29, 33, 92, 159

RA random access 1, 10–12, 15–19, 22, 23, 27, 29, 33, 114, 156–159, 161, 168
- **CBRA** contention-based RA 17
- **CFRA** contention-free RA 17
- **GFRA** grant-free RA 1, 2, 15, 17–19, 22, 33, 35, 39, 62, 79, 84, 89–91, 113, 123, 126, 158–162, 165, 168, 171, 175, 176

RIC restricted isometry constant 28, 160
RIP restricted isometry property 28, 160, 161

SIC succesive interference cancellation 10

UAV unmaned aerial vehicle 6, 154
UE user equipment 1, 2, 6, 9–11, 15–23, 25–27, 29, 30, 33, 62–64, 66, 69, 74, 78, 79, 84, 89–91, 105–107, 110–113, 123, 125–127, 156–162, 171, 172, 175, 176
UER user error rate 29, 62, 82, 83, 110, 112–114
UL uplink 8, 15, 19, 80, 125

V2X vehicular-to-everything 6, 26, 27, 154, 159
VI variational inference 45, 46, 53, 166

VIEF VI with exponential family 45–47, 49, 53, 133, 166

VR virtual reality 6, 7

ZC Zadoff-Chu 18, 27

NOTATIONS

SETS

$[a,b],]a,b[, [a,b[,]a,b]$	real intervals (closed, open, right-open, left-open)
$[N]$	set of natural integers in [1, N]
$\mathbb{N}$	set of natural integers
$\mathbb{Z}$	set of relative integers
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers

VARIABLES

x or X	deterministic scalar
$\boldsymbol{x} = [x_i]_i^\intercal$	deterministic column vector
$\boldsymbol{X} = [x_{ij}]_{i,j}$	deterministic matrix
x or X	random scalar
$\mathbf{x} = [\mathsf{x}_i]_i^\intercal$	random column vector
$\mathbf{X} = [\mathsf{x}_{ij}]_{i,j}^\intercal$	random matrix

SPECIAL MATRICES

$\boldsymbol{I}_N$	identitiy matrix of dimension N
$\mathbf{0}_{N\times M}$	all-zero matrix of dimension $N \times M$
$\mathbf{1}_{N\times M}$	all-one matrix of dimension $N \times M$

DISTRIBUTIONS

$\mathsf{Bern}(p)$	Bernoulli distribution with success probability $p \in [0,1]$

$\mathsf{Beta}(\alpha, \beta)$	Beta distribution with concentrations α and β
$\mathsf{Unif}(\mathcal{S})$	uniform distribution on a set $\mathcal{S}$
$\mathsf{Dirac}(\boldsymbol{M})$	Dirac distribution at matrix point $\boldsymbol{M}$
$\mathsf{Laplace}(\lambda)$	Laplace distribution with scale parameter $\lambda > 0$
$\mathsf{Norm}(\boldsymbol{M}, \boldsymbol{C})$	Gaussian distribution with mean $\boldsymbol{M}$ and covariance $\boldsymbol{C}$
$\mathsf{CNorm}(\boldsymbol{M}, \boldsymbol{C})$	Complex gaussian distribution with mean $\boldsymbol{M}$ and covariance $\boldsymbol{C}$

STATISTICS & RELATED

$F_{\mathsf{X}}(\boldsymbol{X})$	cdf of X evaluated at $\boldsymbol{X}$
$f_{\mathsf{X}}(\boldsymbol{X})$	pdf of X evaluated at $\boldsymbol{X}$
$\mathbb{P}_{\mathsf{X}}(\boldsymbol{X})$	pmf of X evaluated at $\boldsymbol{X}$
$\mathbb{E}[\mathsf{X}] = [\mathbb{E}[\mathsf{x}_{ij}]]_{i,j}$	expectation of X
$\mathbb{V}[\mathsf{X}] = [\mathbb{V}[\mathsf{x}_{ij}]]_{i,j}$	variance of X
$\mathrm{Cov}[\mathsf{x}, \mathsf{y}] = \mathbb{E}[\mathsf{x}\mathsf{y}^*] - \mathbb{E}[\mathsf{x}]\,\mathbb{E}[\mathsf{y}]^*$	covariance operator between x and y
$\mathrm{Cor}[\mathsf{x}, \mathsf{y}] = \mathrm{Cov}[\mathsf{x}, \mathsf{y}] \,/\, \sqrt{\mathbb{V}[\mathsf{x}]\,\mathbb{V}[\mathsf{y}]}$	correlation operator between x and y
$\mathbb{H}_f[\mathsf{X}] = -\int f_{\mathsf{X}} \log f_{\mathsf{X}}$	entropy of X w.r.t. the pdf f
$\mathbb{X}_{f,g}[\mathsf{X}] = -\int f_{\mathsf{X}} \log g_{\mathsf{X}}$	cross-entropy of X between pdfs f and g
$\mathbb{KL}_{\mathsf{X}}[f \Vert g] = \int f_{\mathsf{X}} \log \frac{f_{\mathsf{X}}}{g_{\mathsf{X}}}$	Kullback-Leibler divergence between pdfs f and g on X
$\mathcal{N}(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{C})$	gaussian pdf
$\mathcal{CN}(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{C})$	complex gaussian pdf
$\mathcal{B}(x; \alpha, \beta)$	beta pdf

OPERATORS

$\boldsymbol{X}^{\mathsf{T}}$	transpose of $\boldsymbol{X}$
$\boldsymbol{X}^*$	complex conjugate of $\boldsymbol{X}$
$\boldsymbol{X}^{\mathsf{H}}$	conjugate transpose of $\boldsymbol{X}$
$\lvert\boldsymbol{X}\rvert$	component-wise modulus, i.e. $[\lvert x_{ij}\rvert]_{i,j}$
$\boldsymbol{X} \odot \boldsymbol{Y}$	Hadamard (or component-wise) product, i.e. $[x_{ij} y_{ij}]_{i,j}$
$\boldsymbol{X} \oslash \boldsymbol{Y}$	Hadamard (or component-wise) quotient, i.e. $[x_{ij} / y_{ij}]_{i,j}$
$\boldsymbol{X}^{\odot a}$	Hadamard (or component-wise) exponentiation, i.e. $[x_{ij}^a]_{i,j}$
$\langle \boldsymbol{X} \rangle$	average of $\boldsymbol{X} \in \mathbb{C}^{M \times N}$ i.e. $\sum_{(m,n)} x_{mn} / (MN)$
$\lVert \boldsymbol{X} \rVert_p$	ℓ_p-norm of $\boldsymbol{X} \in \mathbb{C}^{M \times N}$ i.e. $(\sum_{(m,n)} \lvert x_{mn}\rvert^p)^{1/p}$
$\mathbb{1}(\cdot)$	indicator function
$\mathrm{tr}(\cdot)$	trace
$\mathrm{diag}(\cdot)$	vector from square matrix's diagonal or diagonal matrix from vector
$\partial_{\boldsymbol{X}} f(\boldsymbol{X})$	jacobian of f w.r.t. $\boldsymbol{X}$

CONTENTS

List of Figures

List of Tables

List of Algorithms

REMERCIEMENTS

Cette thèse est l'accomplissement d'un travail de près de 4 ans dont la réussite dépasse le cadre de mon doctorat et prend ses racines bien au-delà de mon implication personnelle. Je souhaite donc remercier, par les quelques lignes qui suivront, l'ensemble des personnes qui y ont contribué, volontairement ou non.

J'exprime tout d'abord ma reconnaissance envers mes encadrants de thèse, Jean-Marie Gorce et Malcolm Egan. Grâce à Jean-Marie, j'ai découvert le monde de la recherche durant mes trois années passées en tant qu'étudiant au département Télécommunications de l'INSA Lyon avant de poursuivre en thèse sous sa direction et celle de Malcolm. Leur complémentarité dans leur encadrement et la confiance qu'ils ont su m'accorder durant ces années en me laissant le temps d'explorer, d'apprendre et de me tromper, sont certainement ce que j'ai le plus apprécié.

Je souhaite remercier Mme. Douillard et M. Vukobratović en qualité de rapporteurs de ma thèse pour leurs précieux commentaires et pour les échanges que nous avons eu lors de la session de questions. De même, les discussions avec Mme. Roumy, M. Stefanović et M. Ciblat ont été tout autant appréciées. Enfin, je remercie les membres de mon jury pour avoir été les premiers réels interlocuteurs de mon travail, suite à la période des deux années de confinement liée à la pandémie.

Mes remerciements s'adressent aussi à l'équipe INRIA MARACAS et au laboratoire CITI qui m'ont accueilli et permis d'y trouver facilement ma place par l'intermédiaire des nombreux échanges (in)formels que j'ai pu avoir avec ses membres.

Étant une personne valorisant particulièrement le savoir et sa transmission, ce fut un réel plaisir d'avoir pu enseigné au sein du département

Télécommunications, Services et Usages de l'INSA Lyon auprès de ses étudiantes et étudiants.

Plus personnellement, je tiens à remercier ma famille. Elle m'a supporté, dans tous les sens du terme, pour un travail qui pouvait lui sembler bien souvent opaque, ce qui requiert des efforts certains. Je suis donc très heureux d'avoir pu lui présenter ce pour quoi elle s'est autant investie.

Je me dois de remercier également tous mes amis, qui ont participé tout au long de ma thèse à me maintenir connecter avec cette autre réalité qu'est la vie sociale en dehors d'un laboratoire. Amis d'enfance, de promotion, d'association ou de passage, j'aime croire qu'ils ont tous contribué à créer l'état d'esprit dont j'avais besoin pour me lancer et terminer cette aventure qu'est le doctorat. Je souhaite saluer leur patience, leur bienveillance, leur curiosité et leur tolérance. Je me risque donc à remercier nominativement certains d'entre eux en espérant n'oublier ni ne blesser personne.

Merci à mes amis d'enfance Baptise G. et Fabien C. Merci à mes amis de CPGE, Émeline L., Florence P., Joris T. et Tanguy K. pour qui j'ai une grande admiration. Merci à mes amis de promotion Guillaume T., Pierre F., Maxime P., Vincent H., Yesmine B. R. et Yoan T. pour m'avoir toujours témoigné leur considération en tout instant. Merci à mes bizuths Nolwenn M. et Timothée C. pour avoir été, d'une certaine façon, mes premiers élèves. Merci à Louise C. pour avoir été la première personne avec qui j'ai échangé sérieusement sur la possibilité de me lancer dans une thèse. Merci à Cylia B. pour l'enthousiasme qu'elle m'a montré pour mes très humbles premiers pas dans la recherche. Merci à ma partenaire de badminton Elise B. avec qui j'ai pu échanger de très nombreuses fois sur notre travail de doctorant. Merci à Coralie R., Fiona M., Kainoa A., Laurine T., Mathilde P. C. et Pauline B. pour avoir été à l'origine des bouffées d'air hebdomadaires, et nécessaires, durant la fin de ma thèse. Merci à Aurélie F. et Emma P. pour l'écoute sincère dont elles savent faire preuve. Merci à Angélique D. et Marion D. pour leur bienveillance incarnée et à toute épreuve. Finalement, c'est un bien heureux hasard qu'«Ami» rime avec «Merci».

Enfin, merci à toutes celles et ceux que je n'ai pas mentionnés mais qui sauront s'inclure à ces remerciements !

ABSTRACT

The new 5G's wireless networks have recently started to be deployed all around the world. With them, a large spectrum of services are about to emerge, resulting in new stringent requirements so that 5G targets performance exceeding that of 4G by a factor of 10. The services are centered around the use cases of enhanced mobile broadband (eMBB), ultra reliable and low-latency communication (uRLLC) and massive machine-type communication (mMTC) where each of which has required the ongoing development of key new technologies. Many of these technologies will also play an important role in the emergence of 6G.

In this thesis, the focus is on grant-free RA (GFRA) as an enabler of uRLLC and mMTC. GFRA is a new protocol introduced in 5G new radio (5G-NR) for reducing the data overhead of the random access (RA) procedure. This results in a significant reduction in the latencies of the user equipments (UEs) access to a connected medium via an access point (AP).

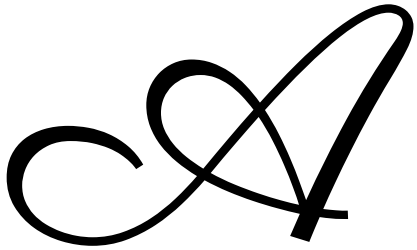
Achieving efficient GFRA is of key importance for many 5G applications, e.g. for large scale internet of things (IoT) wireless networks. The study of new non-orthogonal multiple access (NOMA) signal processing techniques is then considered. Using tools from the theory of compressed sensing (CS), and particularly from Bayesian CS, new algorithms within the family of approximate message passing (AMP) are developed to address the joint active user detection and channel estimation (AUDaCE) problem. The active user detection is crucial to properly identify transmitting UEs within the context of large-scale dense network; the channel estimation is equally important so that an AP can reliably transmit back data to the detected UEs.

In this thesis, in contrast to existing work on this topic, the AUDaCE is studied for wireless networks where the activity of the UEs is assumed to be correlated, as is typical for many large-scale dense networks. To this end, two new activity models are introduced. The first one assumes that the activity of the UEs in the network can be modeled via group-homogeneous activity (GHomA) where devices in the same group have common pairwise correlations and marginal activity probabilities. The second model accounts for more general dependence structure via group-heterogeneous activity (GHetA).

Novel approximate message passing algorithms within the hybrid GAMP (HGAMP) framework are developed for each of the models. With the aid of latent variables associated to each group for modeling the activity probabilities of the UEs, the GHomA-HGAMP algorithm can perform AUDaCE for GFRA leveraging such a group homogeneity. When the activity is heterogenous, i.e. each UE is associated with a latent variable modeling its activity probability correlated with the other variables, it is possible to develop GHetA-HGAMP using the copula theory.

Extensive numerical studies are performed, which highlight significant performance improvements of GHomA-HGAMP and GHetA-HGAMP over existing algorithms (modified generalized AMP (GAMP) and group-sparse HGAMP (GS-HGAMP)), which do not properly account for correlation in activity. In particular, the channel estimation and active user detection capability are enhanced in many scenarios with up to a 4dB improvement with twice less user errors.

As a whole, this thesis provides a systematic approach to AUDaCE for wireless networks with correlated activities using tools from Bayesian CS. We then conclude by showing how it could be used for multi-carrier orthogonal frequency division multiplexing (OFDM) scenarios with possible extensions for grant-free (GF) data transmission leveraging joint data recovery, active user detection and channel estimation (DrAUDaCE).



INTRODUCTION

OVERVIEW OF NEW GENERATION WIRELESS NETWORKS FROM 5G TO ENVISIONED 6G

Mobile wireless telecommunications has been standardized from the beginning of the years 1990 up to now. The second generation (2G) (1991), third generation (3G) (2001), 4G (2009) and now 5G (2018) form the five milestones in the story of the telecommunication standards, orchestrated by international actors such as the European Telecommunications Standards Institute (ETSI), 3rd Generation Partnership Project (3GPP), Institute of Electrical and Electronics Engineers (IEEE) and International Telecommunication Union (ITU). Each standard was developed to enable new telecommunication services brought by breakthrough technologies and changing requirements.

In particular, the former 4G long-term evolution (4G-LTE) popularized the celebrated OFDM waveform which is at the core of current wireless networks. The use of OFDM for modern communication systems has become so important and proved to be so efficient that 5G still relies on it. Its pre-standardization has started in [1, Release 13] and [2, Release 14] with identification of its future applications before being considered distinctly from 4G-LTE in [3, Release 15] and [4, Release 16]. The current

(a) Timeline for Release 17

(b) Timeline for Release 18

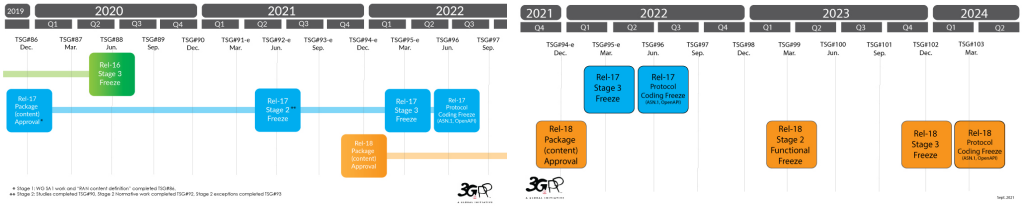


Figure A.1: 5G standardization process is still ongoing.

3GPP's standardization process for 5G is expected to release soon its 17th (see Fig. A.1a) report while Release 18 is still it its infancy (see Fig. A.1b).

5G is then the most recent standard for wireless networks. It has been conceived not only for the classical standard internet data transmission but also with motivations from the industry. Thereby, the user data-rate and the spectral efficiency are no longer the only prevalent key performance indicator (KPI) as they are accompanied by considerations on the network scalability and its connectivity density or the latency and the reliability of the communications.

As of today, the deployment of commercial 5G networks is expanding around the world, with a focus on mobile wireless networks on the side of existing 4G, 3G, 2G networks. The two latter are expected to disappear to give more frequency space to 5G[1], even with the introduction of the band of frequencies higher than 6GHz.

With the advent of 5G, the time has also come for the industry and research communities to actively start thinking the future communication needs, based on the current and envisioned technogical trends. Hence, beyond 5G (B5G) and even 6G wireless networks have become hot discussions and research topics.

The remainder of this chapter is outlined as follows. In Sec. A.II, we review the use cases of 5G, B5G and 6G. The focus is then given to what the multiple access is within this context in Sec. A.III before mentioning some key challenges in Sec. A.IV that will be studied in this thesis. A summary of the contributions is provided in Sec. A.V.

[1]horizon of 2028 in France, see https://reseaux.orange.fr/actualites/arret-2g-3g-en-france

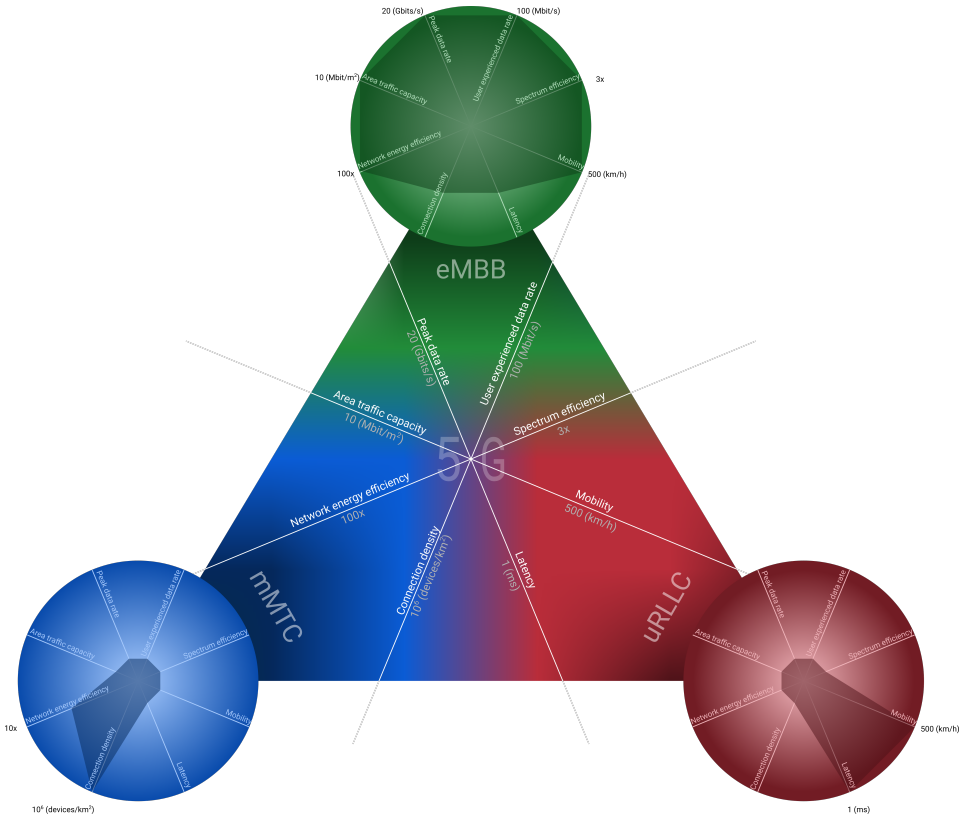


Figure A.2: 5G and its three use cases: eMBB, uRLLC and mMTC. Figure has been inspired by the ones in [5].

II Review of next generation use cases

In order to introduce the reader to the technological context of this thesis, we overview in this section the main use cases of 5G, but also those of B5G and 6G since they are strongly related. The term *use case* will refer to a family of technological applications.

II.1 5G and B5G use cases

In the context of 5G and B5G, the three typical use cases are eMBB, uRLLC and mMTC.

eMBB The use case of eMBB designates a pool of new high-rate driven services including the following:

- high mobility internet connection to guarantee the connection of users with high-speed mobility (highways, trains, planes) [6];
- mobile video streaming services, which is one of the most popular and

key applications of eMBB with the explosion of the consumption of streaming, video and livestream services [6];

- augmented reality (AR) and virtual reality (VR) (aimed to be natively supported) with many applications in the domains of industry with AR-assisted maintenance and assembly [7], education and training [8], entertainment, advertising and gaming [9], virtual environment design [10].

As a whole, eMBB is 5G's use case with faster user-centric applications improving on 4G-LTE mobile communications.

uRLLC It encompasses all the services with stringent requirements, both in terms of latency and reliability [11]. Only these two KPI are considered for services usually classified [12] into

- services where wireless replaces wired systems;
- native uRLLC services.

URLLC applications are numerous and all involve critical systems for which a failure of any kind has serious consequences. Among others, one can cite

- industrial IoT (IIoT), also named "Industry 4.0", where industrial plants and their production units are reorganized and connected together via wireless links to provide a flexible and fast adaptive industrial environment [13];
- autonomous vehicles with vehicular-to-everything (V2X) terrestrial communications [14] and unmaned aerial vehicle (UAV) communications as an enabler of uRLLC [15].
- health where the typical scenario is telesurgery [16] where a physician remotely operates a patient using medical equipments and gets reliable, seamless, haptic and video feedbacks.

mMTC MMTC is the connectivity use case involving dense and massive cellular networks, dominated by machine-to-machine communications, namely IoT. Unlike eMBB, mMTC traffic is sporadic which is a consequence of UEs only transmitting when they need it in order to save their energy. Such low-power wide area network (LPWAN) networks represent the core of mMTC including

- smart cities where sensors and devices are spread over large urban areas to enable their self sustainable management [17] by monitoring physical constants, road traffic, urban infrastructure, power grids and many other relevant data;
- health which leverages the increase in portable healthcare and body sensors;
- IIoT with large-scale industrial plants equipped with wireless sensors either for their monitoring (safety and security considerations) or their control (anomaly detection, automatic decision taking, etc).

II.2 6G use cases

For 6G, the use cases are not yet definitive. However, it seems that it is commonly agreed that they will fill 5G's failed promises or lie in the intersections of the 5G's use cases presented in Sec. A.II.1. Citing [18], "the drivers of 6G will be a confluence of past trends (e.g., densification, higher rates, and massive antennas) and of emerging trends that include new services and the recent revolution in wireless devices (e.g., smart wearables, implants, XR devices, and so on), artificial intelligence (AI), computing, and sensing". In other words, some of the 6G's use cases can be readily obtained by considering jointly 5G's use cases.

Massive uRLLC This is one the two intersections of the 5G's use cases which is considered for 6G: uRLLC and mMTC [18]. Such an intersection is natural, especially when considering IoT for reliable delay-sensitive networks, but also new services enabling the so-called internet of everything (IoE). Typical applications are the same as uRLLC but where the considered networks scale is that of mMTC.

eMBB with uRLLC The second interesection lies at the border of eMBB and, once again, uRLLC [18]. This use case then aims at bringing reliability and latency considerations to eMBB applications. In particular, AR, VR and their combination will need to leverage these KPIs in order to produce a real immersive user experience.

6G native use cases These use cases cannot be put under the umbrellas of massive uRLLC or eMBB/uRLLC because of different technological paradigms. Some of the envisioned applications are integration of satellites as core components of communication systems [19], visible light communications [20] and quantum-powered communications [21].

Metrics	Value (w.r.t. current networks)		
	5G	B5G	6G
Connection density	1e6 devices/km^2 (10×)		
Latency	5 ms (0.5×)	1 ms (0.1×)	$<$ 1 ms ($<$ 0.1×)
Mobility	500 km/h (+33%)		
Peak data rate	10 Gbps (10×)	100 Gbps (100×)	1 Tbps (1000×)
Reliability	99.999%	99.9999%	99.99999%
Spectrum & energy efficiency	10× bps/Hz/m^2/J (area)	100× bps/Hz/m^2/J (area)	1000× bps/Hz/m^3/J (volumetric)
User experienced data rate	100 Mbps (10×)	1 Gbps (100×)	10 Gbps (1000×)

Table A.1: KPIs from 5G to 6G w.r.t. current generation networks up to 4G [5], [18].

II.3 Key performance indicators

From 5G to 6G, the use cases requirements are mainly articulated around KPIs. As rules of thumb,

- 5G is 10 times 4G;
- B5G should be 100 times 4G;
- 6G should be 1000 times 4G.

The detail of some major KPIs is summarized in Tab. A.1.

Rate All the applications under the umbrella of eMBB aim at enhancing the data rates, improving the spectrum efficiency and the network energy efficiency, extending the connectivity coverage, supporting a higher area traffic and allowing a faster connected mobility. Compared to 4G-LTE, all these characteristics are expected to be improved by a large factor. For example, the experienced user rate is expected to reach up to 100Mb/s (resp. 50Mb/s) in the downlink (DL) (resp. uplink (UL)) compared to the 10Mb/s for 4G-LTE's throughput, the area traffic capacity is expected to handle rates per square meters a hundred times higher and the network energy efficiency will be improved by a factor of 100. These enhancements are supposed to answer the future needs of the aforementioned applications.

Reliability Different types of latency and reliability are considered. In [22], a distinction is made between end-to-end, user-plane and control plane latencies while the reliability can be considered at the layer 2, the node, the metadata or the availability levels. Typical values of reliability and latency constraints are $1 - 10^{-x}$ ($x \in \{3, 4, 5, 6, 8, 9\}$) and $c10^{-e}$ms (where e may be as low as 0 and c is a constant).

Scalability With the fast development of IoT, the number of devices per area unit is expected to reach values as high as 1M devices/km^2.

III Multiple access for next generation wireless networks

Now that the new generation use cases and their applications have been presented, an important aspect is the access of devices to an AP so that they can transmit their data.

III.1 Orthogonal multiple access

From 2G to 4G, the following different multiple access (MA) schemes were considered.

- Time division MA (TDMA) where the time is split into non overlapping slots, each of them being exclusively allocated to a UE.
- Frequency division MA (FDMA) where the frequency spectrum is cut into non-overlapping subbands onto which the UEs transmit without interfering with each other.
- Code division MA (CDMA) where each UE is given a binary spreading sequence which is pseudo-orthogonal with others UEs spreading sequences.
- Orthogonal frequency division MA (OFDMA) which is a combination of TDMA and FDMA where a time-frequency grid is formed of time-frequency resource elements, separated in time and orthogonal between frequency carriers.

A key design structure for each of this multiple access schemes is the separation of the UEs over one or several elementary resources (time, frequency, code). Such schemes are grouped under the category of orthogonal MA (OMA) schemes, meaning that the number of dedicated resources must scale linearly with the number of UEs to guarantee collision-free access.

The sustainability of OMA is then compromised for long-term 5G networks as it was mentioned that an increase and densification of mobile networks is expected, pushed by the developement of IoT with LPWAN. The development of NOMA schemes has then been considered to overcome the resource limitations.

III.2 Non-orthogonal multiple access

NOMA is then an important shift in the paradigm of MA. It encompasses any MA techniques that share common resources between the UE. By resources, we mean subcarriers, slots, antennas, power, codes or any other relevant physical resources. Most of these techniques have been extensively surveyed in [23]–[26] and we briefly review some of them hereafter.

Power-domain NOMA Power-domain NOMA (P-NOMA) consists in multiplexing signals in the same resource block using different power levels. At the receiver, the superimposed signals are separated by mitigating successively the interfering signals from the highest power level to the lowest one. This decoding process is known as succesive interference cancellation (SIC). A variant of P-NOMA is also known as rate-splitting NOMA where each signal source (i.e. UE) is split into multiple virtual sources so that the messages encoding costs are reduced [27].

Code-domain NOMA Code-domain NOMA (C-NOMA) consists in multiplexing signals in the same resource block using unique spreading sequences (i.e. pseudo random binary or complex sequences) for each UE. The spreading sequences are designed to be of low density, i.e. the number of non-zero elements is small w.r.t. the number of zeros. Such a technique which consists in incorporating sparse code into the transmitted signals to reduce the possible amount of interference is called low density spreading (LDS).

Compressed sensing-based NOMA CS-based NOMA (CS-NOMA) consists in leveraging the CS theory and the inherent sparsity in the activity pattern of the UEs in mobile networks to enable an efficient MA. Particularly, this NOMA scheme has recently received a lot of interest in the context of RA for IoT, where the low-power characteristic contributes to the sparse activity of the devices. Typical tasks such as active user detection, channel estimation and data recovery are the targeted applications of CS-NOMA, which can be addressed with CS techniques including optimization-based, greedy or bayesian algorithms that will be reviewed in C.I. CS-NOMA will be the chosen NOMA scheme of this thesis.

IV Challenges

It is now the time to open the discussion for the challenges that are considered in this manuscript. From Sec. A.II, uRLLC and mMTC are two use cases which significantly differ from eMBB in that their applications rely on KPIs which are not rate-centric[2]. For these use cases, there is an important problem issued by MA previously introduced in Sec. A.III, which consists in the RA procedure, that will be reviewed more extensively in Sec. B.I.

RA is an important step in every wireless networks that allows UEs to gain access to an AP for future data transmissions. It is therefore very challenging to design it so that the KPIs (scalability, latency, reliability) of uRLLC and mMTC can be met.

In particular, it will be justified in Chap. B that two important signal processing tasks that can enable RA which can comply with uRLLC and mMTC are

1. active user detection which allows a precise identification of UEs by an AP;
2. channel estimation which allows an AP to acquire large-scale information on the environment for improving the reliability of future DL transmissions

These two problems will be studied simultaneously as special cases of AUDaCE in Chap. D and Chap. E

V Contributions

A summary of the contributions of this thesis is given hereafter

1. A comprehensive review of bayesian CS techniques that are useful to AUDaCE is given in Chap. C.
2. Two new models of activity pattern for RA are introduced.
   (a) Group-homogeneous activity in Sec. D.II which relies on latent variables and assumed group structure;

[2]The rate can still be relevant though, but not with the same importance as for eMBB.

(b) Group-heterogeneous activity in Sec. E.II which relies on a simple application of the copula theory for flexible statistical dependence structures in the activity pattern;

3. A systematic approach to AUDaCE is leveraged for efficient grant-free RA within the framework of bayesian CS with the development of two algorithms in the family of hybrid generalized approximate message passing in Sec. D.III and Sec. E.III.

VI Publications & related

VI.1 Conference papers

- L. Chetot, J.-M. Gorce, and J.-M. Kelif, "Fundamental Limits in Cellular Networks with Point Process Partial Area Statistics," in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, Avignon, France: IEEE, Jun. 2019, pp. 1–8, ISBN: 978-3-903176-20-1, URL: https://doi.org/10.23919/WiOPT47501.2019.9144101
- D. Duchemin, L. Chetot, J.-M. Gorce, *et al.*, "Détecteur pour l'accès aléatoire massif entre machines avec connaissance statistique du canal en lien ascendant," p. 5, 2019
- D. Duchemin, L. Chetot, J.-M. Gorce, *et al.*, "Coded random access for massive MTC under statistical channel knowledge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, France: IEEE, Jul. 2019, pp. 1–5, ISBN: 978-1-5386-6528-2, URL: https://doi.org/10.1109/SPAWC.2019.8815491

VI.2 Journal papers

- L. Chetot, M. Egan, and J.-M. Gorce, "Joint Identification and Channel Estimation for Fault Detection in Industrial IoT with Correlated Sensors," *IEEE Access*, 2021, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2021.3106736 (corresponds to Chap. D)
- L. Chetot, M. Egan, and J.-M. Corce, "Active User Detection and Channel Estimation for Grant-Free Random Access with Group-Heterogeneous Activity", *In preparation*, (corresponds to Chap. E)

VI.3 Others

- Poster session, European School of Information Theory (ESIT), 2019
- Poster presentation (online), Communication theory workshop (CTW), 2021
- Presentation (online), GdR ISIS "Statistical learning with missing data", 2021

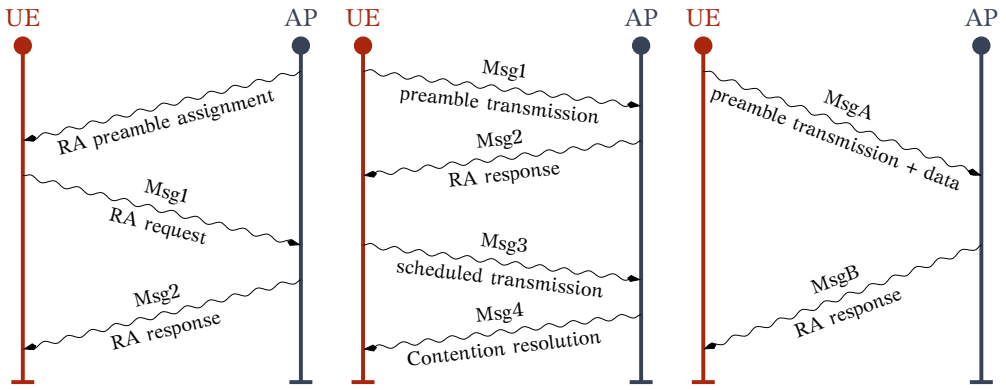
B

State-of-the-Art

In modern telecommunications, users do not typically transmit data continuously. As such, a protocol is required to develop an initial connection between each UE and AP. Since the time that this connection is required is not known a priori, the time of connection is random and the protocol is known as RA.

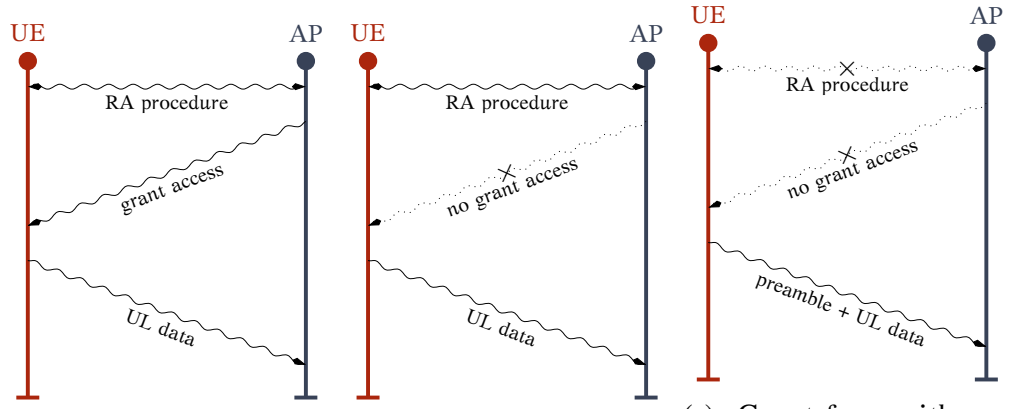
This chapter provides an overview of the state-of-the-art of RA. In Sec. B.I, the RA protocols introduced in the 3GPP standard [3], [4] are overviewed. In particular, GFRA is discussed, which will be the focus in the remainder of the thesis. In Sec. B.II, a baseband model for GFRA UL transmissions is then developed, including the preamble, channel attenuation and noise, as well as activity of the UEs. This model will be exploited for the main technical contributions in Chap. D and Chap. E.

The remainder of the chapter overviews the state-of-the-art for preamble design in Sec. B.III, user identification and channel estimation in Sec. B.IV, as well as further aspects such as resource allocation and payload data decoding in Sec. B.V. Finally, Sec. B.VI concludes, highlighting the key problems that will be addressed in Chap. D and Chap. E.



(a) Contention-free (b) Contention-based (c) Grant-free

Figure B.1: Random access procedures in 5G.



(a) Grant-based after RA (b) Grant-free after RA (c) Grant-free with no RA

Figure B.2: Uplink data transmission in 5G.

I Random access for mMTC and uRLLC

As discussed in Chap. A, an important aspect of 5G is RA of UEs to the network. Its optimization is even crucial for enabling mMTC and uRLLC where the requirements in terms of UE density support, latency and reliability constraints are strong, either individually or jointly. This section is dedicated to a general description of RA aspects in 5G-NR.

I.1 What is random access ?

In 4G and 5G communication systems, each time a UE seeks to transmit data to an AP (or access the network), it must agree with the AP that the data can be successfully sent by completing a RA procedure. The access

is said to be *random* as the time of the first transmission by the UE is unknown to the AP. In 5G-NR, there are the three RA procedures which are summarized in Fig. B.1 and are briefly described hereafter.

I.1.a Contention-based and contention-free random access

As of [3, Release 15], 5G-NR RA protocol is split into two different families: contention-based RA (CBRA) and contention-free RA (CFRA).

Contention-free random access As the names suggest, CFRA prevent contentions, i.e. possible simultaneous accesses of UEs that cannot be discriminated. This is made possible by assigning to each UE in each cell a unique preamble such that when a RA request is made, the AP cannot confuse the UE with any other UE. The number of unique preambles that an AP can allocate is limited and therefore, when no unique preamble can be allocated, the AP instructs the UE to use CBRA.

Contention-based random access When using CBRA, a UE tries to access the AP by first sending a (possibly non unique) preamble (Msg1) and waits for its RA response. In the case of a positive response (Msg2), the UE sends scheduled information (Msg3), which are not useful data and depends on its state. After some fixed duration, the UE receives a Msg4 from the AP which indicates whether a contention has been detected and if not, the UE has successfully completed the CBRA procedure.

I.1.b Grant-free random access

In [4, Release 16], a third RA protocol, known as GFRA, was introduced. As illustrated in Fig. B.1c, GFRA is a 2-step RA protocol which simplifies the CBRA by merging Msg1 with Msg3 to form MsgA. At the same time, MsgB is similar to Msg2 combined with Msg 4.

Such merge of messages is desirable for many 5G applications, in particular those of uRLLC and mMTC. With a 2-step approach, GFRA allows a reduction of the control overhead which is significant when the considered networks are densely and massively populated (mMTC) or when the access and future data transmissions must meet stringent latency requirements (uRLLC).

For these reasons, GFRA is expected to play a central role in 5G-NR and will constitute the main topic of this manuscript.

I.1.c Uplink data transmission

A UE may attempt to transmit its data in three different ways, depending on its needs. The first two methods, shown in Figs. B.2a and B.2b assumes that one of the RA procedures described in Sections B.I.1.a and B.I.1.b has been successfully completed. They differ by either waiting to receive a grant access message from the AP meaning that the UE is allowed to transmit its data or directly sending the data, with no granting from the AP.

The last method assumes that no RA procedure has been followed. The UE sends its data and its preamble without any grants (for access or data). This last scheme referred to as GF transmissions.

I.2 Non-orthogonal multiple access for grant-free random access

I.2.a Zadoff-Chu preambles

All the RA procedures described in Sec. B.I.1 rely on an identification step based on the transmitted preamble. These preambles are built from Zadoff-Chu (ZC) sequences [23] which are of unit euclidean norm and have low cross correlations, or almost-zero inner products. This last property is important since it makes ZC sequences quasi-orthogonal which is helpful when the AP must discriminate and identify transmitting UEs.

However, since the number of UEs is supposed to be extremely large, it is likely that the fixed number of ZC sequences will not be enough to guarantee pseudo-orthogonal RA. Indeed, recall that if a collection of N_{seq} orthogonal sequences, or vectors, is chosen to be the set of possible identification sequences, then the AP can only discriminate up to N_{seq} UE. In [3, Release 15], given the frequency range FR1 (resp. FR2), the RA sequences are vectors of length 839 (resp. 139).

I.2.b Non-orthogonal multiple access

In GFRA, a key problem is to ensure that the AP can identify which users have communicated their preamble. As the number of UE precludes the use of orthogonal preambles, it is desirable to utilize sophisticated signal processing techniques in order to reliably identify the UEs.

Since dense and massive IoT networks are expected to experience very sporadic traffic, due to power-saving system designs so that the UEs composing them will be active a very small fraction of the time and

remain idle during the large remainder fraction. By exploiting such a sparse activity pattern, it is reasonable to consider that resources can be commonly shared between UEs without significant degradations in the RA success, since the resources are likely to not be simultaneously utilized.

Such a sparsity characteristic is of prime interest in the CS theory. Hence, In the context of NOMA for mMTC and uRLLC, the framework of CS-NOMA (see Sec. A.III.2) appears to be very promising for GFRA. In particular, the challenges exposed in Sec. A.IV will be addressed using the framework of bayesian CS, that is introduced in Chap. C.

II System model for grant-free random access

As stated in Chap. A, the candidate RA policy for 5G is GFRA assisted by NOMA. We describe a GFRA (2-step procedure) scheme based on CS-NOMA for the detection part.

In this section, a general mathematical model for uplink transmission in GFRA is introduced. This model will form the basis for optimizing the design of GFRA in the main technical contributions in Chap. D and Chap. E. The model is based on the 5G-NR physical layer in [3], which is overviewed next.

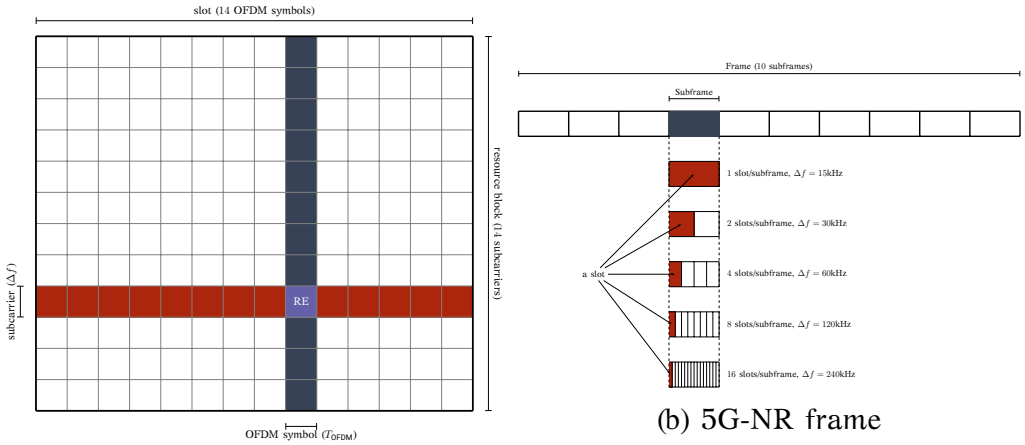
We consider a cellular wireless network with a K-antennas AP serving N single-antenna UEs. We focus on the UL communication direction.

II.1 A review of 5G New Radio physical layer

In 5G-NR, the physical layer is based on OFDM waveforms where the data are modulated onto orthogonal subcarriers across multiple OFDM symbols. This form a so-called OFDM grid, illustrated in Fig. B.3 where a *resource element* corresponds to a time-frequency rectangle indexed by a subcarrier and an OFDM symbol.

In the frequency domain, 12 subcarriers corresponds to a *resource block*. For 5G-NR, there are two operating frequency ranges, denoted by FR1 (from 410MHz to 7.125GHz) and FR2 (24.25GHz to 52.6GHz, aka the mmWave band). Each frequency ranges allows for different channel bandwidths span by an integer number of RB.

In the time domain, the concatenation of 14 OFDM symbols (and their cyclic prefix) corresponds to a *slot*. Following the principles of OFDM,



(a) OFDM grid

(b) 5G-NR frame

Figure B.3: 5G-NR PHY structure.

it is well known that the subcarrier spacing Δf and the duration of an OFDM symbol T_{OFDM} are related by

$$\Delta f = \frac{1}{T_{\text{OFDM}}} \tag{B.1}$$

where the choice of Δf depends of the chosen numerology. From [3], the different numerology modes are given by the formula:

$$\Delta f = 2^{\mu} \times 15\text{kHz} \qquad \text{for} \qquad \mu \in \{0, 1, 2, 3, 4\}. \tag{B.2}$$

Depending on the numerology, a *subframe* contains 2^n slots and 10 subframes form a *frame*.

A transmission is performed over one or multiple antenna ports. An antenna port is defined as the channel that can be infered from any two symbols transmitted on this same antenna port [32]. An antenna port can then correspond to multiple physical antennas. For simplicity, we will assume that the term *antenna* implicitly refers to *antenna port*. Combined with OFDM, each antenna has its own OFDM grid.

II.2 Equivalent baseband uplink transmission

In GFRA, the primary task is for the UEs to reliably transmit their preamble to the AP. When the transmission of the preamble is performed over a time window corresponding to M OFDM symbols, the nth UE may transmit a (possibly random) complex signal denoted by $\mathbf{x}_n \in \mathbb{C}^M$. In what

follows, all the UEs are assumed to transmit on the same subcarrier such that M corresponds to the number of resource elements. This assumption is relevant in the context of IoT, particularly when powered by protocols such as NB-IoT [33] that may operate on a single carrier with narrowband transmissions [34]. The random matrix of all the transmitted signals is then

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix} \in \mathbb{C}^{M \times N}. \tag{B.3}$$

The signal matrix $\mathbf{X}$ is conveyed to the kth AP's antenna over a narrowband flat-fading channel. We denote by h_{nk} the equivalent baseband channel random coefficient between UE n and antenna k and by $\mathbf{h}_k = [\mathsf{h}_{nk}]_{n\in[N]}^{\mathsf{T}}$ the channel random vector. The equivalent baseband received signal thus consists in the linear combination of the UEs' signals

$$\mathbf{z}_k = \sum_{n\in[N]} \mathbf{x}_n \mathsf{h}_{nk} \tag{B.4}$$

$$= \mathbf{X}\mathbf{h}_k \tag{B.5}$$

in the absence of noise. When the channel is noisy, the signal received by the AP is modeled by

$$\mathbf{y}_k = \sum_{n\in[N]} \mathbf{x}_n \mathsf{h}_{nk} + \mathsf{w}_{mk} \tag{B.6}$$

$$= \mathbf{X}\mathbf{h}_k + \mathbf{w}_k \tag{B.7}$$

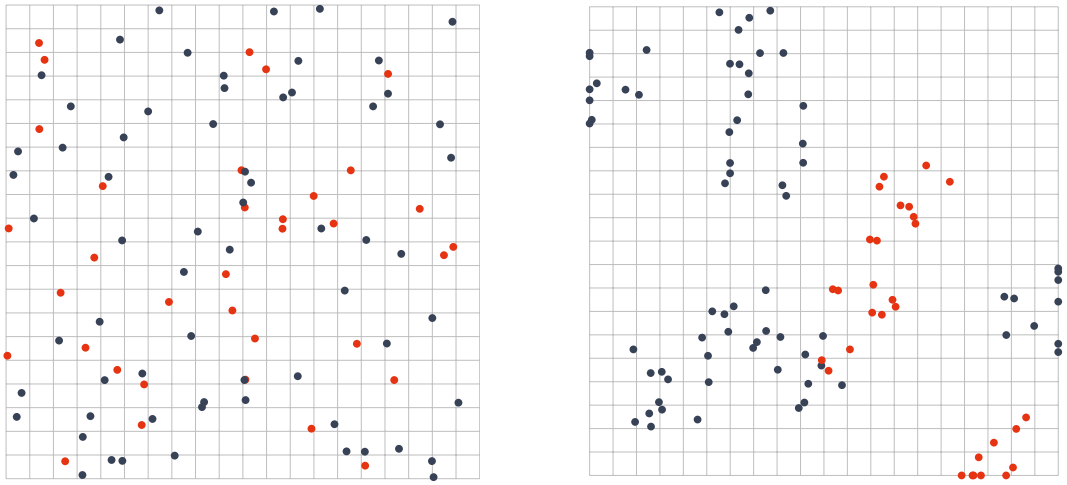
$$= \mathbf{z}_k + \mathbf{w}_k \tag{B.8}$$

where $\mathbf{w}_k = [\mathsf{w}_{mk}]_{m\in[M]}^{\mathsf{T}} \in \mathbb{C}^M$ is the random vector of baseband noise coefficients. Finally the random signal received over the K AP's antenna is the concatenation of the signals (B.8)

$$\mathbf{Z} = \mathbf{X}\mathbf{H} \qquad \text{(noiseless)} \tag{B.9}$$

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \qquad \text{(noisy)} \tag{B.10}$$

where $\mathbf{H} = [\mathbf{h}_k]_{k\in[K]} \in \mathbb{C}^{N\times K}$, $\mathbf{W} = [\mathbf{w}_k]_{k\in[K]} \in \mathbb{C}^{M\times K}$, $\mathbf{Z} = [\mathbf{z}_k]_{k\in[K]} \in \mathbb{C}^{M\times K}$ and $\mathbf{Y} = [\mathbf{y}_k]_{k\in[K]} \in \mathbb{C}^{M\times K}$.



(a) Independent (b) Group-sparse

Figure B.4: Independent and group sparse activity patterns. ● indicates active UE and ● indicates inactive UE. In Fig. B.4a, the active UEs are uniformly located on the grid whereas in Fig. B.4b it is possible to identify groups of active UEs.

II.3 Model of the activity pattern

II.3.a Activity states and pattern

During the RA window, the N UEs belong either to the sets of *active* or *inactive* UEs, depending on whether UEs needs to initiate a RA procedure before transmitting data payloads. The *state*, active or inactive, of the UE $n \in [N]$ is denoted by a binary random variable $\mathsf{s}_n \in \{0, 1\}$ where

$$\begin{cases} \mathsf{s}_n = 0 & \Rightarrow \text{UE is inactive} \\ \mathsf{s}_n = 1 & \Rightarrow \text{UE is active} \end{cases} \tag{B.11}$$

The state of all the UEs during the RA window, is denoted by the binary random vector

$$\mathbf{s} = \begin{bmatrix} \mathsf{s}_1 & \dots & \mathsf{s}_N \end{bmatrix}^{\mathsf{T}}. \tag{B.12}$$

which is called the random activity *pattern*. The probability distribution of $\mathbf{s}$ is described by its probability mass function (pmf) $\mathbb{P}_{\mathbf{s}}$

The characterization of the pmf is of prime importance for GFRA since the probability that a UE succeed its RA will depend on the joint activity of all the UEs in the cell. A number of different models for the pmf $\mathbb{P}_{\mathbf{s}}$

have been considered. While the most common model is independent activity, this can be limiting, as it will be explained in the following survey of relevant models.

II.3.b Probabilistic considerations

Modeling properly the activity pattern is of a prime matter for uRLLC and mMTC at different levels. The most natural consideration holds for activity detection of the UEs. If the considered network is heavily populated but experiences sporadic traffic, detection of active UEs may leverage this behavior. The same applies to resource allocation and channel estimation that can benefit from a sparse activity pattern. It becomes then important to known what kind of sparsity we must deal with.

In the following paragraphs, we review some existing models of the activity pattern based on its pmf $\mathbb{P}_{\mathsf{s}}(\boldsymbol{s})$ and the state pmfs of the UEs $\{\mathbb{P}_{\mathsf{s}_n}(s_n)\}_{n\in[N]}$.

Independent activity states The activity pattern is generally considered to be a collection of the independent random states s_n. This suits well networks where UEs' activities are not related by any means and so may be considered as mutually independent. It translates into the following factorization

$$\mathbb{P}_{\mathsf{s}}(\boldsymbol{s}) = \prod_{n=1}^{N} \mathbb{P}_{\mathsf{s}_n}(s_n). \tag{B.13}$$

A consequence of independent activity states is that the covariance, and hence correlation, is zero

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = 0. \tag{B.14}$$

An example of independent activity is shown in Fig. B.4a.

Independent group-sparse activity pattern A more general model assumes that UEs' activities is *group-sparse*. UEs depends on an underlying structure spliting them into groups. All the UEs that belong to the same group will always be active all together during a RA window. Such a model encapsulates networks in which multiple UEs share a common activity trigger, turning on when an event of interest occurs, as illustrated in Fig. B.4b.

Denotes by $G \in \mathbb{N}$ the number of groups and by $\{\mathfrak{G}_g\}_{g\in[G]}$ the block

index sets such that

$$\forall g \in [G],\ \mathfrak{G}_g \subseteq [N], \qquad \text{(index subset)} \tag{B.15}$$

$$\bigcup_{g=1}^{G} \mathfrak{G}_g = [N], \qquad \text{(union is index set)} \tag{B.16}$$

$$\forall (g, g') \in [G]^2, g \neq g', \mathfrak{G}_g \cap \mathfrak{G}_{g'} = \emptyset \qquad \text{(no overlap).} \tag{B.17}$$

Denoting by $\mathbf{s}^{\mathfrak{G}} = [\mathsf{s}_g^{\mathfrak{G}}]_{g\in[G]}^{\mathsf{T}}$ the activity pattern of the groups, it formally reads

$$\mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) = \prod_{g=1}^{G} \mathbb{P}_{\mathsf{s}_g^{\mathfrak{G}}}(s_g^{\mathfrak{G}}) \prod_{n\in\mathfrak{G}_g} \delta(s_n - s_g^{\mathfrak{G}}) \tag{B.18}$$

or in a simpler form

$$\mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) = \mathbb{P}_{\mathbf{s}^{\mathfrak{G}}}(\boldsymbol{s}^{\mathfrak{G}}) = \prod_{g=1}^{G} \mathbb{P}_{\mathsf{s}_g^{\mathfrak{G}}}(s_g^{\mathfrak{G}}). \tag{B.19}$$

From (B.18), it is clear that such factorization leads to a zero-probability for an activity pattern $\boldsymbol{s} \in \{0,1\}^N$ that does not exhibit a block structure. Also, the no-overlap condition enforces a group activity to be independent from the other groups. Finally, when $G = N$, the independent group-sparse model reduces to the independent activity model presented in the previous paragraph.

In the reminder of this manuscript, we adopt the shorthand notation $\mathbf{s}_g = [\mathsf{s}_{i+\sum_{g'<g}|\mathfrak{G}_{g'}|}]_{i\in[|\mathfrak{G}_g|]}^{\mathsf{T}}$ to denote the random subvector of $\mathbf{s}$ corresponding to the gth group such that

$$\mathbf{s} = \begin{bmatrix} \mathsf{s}_1 \\ \vdots \\ \mathsf{s}_{|\mathfrak{G}_1|} \\ \hline \vdots \\ \hline \mathsf{s}_{1+\sum_{g'=1}^{G-1}|\mathfrak{G}_{g'}|} \\ \vdots \\ \mathsf{s}_{|\mathfrak{G}_G|} \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1 \\ \hline \vdots \\ \hline \mathbf{s}_G \end{bmatrix} \tag{B.20}$$

The pairwise correlation between two activity states is then

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = 0 \qquad \text{if} \qquad \nexists g \in [G],\ (n, n') \in \mathfrak{G}_g^2, \tag{B.21}$$

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = 1 \qquad \text{if} \qquad \exists g \in [G],\ (n, n') \in \mathfrak{G}_g^2. \tag{B.22}$$

General group-sparse activity pattern A direct generalization of the aforementioned independent group-sparse model considers *overlaps* between groups. With the same notations as above, consider the relaxed conditions

$$\forall g \in [G],\ \mathfrak{G}_g \subseteq [N], \qquad \text{(index subset)} \tag{B.23}$$

$$\bigcup_{g=1}^{G} \mathfrak{G}_g = [N], \qquad \text{(union is index set)} \tag{B.24}$$

to which we add the dual formalization

$$\forall n \in [N],\ \exists g \in [G],\ n \in \mathfrak{G}_g \qquad \text{(index in at least 1 group)} \tag{B.25}$$

$$\forall n \in [N],\ \mathfrak{N}_n = \{g \in [G] \mid n \in \mathfrak{G}_g\} \qquad \text{(set group index)} \tag{B.26}$$

leading to a factorization of the activity pattern pmf into

$$\mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) = \sum_{\boldsymbol{s}^{\mathfrak{G}} \in \{0,1\}^G} \mathbb{P}_{\mathbf{s}|\mathbf{s}^{\mathfrak{G}}}(\boldsymbol{s} \mid \boldsymbol{s}^{\mathfrak{G}}) \mathbb{P}_{\mathbf{s}^{\mathfrak{G}}}(\boldsymbol{s}^{\mathfrak{G}}) \tag{B.27}$$

$$= \sum_{\boldsymbol{s}^{\mathfrak{G}} \in \{0,1\}^G} \left(\prod_{n=1}^{N} \mathbb{P}_{\mathsf{s}_n|\mathbf{s}^{\mathfrak{G}}_{\mathfrak{N}_n}}(s_n \mid \boldsymbol{s}^{\mathfrak{G}}_{\mathfrak{N}_n}) \right) \left(\prod_{g=1}^{G} \mathbb{P}_{\mathsf{s}^{\mathfrak{G}}_g}(s^{\mathfrak{G}}_g) \right) \tag{B.28}$$

$$= \sum_{\boldsymbol{s}^{\mathfrak{G}} \in \{0,1\}^G} \left(\prod_{n=1}^{N} \delta(s_n - \max_{g \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g)) \right) \left(\prod_{g=1}^{G} \mathbb{P}_{\mathsf{s}^{\mathfrak{G}}_g}(s^{\mathfrak{G}}_g) \right) \tag{B.29}$$

The non-zero terms are decided by $\delta(s_n - \max_{b \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g))$ which cancels when $\max_{b \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g) \neq s_n$ i.e.

- one of the group states in $\mathfrak{N}_n$ is active, $\max_{b \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g) = 1$, but the UE state is inactive $s_n = 0$;
- all of the group states in $\mathfrak{N}_n$ are inactive, $\max_{b \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g) = 0$, but the UE state is active $s_n = 1$;

and equals 1 when $\max_{b \in \mathfrak{N}_n}(s^{\mathfrak{G}}_g) = s_n$ i.e.

- one of the group states in $\mathfrak{N}_n$ is active, $\max_{b\in\mathfrak{N}_n}(s_g^{\mathfrak{G}}) = 1$, and the UE state is active $s_n = 1$;
- all of the group states in $\mathfrak{N}_n$ are inactive, $\max_{b\in\mathfrak{N}_n}(s_g^{\mathfrak{G}}) = 0$, and the UE state is inactive $s_n = 0$.

The main difference with the previous independent group-sparse model, is the possibility to describe the activity states with an interleaved structure, inducing thiner correlation design between the UE activity random states. Practically speaking and building on this idea of common triggers, this group-sparse model applies well to networks where UEs may have the same underlying structure has in the independent group-sparse model but with additional triggers at larger scales that can spans UEs from different groups.

Multivariate Bernoulli model The multivariate Bernoulli model was studied in [35] and requires no particular group structure. The random activity pattern has support $\{0,1\}^N$ and so its pmf may be written as

$$\mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) = \prod_{\boldsymbol{s}'\in\{0,1\}^N} p(\boldsymbol{s}')^{\prod_{n=1}^N s_n^{s_n'}(1-s_n)^{(1-s_n')}} \tag{B.30}$$

where $\{p(\boldsymbol{s}')\}_{\boldsymbol{s}'\in\{0,1\}^N} \in [0,1]^{2^N}$ is a collection of 2^N probabilities, one for each of the possible patterns in $\{0,1\}^N$. This model is the most general one when it comes to modeling the activity pattern but is also the most difficult to deal with since it requires the knowledge of parameters that grow exponentially with the number N of UEs. Since the focus is on mMTC and uRLLC, N will be large enough to make this model unusable, although it is the one that captures the finest correlated activity structures.

A qualitative comparison of the multivariate Bernoulli model is summarized in Fig. B.5 in terms of capability to capture correlated activity w.r.t. to the complexity involved in the description of the underlying model.

II.3.c Correlated activity

So far a review of some existing activity pattern statistical models has been done. With next generation mMTC and uRLLC networks, a key consideration is related to correlated activity. Applications involving dense and large networks such as smart cities, (I)IoT or V2X are typically concerned by correlated activity.

In smart cities, urban areas are monitored by a network of wireless sensors. Those sensors have been deployed about several locations of interest. Each location may be monitored by a bunch of sensors such that it is correctly covered.

For V2X, it may be expected that area with heavily loaded traffic, especially during certain hours of a day, will produce correlated communications.

Large-scale facilities will benefit from IoT deployment for ensuring the same type of tasks as smart cities, namely monitoring and sensing. For instance, an industrial plant may be equipped with wireless sensors monitoring automated machines and industrial robots. When an event of interest occurs on an equipment, the sensors that monitor it will likelily simultaneously initiate a RA procedure before sending sensored data.

In each of these three scenarios, the notion of correlated activity naturally arises due to the physical nature of the network. From the previous paragraphs we have seen that it is not an easy task to model simple and not fully correlated activity states.

III Design of the preamble matrix

When a UE is active and attempts to access the AP, it transmits its preamble $\boldsymbol{p}_n \in \mathbb{C}^M$; when it is inactive it does not transmit anything. The preamble matrix of all the preambles is known by the AP and is obtained by concatenating them as

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{p}_1 & \dots & \boldsymbol{p}_N \end{bmatrix} \in \mathbb{C}^{M \times N} \tag{B.31}$$

Based on the transmission model in Sec. B.II.2, the AP receives

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{P}\,\mathrm{diag}(\mathbf{s})\tilde{\mathbf{H}} + \mathbf{W} && && (\text{B.32})\\ &= \tilde{\mathbf{P}}\tilde{\mathbf{H}} + \mathbf{W} && \text{with} \quad \tilde{\mathbf{P}} = \boldsymbol{P}\,\mathrm{diag}(\mathbf{s}) && (\text{B.33})\\ &= \boldsymbol{P}\mathbf{H} + \mathbf{W} && \text{with} \quad \mathbf{H} = \mathrm{diag}(\mathbf{s})\tilde{\mathbf{H}} && (\text{B.34}) \end{aligned}$$

where the rows of the random channel matrix $\mathbf{H}$ are zero when the corresponding UEs are inactive.

As stated in Sec. B.I.2.a, ZC sequences are used for generating the RA preambles. Other ways of constructing these preambles may be considered, especially when considering CS-NOMA. The preamble matrix $\boldsymbol{P}$ is

generally refered to a frame in frame theory and the sensing or measurement matrix in CS.

An important criterium for designing good measurement matrix $\boldsymbol{P} \in \mathbb{C}^{M\times N}$ is given from the CS theory and is known as the restricted isometry property (RIP) [36], [37] and formalizes as

$$\forall \boldsymbol{h} \in \mathbb{C}^N,\ \|\boldsymbol{h}\|_0 \leq \nu,\ (1-\delta_\nu)\|\boldsymbol{h}\|_2^2 \leq \|\boldsymbol{P}\boldsymbol{h}\|_2^2 \leq (1+\delta_\nu)\|\boldsymbol{h}\|_2^2 \tag{B.35}$$

where $\nu \in [N]$ is the sparsity level and $\delta_\nu > 0$ is the restricted isometry constant (RIC). If $\boldsymbol{P}$ satisfies the RIP for some ν and RIC δ_ν, exact recovery of ν'-sparse signals with $\nu' \leq \nu$ can be achieved for some specific ν and ν'. Interpretations of the RIP includes that all the eigenvalues of $\boldsymbol{P}$ must be contained in a disc centered at 1 with radius δ_ν and that any submatrices $\boldsymbol{P}_\nu$ formed with ν columns of $\boldsymbol{P}$ slightly deform the non-zero subvector $\boldsymbol{h}_\nu$ of $\boldsymbol{h}$.

General deterministic constructions of RIP-matrices [38], [39], and hence preambles, remains an open problem, especially because computing the RIC is NP-hard. However, deterministic designs were considered based on the less complex design criterium of coherence minimization. The *coherence* of a matrix is defined as

$$\mu(\boldsymbol{P}) = \max_{1\leq n<n'\leq N} \left|\boldsymbol{p}_n^{\mathsf{H}}\boldsymbol{p}_{n'}\right| \tag{B.36}$$

and good matrices, namely *Grassmannian frames* [40], [41], attempts to solve the following packing problem

$$\boldsymbol{P}^\star = \underset{\boldsymbol{P}\in\mathbb{C}^{M\times N}}{\arg\min}\, \mu(\boldsymbol{P}). \tag{B.37}$$

Existing constructions of such Grassmannian frames are usually classified into equiangular (tight) frames [42]–[45]. Designs of such matrices have received a lot of attention, formerly in the context multi-antenna communication systems [41], [46]–[48] and more recently in the context of multi-user detection [49], [50], i.e. data transmission in presence of a known number of users (which should not be confused with active user detection) and grant-free communications [51].

On the other hand, random designs of RIP-matrices have been proved to satisfy the RIP with high probability, especially for subgaussian matrices [37, and references therein]. Examples include but are not limited

to gaussian ($p_{mn} \sim \mathsf{CNorm}(0, 1/M)$), normalized gaussian (gaussian with unit-norm columns) and Rademacher matrices ($p_{mn} \sim \mathsf{Unif}\Big(\{-1/\sqrt{M}, +1/\sqrt{M}\}\Big)$).

IV Active user detection and channel estimation

We review in this section the different signal processing problems and decision making that can be addressed within the transmission and RA scheme described in Sec. B.II.

IV.1 Active user detection

Based on (B.33), if we assume that the channel $\tilde{\mathbf{H}}$ is known to be $\tilde{\boldsymbol{H}}$, the problem of active user detection consists in estimating which UE is active i.e. find the support of $\tilde{\mathbf{P}}$, or equivalently, the support of the activity pattern $\mathbf{s}$. The theoretical optimal detector is well formulated as the maximum a posteriori (MAP) detector

$$\boldsymbol{s}^\star = \underset{\boldsymbol{s}\in\{0,1\}^N}{\arg\max}\, \mathbb{P}_{\mathbf{s}|\mathbf{Y},\tilde{\mathbf{H}}}(\boldsymbol{s} \mid \boldsymbol{Y}, \tilde{\boldsymbol{H}}) \tag{B.38}$$

where the posterior pmf, once the Bayes' theorem is applied, factorizes as

$$\mathbb{P}_{\mathbf{s}|\mathbf{Y},\tilde{\mathbf{H}}}(\boldsymbol{s} \mid \boldsymbol{Y}, \tilde{\boldsymbol{H}}) = f_{\mathbf{Y}|\tilde{\mathbf{H}}}(\boldsymbol{Y} \mid \tilde{\boldsymbol{H}})^{-1} f_{\mathbf{Y}|\mathbf{s},\tilde{\mathbf{H}}}(\boldsymbol{Y} \mid \boldsymbol{s}, \tilde{\boldsymbol{H}}) \mathbb{P}_{\mathbf{s}}(\boldsymbol{s}). \tag{B.39}$$

Note that $\mathbb{P}_{\mathbf{s}}(\boldsymbol{s})$ is not conditioned on $\tilde{\mathbf{H}} = \tilde{\boldsymbol{H}}$ since $\mathbf{s}$ and $\tilde{\mathbf{H}}$ are independent.

Complexity It is clear from (B.38), that there is $O(2^N)$ tests to find the optimal activity pattern $\boldsymbol{s}^\star$. When N becomes large, the number of tests becomes prohibitive to be solved by classical computing means. Generally this problem is known to be NP-hard.

Metrics In order to assess the performance of the activity detector, the usual metrics of interest are the following.

- The user error rate (UER), which measures the probability to incorrectly detect a UE's state and is given by

$$P_{\mathsf{ue}} = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(\hat{\mathsf{s}}_n \neq \mathsf{s}_n)\right]. \tag{B.40}$$

- The false alarm rate (FAR), which measures the probability to incorrectly detect an inactive UE, i.e. an inactive UE is detected active. It is given by

$$P_{\mathsf{fa}} = \mathbb{E}\left[\frac{1}{N - \|\mathbf{s}\|_0} \sum_{n=1}^{N} \mathbb{1}(\hat{\mathsf{s}}_n > \mathsf{s}_n)\right]. \tag{B.41}$$

- The missed detection rate (MDR), which measures the probability to incorrectly detect an active UE, u.e. an active UE is detected inactive. It is given by

$$P_{\mathsf{md}} = \mathbb{E}\left[\frac{1}{\|\mathbf{s}\|_0} \sum_{n=1}^{N} \mathbb{1}(\hat{\mathsf{s}}_n < \mathsf{s}_n)\right]. \tag{B.42}$$

IV.2 Channel estimation

The problem of estimating the channel is crucial for mMTC and uRLLC systems for reliably transmitting data. Given an activity pattern $\mathbf{s} = \boldsymbol{s}$ and the corresponding realization $\mathbf{Y} = \boldsymbol{Y}$ given by (B.34), a classical estimator of the channel matrix the minimum mean squared error (MMSE) estimator

$$\boldsymbol{H}^{\star} = \underset{\boldsymbol{H} \in \mathbb{C}^{N \times K}}{\arg\min}\, \mathbb{E}\left[\|\boldsymbol{H} - \mathbf{H}\|_2^2 \mid \mathbf{Y} = \boldsymbol{Y}, \mathbf{s} = \boldsymbol{s}\right] \tag{B.43}$$

which is obtained by finding the matrix $\boldsymbol{H}^{\star}$ minimizing the expected estimation error using euclidean distance, represented here by the ℓ_2-norm of the difference. Under the assumptions that $\mathbb{E}[\boldsymbol{H} \mid \mathbf{Y} = \boldsymbol{Y}]$ and $\mathbb{E}\left[\mathbf{H}^{\mathsf{H}}\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}\right]$ exist, one can find such a minimum $\boldsymbol{H}^{\star}$. Indeed,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{H} - \mathbf{H}\|_2^2 \mid \mathbf{Y} = \boldsymbol{Y}, \mathbf{s} = \boldsymbol{s}\right] = {} & \operatorname{tr}(\boldsymbol{H}^{\mathsf{H}}\boldsymbol{H}) \\ & - \operatorname{tr}(\boldsymbol{H}^{\mathsf{H}}\mathbb{E}[\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}]) \\ & - \operatorname{tr}(\mathbb{E}[\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}]^{\mathsf{H}}\boldsymbol{H}) \\ & + \operatorname{tr}(\mathbb{E}\left[\mathbf{H}^{\mathsf{H}}\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}\right])\end{aligned} \tag{B.44}$$

Differentiating w.r.t. $\boldsymbol{H}^{\mathsf{H}}$ and zeroing leads to the optimal $\boldsymbol{H}^{\star}$

$$\boldsymbol{H}^{\star} - \mathbb{E}[\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}, \mathbf{s} = \boldsymbol{s}] = 0, \tag{B.45}$$

$$\boldsymbol{H}^{\star} = \mathbb{E}[\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}, \mathbf{s} = \boldsymbol{s}], \tag{B.46}$$

which is nothing but the a posteriori expected value of $\mathbf{H}$ given $\mathbf{Y} = \boldsymbol{Y}$. Its computation reads

$$\boldsymbol{H}^{\star} = \mathbb{E}[\mathbf{H} \mid \mathbf{Y} = \boldsymbol{Y}, \mathbf{s} = \boldsymbol{s}] \tag{B.47}$$

$$= \int_{\mathbb{C}^{N \times K}} \boldsymbol{H} f_{\mathbf{H}|\mathbf{Y}}(\boldsymbol{H} \mid \boldsymbol{Y}, \boldsymbol{s}) \mathrm{d}\boldsymbol{H} \tag{B.48}$$

Using Bayes' theorem, the posterior expectation becomes

$$\boldsymbol{H}^{\star} = f_{\mathbf{Y}|\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{s})^{-1} \int_{\mathbb{C}^{N \times K}} \boldsymbol{H} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}, \boldsymbol{s}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathrm{d}\boldsymbol{H} \tag{B.49}$$

$$= \frac{\int_{\mathbb{C}^{N \times K}} \boldsymbol{H} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}, \boldsymbol{s}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathrm{d}\boldsymbol{H}}{\int_{\mathbb{C}^{N \times K}} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}, \boldsymbol{s}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathrm{d}\boldsymbol{H}} \tag{B.50}$$

Complexity Computing (B.50) can be achieved in a closed form if the posterior or the prior and likelihood densities allow a tractable derivation. If it is not the case, the high dimensional integral can be computed by Monte-Carlo integration but the complexity scales prohibitively as the number of samples $\{\boldsymbol{H}_i\}_i$ grows.

Metrics Measuring the quality of the MMSE estimation is often accomplished by evaluating the normalized mean squared error (NMSE)

$$\mathbb{E}\left[\frac{\left\|\mathbf{H} - \hat{\mathbf{H}}\right\|_2^2}{\|\mathbf{H}\|_2^2}\right] = \iint_{\mathbb{C}^{N \times K} \times \mathbb{C}^{N \times K}} \frac{\left\|\boldsymbol{H} - \hat{\boldsymbol{H}}\right\|_2^2}{\|\boldsymbol{H}\|_2^2} f_{\hat{\mathbf{H}}|\mathbf{H}}(\hat{\boldsymbol{H}} \mid \boldsymbol{H}) f_{\mathbf{H}}(\boldsymbol{H}) \mathrm{d}\hat{\boldsymbol{H}} \mathrm{d}\boldsymbol{H} \tag{B.51}$$

where the expectation is taken w.r.t. to the true channel random matrix $\mathbf{H}$ and the estimated one $\hat{\mathbf{H}}$.

IV.3 Joint problem statement

The joint problem of AUDaCE is more difficult to setup. A natural formulation of the problem would be under a MAP estimator of the form

$$\boldsymbol{s}^\star, \boldsymbol{H}^\star = \underset{\boldsymbol{s}\in\{0,1\}^N, \boldsymbol{H}\in\mathbb{C}^{N\times K}}{\arg\max} f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) \tag{B.52}$$

Another one would be to consider the proxy block matrix $\mathbf{V}(\mathbf{s},\mathbf{H}) = [\mathbf{s},\mathbf{H}]$ and its MMSE estimator

$$\boldsymbol{V}(\mathbf{s},\mathbf{H})^\star = \mathbb{E}[\mathbf{V}(\mathbf{s},\mathbf{H})] = \int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \boldsymbol{V}(\boldsymbol{s},\boldsymbol{H}) f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s},\boldsymbol{H} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{H}. \tag{B.53}$$

Both estimators rely on the joint posterior density

$$\begin{aligned} f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s},\boldsymbol{H} \mid \boldsymbol{Y}) &= f_{\mathbf{Y}}(\boldsymbol{Y})^{-1} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) && \text{(B.54)} \\ &= \frac{f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}}(\boldsymbol{s})}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) \mathrm{d}\boldsymbol{H}} && \text{(B.55)} \end{aligned}$$

where we use the fact that the system variables $\mathbf{s}, \mathbf{H}, \mathbf{Y}$ form the following Markov chain

$$\mathbf{s} \to \mathbf{H} \to \mathbf{Y}. \tag{B.56}$$

Note that, depending on the likelihood and priors densities, the optima (B.52) and (B.53) may be different.

A final way for considering the AUDaCE problem consists in considering a variant of the density (B.55) given by

$$f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s},\boldsymbol{H} \mid \boldsymbol{Y}; \kappa) = \frac{f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H})^\kappa f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s})^\kappa \mathbb{P}_{\mathbf{s}}(\boldsymbol{s})^\kappa}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y} \mid \boldsymbol{H})^\kappa f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s})^\kappa \mathbb{P}_{\mathbf{s}}(\boldsymbol{s})^\kappa \mathrm{d}\boldsymbol{H}} \tag{B.57}$$

where $\kappa = 1$ corresponds to the density for the MMSE case and $\kappa \to +\infty$ corresponds to the MAP since the density will concentrate about its mode [52], [53].

Complexity The AUDaCE problem is a CS problem since it aims at jointly estimating the support of some input signal (the channel matrix $\boldsymbol{H}^\star$) and its support (the activity pattern $\boldsymbol{s}^\star$). Therefore, it is NP-hard and cannot

be solved by any of the two methods above.

V Other relevant aspects for grant-free random access

Even if the focus of this manuscript is on GFRA, there exist other relevant problem strongly connected to it.

As mentioned in Sec. B.I.1.c, it is possible to consider grant-free uplink data transmission which is very similar to GFRA except that the preamble matrix $\boldsymbol{P}$ is changed to a random signal matrix $\mathbf{X}$ that may also contains few information symbols. Hence, the problem of joint AUDaCE introduced in Sec. B.IV requires a joint data recovery and AUDaCE where the transmitted signal matrix $\mathbf{X} = \boldsymbol{X}$ needs to be estimated along the channel matrix and activity pattern based on the received signal $\boldsymbol{Y}$. This approach was considered in [54] with an independent activity pattern.

Another important problem is the optimization of resource allocation in the context of GFRA with possibly correlated activity of the UEs. In [55], the rate allocation of UEs is investigated and performed based on some heuristics assuming their activity is correlated. The knowledge of the underlying correlation structure is described based on the joint pmf of the activity pattern. In [56], a similar problem is studied where a matrix of pairwise activity probabilities is optimized to maximize the expected throughput.

VI Conclusion

Based on the review of RA in 5G, it appears that GFRA is a key enabler to mMTC and uRLLC, especially to fasten further data transmissions by reducing control overhead. In particular, AUDaCE plays an important role in GFRA to reliably identify active UEs and preparing later transmission by estimating the channel between the AP and the UEs. However, AUDaCE was only considered with independent and group-sparse activity patterns due to tractability in their model. In Chapters D and E, we will consider AUDaCE applied to GFRA with more flexible correlation model. This will be made possible using CS-NOMA techniques overviewed in Chap. C.

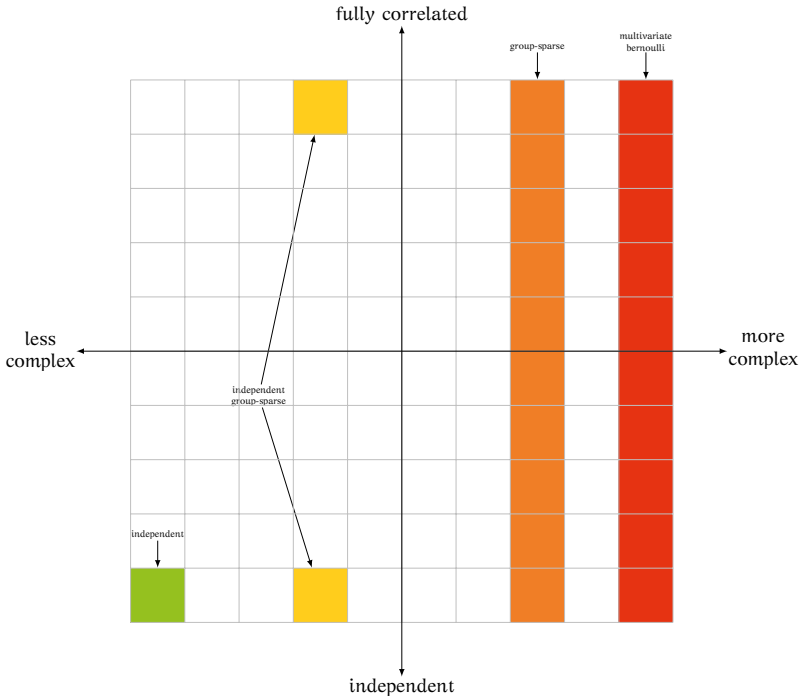


Figure B.5: Qualitative comparison of different activity pattern models. The independent model is the simplest one, both in terms of modeling complexity (bernoulli distribution) and correlation (independence between the activity states). The independent group-sparse activity pattern model requires the introduction of group states variables, each of them connected to non-overlapping subsets of the activity states. The more general group-sparse model is similar but allows the group states to share common activity state leading to potentially very dependent and correlated structure between the activity states. Finally, the multivariate bernoulli is the most complete model since it theoretically allows for any dependence structure between the activity states, at the cost of a very complex modeling.

C

ALGORITHMIC BACKGROUND OF BAYESIAN COMPRESSED SENSING

This chapter provides an overview of the theoretical and algorithmic tools that will be used to address the joint AUDaCE problem in the context of GFRA based on CS-NOMA.

In Sec. C.I, an overview of the CS theory is presented emphasizing non-bayesian methods. In Sec. C.II, we shift to the bayesian framework introducing the belief propagation (BP) algorithm and the graphical models that will be essential in the understanding of the models of Chapters D and E. Finally, the framework of AMP and its generalization, which is at the heart of the solution proposed in Chapters D and E, is presented in Sec. C.III.

I NON-BAYESIAN COMPRESSED SENSING

A classical signal processing problem is CS [37], [57]. This problem consists in the recovery of a sparse signal $\boldsymbol{x} \in \mathbb{C}^N$ from an observed signal $\boldsymbol{y} \in \mathbb{C}^M$. The transformation of $\boldsymbol{x}$ into $\boldsymbol{y}$ is described by the following

generalized linear model:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{x}) \tag{C.1}$$

where $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ is the so-called measurement matrix and $\boldsymbol{f} : \mathbb{C}^M \to \mathbb{C}^N$ is some non-linear mapping acting component-wisely on its argument. A classical particular case is that of the noisy channel

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \tag{C.2}$$

where $\boldsymbol{w} \in \mathbb{C}^M$ is a noise vector. Under the assumption that $M < N$, this problem is underdetermined and may be impossible to solve for a general scenario. However, given that $\boldsymbol{x}$ is sparse, i.e. its support has few non-zero entries, one can expect the recovery to be made possible using this side information.

Let $(N, M) \in \mathbb{N}_+^2$ such that $M \leq N$. Let $(\boldsymbol{x}^\star, \boldsymbol{y}) \in \mathbb{C}^N \times \mathbb{C}^M$, $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ and $\boldsymbol{f} : \mathbb{C}^N \to \mathbb{C}^M$ such that

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star. \tag{C.3}$$

Denote by $\|\boldsymbol{x}\|_0$ the ℓ_0 pseudo-norm which counts the number of non-zero entries of $\boldsymbol{x}$. The noiseless CS problem is thus generally formulated as follows:

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}\in\mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_0 \qquad \text{s.t.} \qquad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{C.4}$$

When the observation is noisy, one can relax the equality constraint to rewrite the noiseless CS problem in its noisy version

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}\in\mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_0 \qquad \text{s.t.} \qquad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_p \leq \epsilon \tag{C.5}$$

where $p \in \mathbb{R}_+$. In both the noiseless (C.4) and noisy cases (C.5), solving such an constrained optimization problem is generally NP-hard and so cannot be achieved exactly by classical means.

I.1 Optimization methods

Techniques have been developped to approximate the optimal solution to (C.4) and (C.5) by relaxing the constraint on the ℓ_0 pseudo norm. All of

these methods consider relaxed variants of the problem (C.5) where the ℓ_0 pseudo-norm is replaced by a ℓ_q norm. Formally, let $(p, q) \in \mathbb{N}_*^2$ and $\epsilon > 0$. The relaxed noisy CS problem thus reads

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{C}^N}{\arg\min} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_q \qquad \text{s.t.} \qquad \|\boldsymbol{x}\|_p \leq \epsilon \tag{C.6}$$

The case $p = 1$ and $q = 2$

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{C}^N}{\arg\min} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \qquad \text{s.t.} \qquad \|\boldsymbol{x}\|_1 \leq \epsilon \tag{C.7}$$

is particularly famous and is known as the least absolute shrinkage and selection operator (LASSO) [58] where the reconstructed signal is allowed to have its ℓ_1-norm contained into the ℓ_1 ball $\{\boldsymbol{x} \in \mathbb{C}^N \mid \|\boldsymbol{x}\|_1 \leq \epsilon\}$.

Another approximation of (C.5) which is similar to LASSO, is basis pursuit denoising (BPDN) [59]

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_1 \qquad \text{s.t.} \qquad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon \tag{C.8}$$

where, contrary to LASSO, the objective function is the ℓ_1-norm of the input signal that promotes sparsity with the constraint that the ℓ_2-norm residual error lies in the ℓ_2 ball $\{\boldsymbol{x} \in \mathbb{C}^N \mid \|\boldsymbol{x}\|_2 \leq \epsilon\}$. There also exists a version of BPDN, more tailored to Eq. C.4, which adopts a stricter formulation

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_1 \qquad \text{s.t.} \qquad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{C.9}$$

where the constraint turns to be an equality [60], [61].

LASSO and BPDN benefit from turning the original CS problem into a convex one that makes it solvable.

I.2 Greedy methods

Greedy methods applied to CS usually consist of algorithms where the signal's support is iteratively recovered by activating one or many components at the same time given some criteria (maximum correlation, maximum absolute value, threshold exceeding, maximum likelihood) and the signal is estimated using the updated support. Well known greedy methods derivate from the matching pursuit (MP) algorithm [62] (orthogonal

MP (OMP) [63], stagewise OMP (StOMP) [64], multipath MP (MMP) [65], compressive sampling MP (CoSaMP) [66], fast bayesian MP (FBMP) [67]).

I.3 Iterative thresholding methods

Iterative thresholding methods for CS are similar to greedy methods since they update iteratively an estimate of the input signal. The update of such algorithms can usually be written as

$$\hat{\boldsymbol{x}}_i = \mathrm{T}(\hat{\boldsymbol{x}}_{i-1} + \boldsymbol{A}^{\mathsf{H}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})) \tag{C.10}$$

where T is a thresholding operator.

In the case of iterative hard thresholding algorithm (IHTA) [68], [69], the operator is *hard* and is given by

$$\mathrm{T}(\boldsymbol{u}; n_s) = \boldsymbol{u} \odot \Big(\underset{\boldsymbol{s}\in\{0,1\}^N, \|\boldsymbol{s}\|_0 = n_s}{\arg\max} \|\boldsymbol{u} \odot \boldsymbol{s}\|_1 \Big) \tag{C.11}$$

where it returns a sparse vector $\mathrm{T}(\boldsymbol{u}; n_s)$ equal to the vector with the n_s largest components of $\boldsymbol{u}$ and the others equal to 0.

In the case of iterative soft thresholding algorithm (ISTA) [69], [70], it is *soft* and reads

$$\mathrm{T}(\boldsymbol{u}; \boldsymbol{t}) = \frac{\boldsymbol{u}}{|\boldsymbol{u}|} \odot \max(|\boldsymbol{u}| - \boldsymbol{t}, \boldsymbol{0}) \tag{C.12}$$

where all the operations act component-wise on their input and $\boldsymbol{t}$ is a vector of real thresholds. The soft operator may be seen as a rectified linear unit (ReLU) component-wisely applied to its input, so that it sets to 0 the components with magnitude lower than their corresponding threshold and let the other unchanged. Another version of ISTA that fasten the convergence rate of the algorithm can be found in [71].

I.4 Bayesian compressed sensing

The methods presented so far roughly consists in recovering a signal based on minimization problems involving objective with a constraint on the fidelity w.r.t. to the output signal and a constraint related to the signal sparsity.

One could also consider CS problems with additional constraints. From the perspective of an optimization problem, this would translate

into adding supplementary constraints or penaltys to the the original problem Eq. C.5. As long as the new constraints explicitly depends on the signal $\boldsymbol{x}$ and its sparsity, adding them may be done in a simple way (e.g. consider group-LASSO [72]). However, when the new constraints are more intricate, e.g. because latent variables are involved or the dependence of the constraints on $\boldsymbol{x}$ is different from Eq. C.5, their inclusion may become an issue.

In the context of wireless communications, and especially in the context of GFRA and AUDaCE, problems are more naturally formulated in the Bayesian framework as in Eq. B.52 and Eq. B.53.

In fact, it is possible in some cases to cast CS problems into the Bayesian framework. For instance, equivalent MAP formulations of BPDN and LASSO [53] are

$$\hat{\boldsymbol{x}}_{\text{BPDN}} = \underset{\boldsymbol{x}\in\mathbb{C}^N}{\arg\max} \prod_{n\in[N]} \frac{\lambda}{2} e^{-\lambda|x_n|} \prod_{m\in[M]} \delta(y_m - [\boldsymbol{A}\boldsymbol{x}]_m) \tag{C.13}$$

$$\hat{\boldsymbol{x}}_{\text{LASSO}} = \underset{\boldsymbol{x}\in\mathbb{C}^N}{\arg\max} \prod_{n\in[N]} \frac{\lambda}{2} e^{-\lambda|x_n|} \prod_{m\in[M]} \frac{1}{\pi\tau_w} e^{-\frac{|y_m - [\boldsymbol{A}\boldsymbol{x}]_m|^2}{\tau_w}} \tag{C.14}$$

applying to Eqs. (C.7) and (C.9) monotonic mapping to their objective functions and constraints. In these formulations, the objective is composed of multiple factors, each corresponding to the LASSO and BPDN constraints. This is reminiscent of Eq. B.55 so that in Sec. C.II, this factorized structure is leveraged to introduce the framework of BP before introducing Bayesian techniques for the CS problem in Sec. C.III.

II Belief propagation

In this section, an introduction to graphical models and their use for BP is proposed. Graphical models are powerful when it comes to describe the dependency between the system variables and the factors that constitute the objective function of a bayesian problem. The latter may be solved using the framework of BP, which is a family of algorithms that relies on these graphical models to approximate the optimal solution.

II.1 Factor graphs

We start by introducing the graphical models.

(a) $f_{\mathsf{a}_1}(a_1)f_{\mathsf{a}_2}(a_2)f_{\mathsf{a}_3}(a_3)$ (b) $f_{\mathsf{a}_1}(a_1)f_{\mathsf{a}_2,\mathsf{a}_3}(a_2,a_3)$ (c) $f_{\mathsf{a}_1,\mathsf{a}_2,\mathsf{a}_3}(a_1,a_2,a_3)$

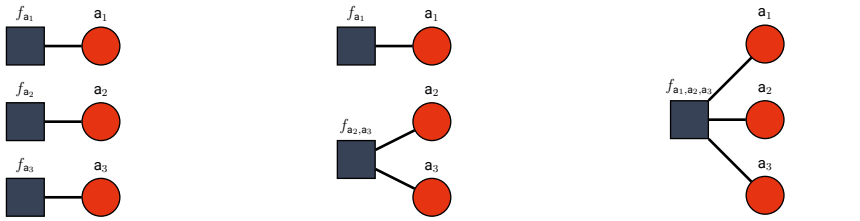


Figure C.1: Factor graphs based on the possible factorizations given in (C.17).

II.1.a Factorization

Consider a joint density function $f_{\mathbf{a}}$ of the random vector $\mathbf{a} = [a_i]_{i\in[I]}$ and assume it can be written into the following product

$$f_{\mathbf{a}}(\boldsymbol{a}) = \prod_{j\in[J]} f_j(\boldsymbol{a}_{\mathfrak{N}(j)}) \tag{C.15}$$

where f_j is the j-th factor of $f_{\mathbf{a}}(\boldsymbol{a})$ and

$$\mathfrak{N}(j) = \{i \in [I] \mid \mathbf{a}_i \text{ is a variable of } f_j\} \tag{C.16}$$

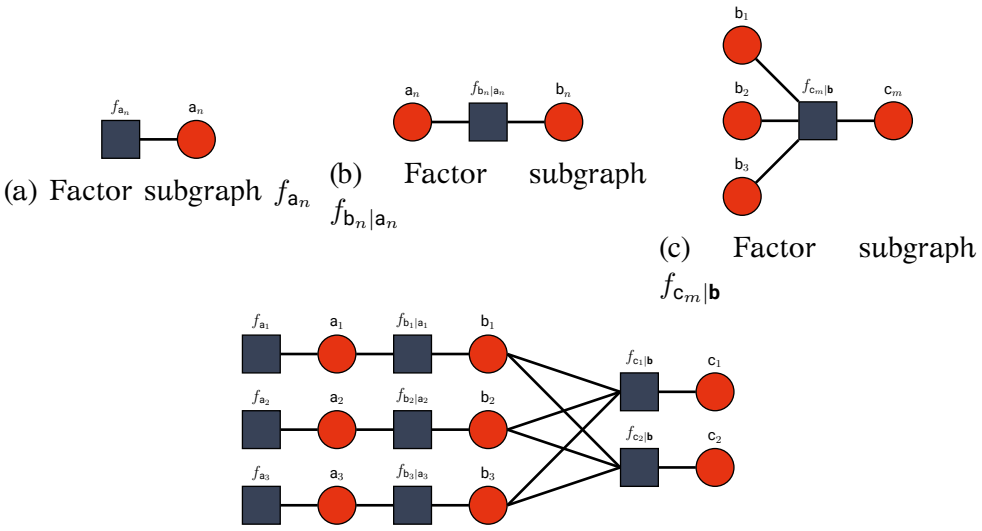
is its corresponding *neighbor* variable set. Note that such factorization of $f_{\mathbf{a}}$

1. always exists since one can consider the case $J = 1$ with $f_1 = f_{\mathbf{a}}$
2. may not be unique e.g. in the case $I = 3$ and $\{\mathsf{a}_i\}_{i\in[I]}$ are mutually independent, one can write 5 different factorizations

$$\begin{aligned} f_{\mathsf{a}_1,\mathsf{a}_2,\mathsf{a}_3}(a_1,a_2,a_3) &= f_{\mathsf{a}_1}(a_1)f_{\mathsf{a}_2}(a_2)f_{\mathsf{a}_3}(a_3) && \text{(C.17a)}\\ &= f_{\mathsf{a}_1}(a_1)f_{\mathsf{a}_2,\mathsf{a}_3}(a_2,a_3) && \text{(C.17b)}\\ &= f_{\mathsf{a}_2}(a_2)f_{\mathsf{a}_1,\mathsf{a}_3}(a_1,a_3) && \text{(C.17c)}\\ &= f_{\mathsf{a}_3}(a_3)f_{\mathsf{a}_1,\mathsf{a}_2}(a_1,a_2) && \text{(C.17d)} \end{aligned}$$

II.1.b Graphical model

A graphical model of the factorization (C.15) can be drawn in order to represent the functional dependencies of each factor j to its variables i and is named *factor graph*. Formally, a factor graph [73] $\mathfrak{G}(\mathfrak{F},\mathfrak{V},\mathfrak{E})$ is an undirected multipartite graph with vertices split into



(a) Factor subgraph f_{a_n} (b) Factor subgraph $f_{\mathsf{b}_n|\mathsf{a}_n}$ (c) Factor subgraph $f_{\mathsf{c}_m|\mathbf{b}}$

Figure C.2: Factor graph of Eq. C.19.

1. the set of *factor nodes* $\mathfrak{F} = \{f_j\}_{j\in[J]}$;
2. the set of *variable nodes* $\mathfrak{V} = \{\mathsf{a}_i\}_{i\in[I]}$.

An edge in $\mathfrak{E}$ connect the factor node $f_j \in \mathfrak{F}$ and the variable node a_i if and only if f_j is a density depending on a_i. As a convention, a factor and variable node will be respectively drawn like ■ and ●.

II.1.c Examples

Based on the definition given in Sec. C.II.1.b, we give some examples to illustrate the concept of factor graph.

The first simple examples of factor graphs are drawn in Fig. C.1 after Eq. C.17. A second and more complicated factor graph may be considered based on the following Markov chain

$$\mathbf{a} \in \mathbb{C}^N \to \mathbf{b} \in \mathbb{C}^N \to \mathbf{c} \in \mathbb{C}^M \tag{C.18}$$

where $M < N$. Assume that the joint density of the $N \times N \times M$ complex random variables factorizes as

$$f_{\mathbf{a},\mathbf{b},\mathbf{c}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \left(\prod_{n\in[N]} f_{\mathsf{a}_n}(a_n)\right)\left(\prod_{n\in[N]} f_{\mathsf{b}_n|\mathsf{a}_n}(b_n \mid a_n)\right)\left(\prod_{m\in[M]} f_{\mathsf{c}_m|\mathbf{b}}(c_m \mid \boldsymbol{b})\right) \tag{C.19}$$

where one can identify 3 groups of factors $\{f_{\mathsf{a}_n}\}_{n\in[N]}$, $\{f_{\mathsf{b}_n|\mathsf{a}_n}\}_{n\in[N]}$ and

$\{f_{\mathsf{c}_m|\mathbf{b}}\}_{m\in[M]}$.

Factor f_{a_n} Each factor in $\{f_{\mathsf{a}_n}\}_{n\in[N]}$ depends only on one random variable and would have the simple factor graph of Fig. C.2a.

Factor $f_{\mathsf{b}_n|\mathsf{a}_n}$ Each factor in $\{f_{\mathsf{b}_n|\mathsf{a}_n}\}_{n\in[N]}$ depends the random variable b_n but also on the conditioning random variable a_n. The corresponding factor subgraph is then depicted as in Fig. C.2b.

Factor $f_{\mathsf{c}_m|\mathbf{b}}$ Each factor in $\{f_{\mathsf{c}_m|\mathbf{b}}\}_{m\in[M]}$ depends on the random variable c_m and the complete conditionning random vector $\mathbf{b} = [\mathsf{b}]_{n\in[N]}$. The corresponding factor subgraph is drawn in Fig. C.2c.

Complete factor graph The complete factor graph Fig. C.2 is finally obtained by connecting each factor subgraphs based on the common variable nodes they share.

II.2 Bayesian inference

Factor graphs are very useful to describe the structure of bayesian inference problems since, as in Eqs. (B.52), (B.53), (C.13) and (C.14), the objective depends on some joint density on the system random variables. We briefly summarize the different types of bayesian inference problems that can be studied and explicitly show that all of them exhibit a factorization structure that will be leveraged in Sec. C.II.3.

Note In what follows, the problems are optimized on the variable $\boldsymbol{x}$ that belongs to any relevant set, the latter being implicit in the formulation. This set can be unconstrained (e.g. $\mathbb{R}^N, \mathbb{C}^N$) or constrained (e.g. $[0,1]^N$) depending on what would be a practical instance of the generic problems presented in Sections C.II.2.a to C.II.2.c.

II.2.a Maximum a posteriori inference

For instance, consider a MAP estimation problem [74] where one aims at estimating some hidden data $\mathbf{x}$ from the observation $\mathbf{y}$. Such a problem can be formalized as

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y}) \tag{C.20}$$

where the *posterior* density $f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y})$ is often unknown. After Bayes' theorem, the MAP estimation problem can be rewritten as

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x}) \tag{C.21}$$

after removing the terms independent of the mute variable $\boldsymbol{x}$. Contrary to the posterior, the *likelihood* density $f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x})$ is assumed to be known most of the time, e.g. with some experimental evidences, and so is the *prior* density $f_{\mathbf{x}}(\boldsymbol{x})$. If the prior is unknown, it is common to assume uniformity in the hidden data, turning the MAP estimation problem into a maximum likelihood (ML) one since the prior density in (C.21) becomes independent of the mute variable $\boldsymbol{x}$:

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}). \tag{C.22}$$

One can further consider that latent (hidden) variables $\mathbf{z}$ are involved in the inference process, extending the MAP problem to

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} \int f_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{z} \tag{C.23}$$

$$= \arg\max_{\boldsymbol{x}} \int f_{\mathbf{y}|\mathbf{x},\mathbf{z}}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}) f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}. \tag{C.24}$$

All of the proposed forms of the MAP estimation problem demand maximization over the set of variables contained in $\boldsymbol{x}$, and sometimes additional computation steps induced by latent variables. When the domain of $\boldsymbol{x}$ (and sometimes $\boldsymbol{z}$) is small, the maximization is acceptable. However, when the domain is high-dimensional and subject to stringent constraints, the MAP estimation problem turns out to be intractable and requires heuristics methods in order to be adressed.

II.2.b Minimum mean squared error inference

Another very common inference problem is the MMSE estimation problem [74]

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} \mathbb{E}\left[\|\mathbf{x} - \boldsymbol{x}\|_2^2 \mid \mathbf{y} = \boldsymbol{y}\right] \tag{C.25}$$

which consists of estimation for the vector $\boldsymbol{x}^\star$ minimizing the expected squared euclidian distortion given the observed data $\boldsymbol{y}$. Provided that

$$\mathbb{E}[\|\mathbf{x}\|_2 \mid \mathbf{y} = \boldsymbol{y}] < \infty \tag{C.26a}$$
$$\mathbb{E}\left[\|\mathbf{x}\|_2^2 \mid \mathbf{y} = \boldsymbol{y}\right] < \infty \tag{C.26b}$$

we can write

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x} - \boldsymbol{x}\|_2^2 \mid \mathbf{y} = \boldsymbol{y}\right] \\
&= \mathbb{E}\left[\|\mathbf{x}\|_2^2 + \|\boldsymbol{x}\|_2^2 - \mathbf{x}^\mathsf{H}\boldsymbol{x} - \boldsymbol{x}^\mathsf{H}\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}\right] && \text{(C.27)} \\
&= \mathbb{E}\left[\|\mathbf{x}\|_2^2 \mid \mathbf{y} = \boldsymbol{y}\right] - \mathbb{E}\left[\mathbf{x}^\mathsf{H} \mid \mathbf{y} = \boldsymbol{y}\right]\boldsymbol{x} - \boldsymbol{x}^\mathsf{H}\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] + \|\boldsymbol{x}\|_2^2 && \text{(C.28)}
\end{aligned}$$

which is minimized by differentiating (in the sense of the Wirtinger derivative, [75]) w.r.t. $\boldsymbol{x}^\mathsf{H}$

$$\partial_{\boldsymbol{x}^\mathsf{H}}\mathbb{E}\left[\|\mathbf{x} - \boldsymbol{x}\|_2^2 \mid \mathbf{y} = \boldsymbol{y}\right] = 0 \Rightarrow -\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] + \boldsymbol{x} = 0 \tag{C.29}$$

and so

$$\boldsymbol{x}^\star = \mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}]\,. \tag{C.30}$$

This last a posteriori expectation $\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}]$ reads

$$\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] = \int \boldsymbol{x} f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y})\mathrm{d}\boldsymbol{x}. \tag{C.31}$$

As for the MAP estimation problem, the posterior density $f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y})$ is generally unknown and replaced using the Bayes theorem by

$$\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] = \frac{\int \boldsymbol{x} f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.}{\int f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.} \tag{C.32}$$

where the likelihood and prior densities are made explicit. Considering latent variables in the model leads to MMSE estimation of the form

$$\mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] = \frac{\iint \boldsymbol{x} f_{\mathbf{y}|\mathbf{x},\mathbf{z}}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}) f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}}{\iint f_{\mathbf{y}|\mathbf{x},\mathbf{z}}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}) f_{\mathbf{x},\mathbf{z}}(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}}. \tag{C.33}$$

II.2.c Variational inference

Some inference problems focus on estimating densities of the variables rather than the variables themselves [76, see section "Variational Methods"]. An example of such an inference problem is

$$g^\star = \arg\min_{g\in\mathcal{D}} \mathbb{KL}\left[g \middle\| f_{\mathbf{x}|\mathbf{y}}\right] \quad \text{s.t.} \quad \mathbf{y} = \boldsymbol{y} \tag{C.34}$$

where one aims at approximating the posterior $f_{\mathbf{x}|\mathbf{y}}$ by a density g taken from a space $\mathcal{D}$, e.g. the exponential family of distributions (VI with exponential family (VIEF)), and given the observed data $\boldsymbol{y}$. A famous example of such variational inference (VI) algorithms is expectation propagation (EP) [77]–[80]. Since

$$\mathbb{KL}\left[g \middle\| f_{\mathbf{x}|\mathbf{y}}\right] = -\mathbb{H}_g[\mathbf{x}] - \mathbb{X}_{g,f_{\mathbf{x}|\mathbf{y}}}[\mathbf{x}] \tag{C.35}$$

using Bayes' theorem on $f_{\mathbf{x}|\mathbf{y}}$, leads to rewrite

$$g^\star = \arg\max_{g\in\mathcal{D}} \mathbb{H}_g[\mathbf{x}] + \mathbb{X}_{g,f_{\mathbf{y}|\mathbf{x}}}[\mathbf{x}] + \mathbb{X}_{g,f_{\mathbf{x}}}[\mathbf{x}] \quad \text{s.t.} \quad \mathbf{y} = \boldsymbol{y} \tag{C.36}$$

where the constant terms, i.e. those that do not depend on g, were removed. This maximization problem thus formulated approximates the posterior density by a density g which must satisfy the constraints contained in the factors $\mathbb{X}_{g,f_{\mathbf{y}|\mathbf{x}}}[\mathbf{x}]$ and $\mathbb{X}_{g,f_{\mathbf{x}}}[\mathbf{x}]$. The constraints being imposed simultaneously on g, the maximization may be too complicated. To relax this constraint, one can consider a maximization of the form

$$(g_\ell^\star, g_p^\star) = \arg\max_{(g_\ell,g_p)\in\mathcal{D}^2} \mathbb{H}_{g_\ell g_p}[\mathbf{x}] + \mathbb{X}_{g_\ell,f_{\mathbf{y}|\mathbf{x}}}[\mathbf{x}] + \mathbb{X}_{g_p,f_{\mathbf{x}}}[\mathbf{x}] \quad \text{s.t.} \quad \mathbf{y} = \boldsymbol{y} \tag{C.37}$$

where the maximization over g has been replaced by a joint maximization over independent densities g_ℓ and g_p for the likelihood and prior densities respectively.

II.3 Approximation by belief propagation

BP is an inference algorithm which was introduced in [81], [82] by Judea Pearl as a technique to propagate belief about data in an acyclic directed graph (e.g. hierarchical networks or trees). It was later considered for more generic undirected graphs that may contain cycles or loops and is

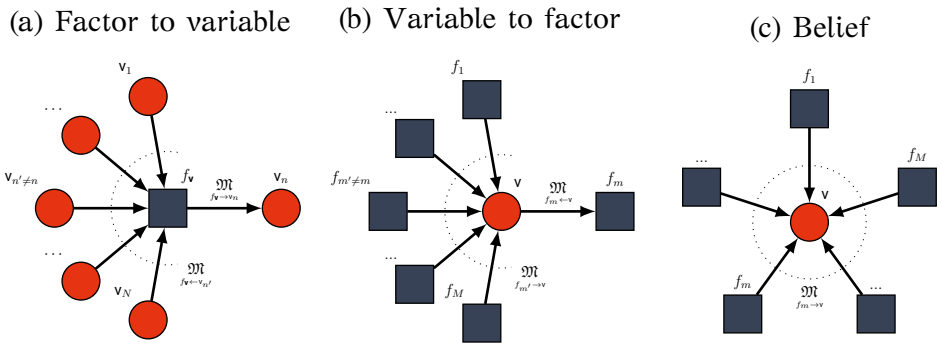


Figure C.3: Belief propagation update rules.

then refered to as loopy BP (LBP). BP has a wide spectrum of applications, especially in the domain of bayesian inference applied to signal processing.

The MAP (Sec. C.II.2.a), MMSE (Sec. C.II.2.b) and VI (Sec. C.II.2.c) estimation problems becomes intractable to solve with more complicated factorization of the posterior and so requires the use of heuristic algorithms to be solved. BP is one them and works as follows. Based on a factorization of the posterior density and the corresponding factor graph, BP iteratively approximates the posterior densities $\left\{f_{\mathsf{x}_n|\mathbf{y}}(x_n \mid \boldsymbol{y})\right\}_n$ of each hidden variable x_n with a local approach unlike the original problems which seek for global solutions.

This process of approximation is described by a message-passing algorithm based on the factor graph. Each node in the factor graph exchanges messages with its neighbors, denoted by $\underset{f\leftrightarrow v}{\mathfrak{M}}$, where the left index denotes a factor, the right index denotes a variable and the arrow indicates the propagation direction of the message ($\rightarrow$ from factor to variable and $\leftarrow$ from variable to factor). The message rules described below apply to the MAP, MMSE and VIEF problems.

Message from a factor node to a variable node The message sent by a factor node f to a variable node n is a density which consists in the following steps:

1. Aggregation of the messages coming from the neighbor variable nodes $\{\mathsf{v}_n\}_n$.

$$\prod_{n'\in[N]} \underset{f_{\mathbf{v}}\leftarrow\mathsf{v}_{n'}}{\mathfrak{M}}(v_{n'}) \tag{C.38}$$

2. Product with the local factor.

$$f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N]} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \tag{C.39}$$

3. Marginalization of this product w.r.t. to all the variables except v_n (and projection for VIEF).

$$\max_{\boldsymbol{v}_{\setminus n}} f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N]} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \qquad \text{(MAP)} \tag{C.40a}$$

$$\int f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N]} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \mathrm{d}v_{n'} \qquad \text{(MMSE)} \tag{C.40b}$$

$$\mathrm{Proj}\left(\int f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N]} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \mathrm{d}v_{n'} \right) \qquad \text{(VIEF)} \tag{C.40c}$$

where $\mathrm{Proj}(f) = \arg\min_{g \in \mathcal{F}} \mathbb{KL}[g \| f]$.

4. Remove contribution of the message coming from the destination variable node v_n.

$$\underset{f_{\mathbf{v}} \to \mathrm{v}_n}{\mathfrak{M}} (v_n) \propto \begin{cases} \max_{\boldsymbol{v}_{\setminus n}} f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N] \setminus \{n\}} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) & \text{(MAP)} \\ \int f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N] \setminus \{n\}} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \mathrm{d}v_{n'} & \text{(MMSE)} \\ \dfrac{\mathrm{Proj}\left(\int f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n' \in [N]} \underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_{n'}}{\mathfrak{M}} (v_{n'}) \mathrm{d}v_{n'} \right)}{\underset{f_{\mathbf{v}} \leftarrow \mathrm{v}_n}{\mathfrak{M}} (v_n)} & \text{(VIEF)} \end{cases} \tag{C.41}$$

Note that for VIEF, the denominator is not included into the projection operation.

Message from a variable node to a factor node A message sent by a variable to a factor node simply consists in the aggregation of the incoming messages except the one coming from the target factor node as depicted in Fig. C.3b. It reads

$$\underset{f_m \leftarrow \mathrm{v}}{\mathfrak{M}} (v) \propto \prod_{m' \in [M] \setminus \{m\}} \underset{f_{m'} \to \mathrm{v}}{\mathfrak{M}} (v) \tag{C.42}$$

Algorithm C.1 Belief propagation

init:

  Messages $\left\{ \underset{f \leftarrow \mathsf{v}}{\mathfrak{M}} \right\}$ for all factor and variable nodes.

end

repeat:

  Select, e.g. at random, a variable node $\mathsf{v} \in \mathfrak{F}$.

  Compute the messages $\left\{ \underset{f' \to \mathsf{v}}{\mathfrak{M}} \right\}_{f' \in \mathfrak{N}(\mathsf{v})}$ based on Eq. C.41.

  Compute the belief $\underset{\mathsf{v}}{\mathfrak{B}}$ at variable node v based on Eq. C.43.

  Send for every neighboring factor $f \in \mathfrak{N}(\mathsf{v})$ the message $\underset{f \leftarrow \mathsf{v}}{\mathfrak{M}}$ based on Eq. C.42.

until: convergence of the messages

Estimate independently each variable $\mathsf{v} \in \mathfrak{F}$ with their belief $\underset{\mathsf{v}}{\mathfrak{B}}$ if MAP or MMSE is considered.

Belief of a variable node The belief of a variable node consists of the same of the message sent by the variable node to any factor node except that all the impinging messages are considered.

$$\underset{\mathsf{v}}{\mathfrak{B}}(v) \propto \prod_{m \in [M]} \underset{f_m \to \mathsf{v}}{\mathfrak{M}}(v). \tag{C.43}$$

Also, the belief is central in BP since it represents the approximate density of the variable's posterior density. In the case of MAP and MMSE estimations, the variable v_n is finally estimated with

$$\hat{v}_n = \arg\max_{v} \underset{\mathsf{v}_n}{\mathfrak{B}}(v) \qquad \text{(MAP)}, \tag{C.44}$$

$$\hat{v}_n = \mathbb{E}\left[\mathsf{v}; \underset{\mathsf{v}_n}{\mathfrak{B}} \right] \qquad \text{(MMSE)}. \tag{C.45}$$

Complete algorithm The complete BP algorithm is described in Algo. C.1. Notice that this is a very general suggestion of how BP can be applied and that more optimized strategies in terms of message exchanging could be considered, based on the structure of the factor graph. For instance in Fig. C.2, instead of chosing randomly the variable nodes from which to send messages, one can consider sending the messages starting from

Algorithm C.2 AMP

input: $\boldsymbol{y}, \boldsymbol{A}, \tau_w, I_{\max}$ and parameters of $f_{\mathsf{x}}(\boldsymbol{x})$

init:

$i = 0$

$\boldsymbol{\tau}_w = \tau_w \mathbf{1}_{M\times 1}$

$\tau_{\hat{\boldsymbol{x}},i}$ and $\hat{\boldsymbol{x}}_i$ (e.g. $\mathbb{V}[\mathsf{x}; f_{\mathsf{x}}(\boldsymbol{x})]$ and $\mathbb{E}[\mathsf{x}; f_{\mathsf{x}}(\boldsymbol{x})]$)

end

for $i \in [I_{\max}]$ **do**:

$\tau_{\hat{\boldsymbol{p}},i} = \boldsymbol{A}^{\odot 2}\tau_{\hat{\boldsymbol{x}},i-1}, \quad \hat{\boldsymbol{p}}_i = \boldsymbol{A}\hat{\boldsymbol{x}}_{i-1} - \tau_{\hat{\boldsymbol{p}},i} \odot (\boldsymbol{y} - \hat{\boldsymbol{p}}_{i-1}) \oslash (\boldsymbol{\tau}_w + \tau_{\hat{\boldsymbol{p}},i-1})$

$\tau_{\hat{\boldsymbol{r}},i} = (\boldsymbol{A}^{\odot 2}(\boldsymbol{\tau}_w + \tau_{\hat{\boldsymbol{p}},i})^{\odot -1})^{\odot -1}, \quad \hat{\boldsymbol{r}}_i = \hat{\boldsymbol{x}}_{i-1} + \tau_{\hat{\boldsymbol{r}},i} \odot \left(\boldsymbol{A}^{\mathsf{H}}(\boldsymbol{y} - \hat{\boldsymbol{p}}_i) \oslash (\boldsymbol{\tau}_w + \tau_{\hat{\boldsymbol{p}},i})\right)$

$\tau_{\hat{\boldsymbol{x}},i} = \mathbb{V}[\mathsf{x}; f_{\mathsf{x}}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_i, \mathrm{diag}(\tau_{\hat{\boldsymbol{r}},i}))], \quad \hat{\boldsymbol{x}}_i = \mathbb{E}[\mathsf{x}; f_{\mathsf{x}}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_i, \mathrm{diag}(\tau_{\hat{\boldsymbol{r}},i}))]$

end

return:

The last estimate $\hat{\boldsymbol{x}}_{I_{\max}}$

end

the factor nodes $\{f_{\mathsf{a}_n}\}_n$, following the path $f_{\mathsf{a}_n} \to \mathsf{a}_n \to f_{\mathsf{b}_n|\mathsf{a}_n} \to \mathsf{b}_n \to f_{\mathsf{c}_m|\mathsf{b}_n} \to \mathsf{c}_m$ before going the other way around.

The convergence of BP has been proven when the factor graph exhibits a tree-like structure (see e.g. [83]). However, in the case of LBP, there is no convergence guarantees in the general case although some few works were done towards this direction [84], [85]. In the special case of VIEF and EP with gaussian-based messages [77], numerical evidences of convergence were found.

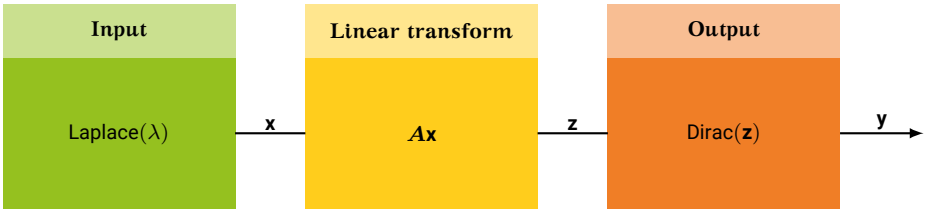
Remarks As stated in this subsection, the BP performs local approximations of the posterior densities of all the variables. This is typically visible from (C.43) and (C.41) where the belief aggregates the messages from the factor nodes which are local solutions of one of the original problems.

III Approximate Message Passing Algorithms

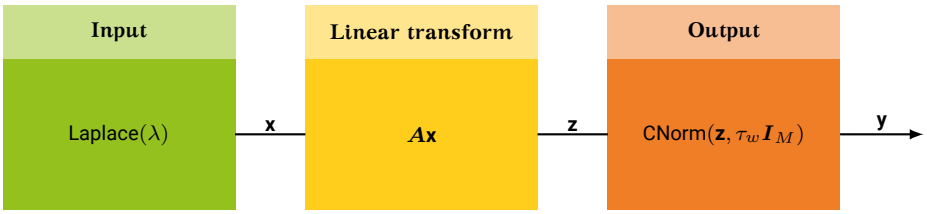
III.1 AMP

With the advent of CS, Donoho, Maleki and Montanari brought a new algorithm named AMP. Introduced in their seminal papers [53], [86], [87]

(a) BPDN



(b) LASSO



(c) Bayes optimal

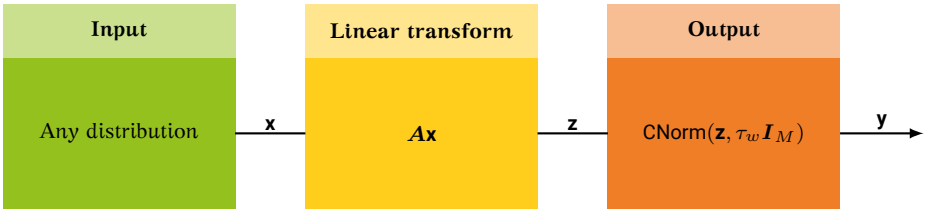


Figure C.4: Channel models for BPDN-AMP, LASSO-AMP and Bayes optimal-AMP.

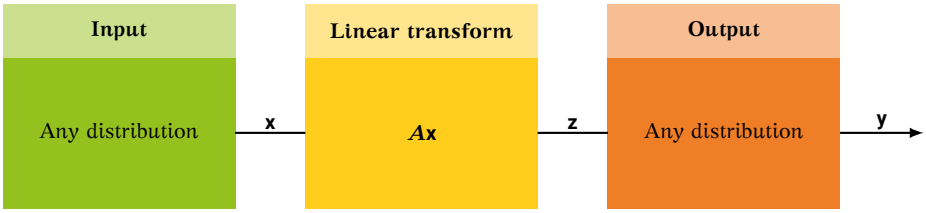


Figure C.5: Channel model for GAMP.

and in the thesis of Maleki [88], AMP was shown to be a promising algorithmic framework for studying CS problems through the eyes of bayesian inference. By studying the relaxed versions of the original CS problem that are BPDN and LASSO Sec. C.I.1, they derived a LBP, aka message-passing, algorithm to address the estimation of a sparse system input signal given the system output and its measurement matrix.

More precisely, they studied the problems under the MAP formulations in Eqs. (C.13) and (C.14). In [53], Donoho et al even suggested a third model, sometimes refered as *Bayes optimal* [89], that takes into account general prior knowledge on the signal input:

$$\hat{\boldsymbol{x}}_{\text{Bayes}} = \underset{\boldsymbol{x}\in\mathbb{C}^N}{\arg\max} \prod_{n\in[N]} f_{\mathsf{x}_n}(x_n) \prod_{m\in[M]} \frac{1}{\pi\tau_w} e^{-\frac{|y_m-[\boldsymbol{A}\boldsymbol{x}]_m|^2}{\tau_w}}. \tag{C.46}$$

These three new optimization problems are reminiscent of MAP estimators with Laplace (BPDN and LASSO) or general prior (Bayes optimal) distribution for the input signal and with Dirac (for BPDN) or Gaussian (LASSO and Bayes optimal) output channel. Their schematic representations are shown in Fig. C.4. Also, the objective functions exhibit a factorized structure allowing to draw for each problem a factor graph and derive a tailored LBP algorithm.

However, the loopy structures of the corresponding BP algorithms produce message probability density functions (pdfs) that are still too difficult to compute so that an approximation is needed. To do so, it is assumed that:

1. subGaussian measurement matrices are used, i.e that the entries of $\boldsymbol{A}$ are assumed to be i.i.d. random variables satisfying

$$\exists(C,\kappa)\in\mathbb{R}_{+,*},\ \forall(m,n)\in[M]\times[N],\ 1-F_{\mathsf{a}_{mn}}(a)\leq Ce^{-\kappa|a|^2}; \tag{C.47}$$

2. the input signal dimension N tends to infinity and the output signal dimension M linearly scales with N, such that $M(N)/N \xrightarrow{N\to+\infty} \eta$ where η is finite and constant.

Based on these assumptions, they show that, using a variant of the Berry-Esséen theorem, some of the BP messages can be approximated by Gaussian pdfs.

Propagating the approximation in the other messages finally leads to three versions of AMP, one for BPDN, one for LASSO and one for Bayes

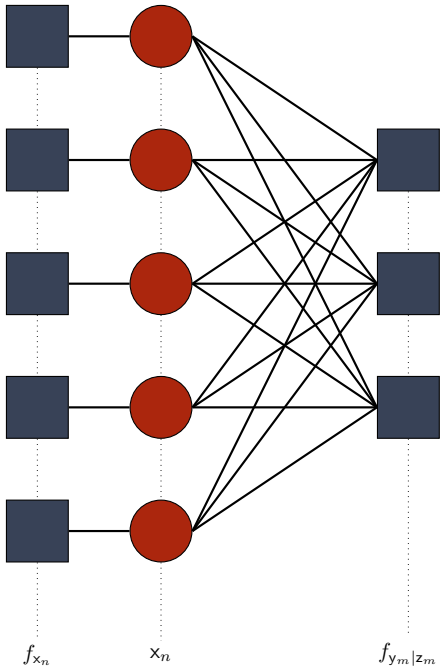


Figure C.6: Underlying factor graph of GAMP.

optimal. This last version of AMP generalizes both BPDN-AMP and LASSO-AMP since they are particular cases with respectively Dirac and Laplace priors on the input signal. It is given in Algo. C.2. Since AMP is a particular case of GAMP, presented in Sec. C.III.2, the derivation of the updates is not given and could be infered from that of Algo. C.3. A similar derivation can also be found in [90].

III.2 GAMP

A generalization of AMP is GAMP [91]. The main difference lies in the fact that AMP was originally derived in the case of a (complex) Gaussian output channel only whereas GAMP is more general by allowing any type of output channel, both for an MMSE and MAP-based estimation. In fact, it goes even further by approximating the distribution of the system input $\mathbf{x}$ by a Gaussian random vector where the mean and variance are derived in the rest of this subsection.

The first proof of GAMP equations is given in [91] with a focus on the real ($\mathbb{R}$) case although it was said that the result also applies to the complex ($\mathbb{C}$) case. This proof relies on Taylor expansions associated to approximation by removing terms of higher order.

Algorithm C.3 GAMP

input: $\boldsymbol{y}, \boldsymbol{A}, I_{\max}$ and parameters of $f_{\mathsf{x}}(\boldsymbol{x}), f_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y} \mid \boldsymbol{z})$

init:

$i = 0$

$\hat{\boldsymbol{v}}_i = \mathbf{0}_{N\times 1}$

$\hat{\boldsymbol{\tau}}_{\mathsf{x},i}$ and $\hat{\boldsymbol{x}}_i$ (e.g. $\mathbb{V}[\mathbf{x}; f_{\mathsf{x}}(\boldsymbol{x})]$ and $\mathbb{E}[\mathbf{x}; f_{\mathsf{x}}(\boldsymbol{x})]$)

end

for $i \in [I_{\max}]$ **do**:

$\hat{\boldsymbol{\tau}}_{\mathsf{p},i} = \boldsymbol{A}^{\odot 2}\hat{\boldsymbol{\tau}}_{\mathsf{x},i-1}, \quad \hat{\boldsymbol{p}}_i = \boldsymbol{A}\hat{\boldsymbol{x}}_{i-1} - \hat{\boldsymbol{\tau}}_{\mathsf{p},i} \odot \hat{\boldsymbol{v}}_{i-1}$

$\hat{\boldsymbol{\tau}}_{\mathsf{z},i} = \mathbb{V}\left[\mathbf{z}; f_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y} \mid \boldsymbol{z})\mathcal{CN}(\boldsymbol{z}; \hat{\boldsymbol{p}}_i, \mathrm{diag}(\hat{\boldsymbol{\tau}}_{\mathsf{p},i}))\right], \quad \hat{\boldsymbol{z}}_i = \mathbb{E}\left[\mathbf{z}; f_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y} \mid \boldsymbol{z})\mathcal{CN}(\boldsymbol{z}; \hat{\boldsymbol{p}}_i, \mathrm{diag}(\hat{\boldsymbol{\tau}}_{\mathsf{p},i}))\right]$

$\hat{\boldsymbol{\tau}}_{\mathsf{v},i} = (\hat{\boldsymbol{\tau}}_{\mathsf{p},i} - \hat{\boldsymbol{\tau}}_{\mathsf{z},i}) \oslash (\hat{\boldsymbol{\tau}}_{\mathsf{p},i})^{\odot 2}, \quad \hat{\boldsymbol{v}}_i = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{p}}_i) \oslash \hat{\boldsymbol{\tau}}_{\mathsf{p},i}$

$\hat{\boldsymbol{\tau}}_{\mathsf{r},i} = \mathbf{1}_{N\times 1} \oslash (\boldsymbol{A}^{\odot 2}\hat{\boldsymbol{\tau}}_{\mathsf{v},i}), \quad \hat{\boldsymbol{r}}_i = \hat{\boldsymbol{x}}_{i-1} + \hat{\boldsymbol{\tau}}_{\mathsf{r},i} \odot (\boldsymbol{A}^{\mathsf{H}}\hat{\boldsymbol{v}}_i)$

$\hat{\boldsymbol{\tau}}_{\mathsf{x},i} = \mathbb{V}[\mathbf{x}; f_{\mathsf{x}}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_i, \mathrm{diag}(\hat{\boldsymbol{\tau}}_{\mathsf{r},i}))], \quad \hat{\boldsymbol{x}}_i = \mathbb{E}[\mathbf{x}; f_{\mathsf{x}}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_i, \mathrm{diag}(\hat{\boldsymbol{\tau}}_{\mathsf{r},i}))]$

end

return:

The last estimate $\hat{\boldsymbol{x}}_{I_{\max}}$

end

Another proof of GAMP is given in [92] where the focus was given to the complex case for MMSE estimation. The proof is also "simpler" in that it does not require the use of Taylor expansions, that would be painful to write for the complex case with the use of Wirtinger derivatives. In fact, it uses the framework of VI presented in Sec. C.II.2.c with the particular case of VIEF with Gaussian distribution, which is the framework of EP.

In order to provide a self-contained document to the readers, a more detailed proof following [92] is written in Sec. G.III, without claiming any novelty. The final algorithm is given in Algo. C.3 and is described by the factor graph Fig. C.6.

The main outcomes of GAMP are the following. First, the *a posteriori* distribution of the system input is approximated by

$$\forall n \in [N], \quad f_{\mathsf{x}_n|\mathsf{y}}(x_n \mid \boldsymbol{y}) \approx f_{\mathsf{x}_n}(x_n)\mathcal{CN}(x_n; \hat{r}_n^i, \hat{\tau}_{\mathsf{r},n}^i) \tag{C.48}$$

where the variables $\{(\hat{r}_n^i, \hat{\tau}_{\mathsf{r},n}^i)\}_{n\in[N]}$ corresponds to some of the GAMP variables introduced in Sec. G.III. Based on this approximation, it is

possible to compute analytically separate estimate of each variable (based on a MAP or MMSE estimator), provided that the priors pdfs $\{f_{\mathsf{x}_n}\}_{n\in[N]}$ are not too complicated.

Second, the initial inference problem which was intractable, has been turned into a suboptimal yet simple algorithm, with tractable updates of the intermediate variables ($\{(\hat{r}_n^i, \hat{\tau}_{\mathsf{r},n}^i)\}_{n\in[N]}$, $\{(\hat{p}_m^i, \hat{\tau}_{\mathsf{p},m}^i)\}_{m\in[M]}$ and $\{(\hat{v}_m^i, \hat{\tau}_{\mathsf{v},m}^i)\}_{m\in[M]}$). When the priors $\{f_{\mathsf{x}_n}\}_{n\in[N]}$ and posteriors $\{f_{\mathsf{y}_m|\mathsf{z}_m}\}_{m\in[M]}$ are tractable, the principal variables updates ($\{(\hat{x}_n^i, \hat{\tau}_{\mathsf{x},n}^i)\}_{n\in[N]}$ and $\{(\hat{z}_m^i, \hat{\tau}_{\mathsf{z},m}^i)\}_{m\in[M]}$) can be performed analytically.

In [93], GAMP is enhanced with damped updates of the variables in order to improve its convergence and robustness against ill-conditionned measurement matrices $\boldsymbol{A}$ [94]. It consists in applying the following function component-wise to each GAMP variables (means and variances)

$$\text{damp}(\chi_{\text{old}}, \chi_{\text{new}}; \theta) = (1-\theta)\chi_{\text{old}} + \theta\chi_{\text{new}} \tag{C.49}$$

where $\theta \in [0, 1]$ is the damping weight (0 no update and 1 no memory) and χ_{old} (resp. χ_{new}) denotes the value of a variable χ before (resp. after) the update. For improved readibility of the algorithms, damped updates will not be explicitly written although they are used. Based on our own extensive numerical study, a value of $\theta = 0.9$ gives stable convergence of GAMP with normalized centered Gaussian mesurement matrix.

III.3 Hybrid GAMP

Both AMP and GAMP are very powerful algorithms to address CS problems as stated in Sec. C.I. However, more complicated structures may arise in the system model such as

- more variables, different from **y** and **x**, may be required to best describe the system;
- the factors depending on the system variables do not exhibit a nice separable structure;
- the factors may be at the same time strongly dependent on a subset of the system variables and weekly dependent on the reminder of the variables.

In this situation, it becomes difficult to derive a GAMP instance that would fit to the whole system with such structures. However, if the

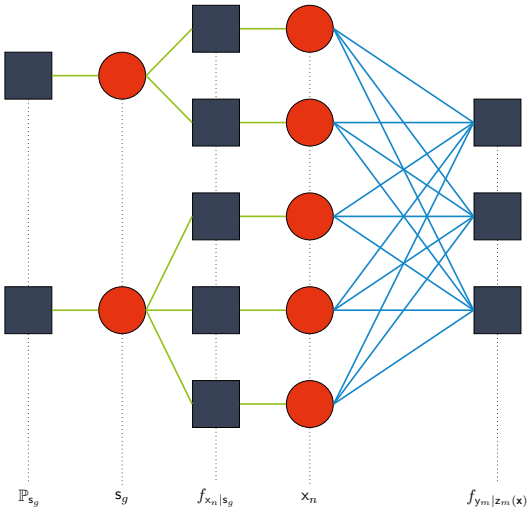


Figure C.7: Underlying factor graph of GS-HGAMP. Green edges indicate strong edges and blue edges indicate weak edges.

system can be decomposed into multiple parts, one that fits GAMP and the others with the difficult structures, it would be interesting to see if one can still leverage GAMP benefits to simplify the estimation process.

This problem was studied in [95] where the authors considered the CS problems where the factorization of the joint system variables $\mathbf{v}$ could be written as

$$f_{\mathbf{v}}(\boldsymbol{v}) = \prod_a f_a(\boldsymbol{v}) \prod_b f_b(\boldsymbol{v}) \tag{C.50}$$

where the factors $\{f_a\}_a$ are the factors that equally and linearly depend on subsets of the system variables while the factors $\{f_a\}_a$ strongly depend on a subset of the variables contained in $\mathbf{v}$. In other words, in the factor graph associated with Eq. C.50, the factor nodes $\{f_a\}_a$ are densely connected to the variables nodes it depends with *weak* edges and the factor nodes $\{f_b\}_b$ only share *strong edges* with few variable nodes.

As mentioned in [95], this situation typically arises in the context of CS with group-sparsity where the components x_n of the the input signal $\mathbf{x}$ depends on a binary random pattern $\mathbf{s}$, so that the factorization of the joint pdf of the system variables reads

$$f_{\mathbf{y},\mathbf{x},\mathbf{s}}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}) = \prod_{m\in[M]} f_{\mathsf{y}_m|\mathbf{x}}(y_m \mid \boldsymbol{x}) \prod_{n\in[N]} f_{\mathsf{x}_n|\mathbf{s}}(\boldsymbol{x} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) \tag{C.51}$$

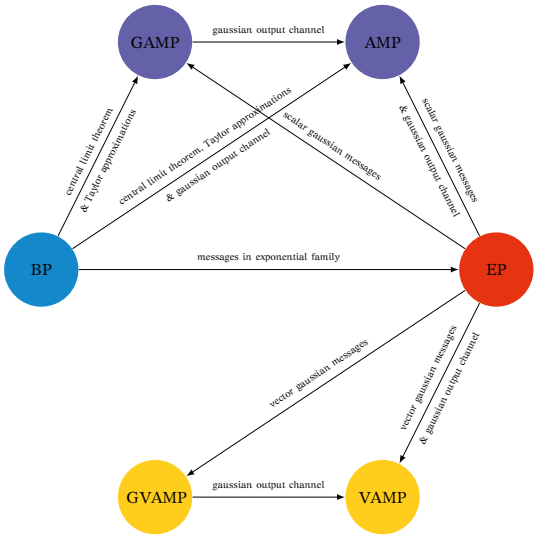


Figure C.8: Some relationships between AMP-based algorithms.

The corresponding factor graph is depicted in Fig. C.7 where one can see that the factors f_{s_g} and $f_{\mathsf{x}_n|\mathsf{s}_g}$ strongly depend on the variables s_g. For such a problem, it would still be possible to use GAMP for the dense part of the factor that involves weak edges but the strong edges and the messages they convey will be treated with BP. The rationale of using GAMP for weak edges only lies in the fact that, as seen in Sec. C.III.2, GAMP relies on Gaussian approximations and the central limit theorem which are valid approximations when the factors involve linear combination of the system variables where the combination weights are of the same order.

A natural question that arises is the connection between GAMP and BP messages. From Sec. C.II, BP messages are pdfs whereas GAMP messages are means and variances. Passing from GAMP to BP messages is rather straightforward since one can replace the messsage means and variances by the corresponding Gaussian pdf. Conversely, passing from BP to GAMP messages is ensured through the computation of the means and variances w.r.t. $f^{\text{in}}_{\mathsf{x}_n}$ and $f^{\text{out}}_{\mathsf{x}_n}$ since these two pdfs will depend on the incoming BP messages. For mathematical details about such a connection, see [95, Sec. IV].

III.4 Extensions

In order to provide the reader with extensions of AMP, we shortly review some of its variants.

AMP and GAMP depend on the knowledge of hyperparameters to

compute their message means and variances, particularly the mean and variances based on $f^{\text{in}}_{\mathsf{x}_n}$ and $f^{\text{out}}_{\mathsf{x}_n}$. A well known framework for iterative parameter estimation is expectation maximization (EM). It has been successfully adapted to multiple instances of GAMP, named EM-GAMP to estimate the noise variance, sparsity rates of the input signal or Gaussian mixture weights [96]–[99].

Another version of AMP derived from EP with vector Gaussian messages was also considered. This vector AMP (VAMP) algorithm [100] has been generalized as generalized VAMP (GVAMP) in [101] based on the expectation consistent algorithm (similar to EP) [102].

Both GAMP and GVAMP were extended to the bilinear estimation case [52], [103]–[108], i.e. considering that the measurement matrix $\mathbf{A}$ is assumed to be random and must be estimated simultaneously with the system random input $\mathbf{x}$ based on the system output $\mathbf{y} = \boldsymbol{y}$.

GAMP can also be applied to multi-layered, or nested, systems [109], [110] that consist of a sequence of subsytems of the form Eq. C.5 where the output of system ℓ is the input of system $\ell + 1$.

Finally, work has been done to optimize some the operations of AMP and VAMP using deep networks, e.g. [111], [112].

D

HYBRID GENERALIZED APPROXIMATE MESSAGE PASSING FOR AUDACE WITH GROUP-HOMOGENEOUS ACTIVITY

I INTRODUCTION

To improve efficiency and minimize economic losses, a key challenge in large-scale industrial plants is rapid identification of faulty or degraded machines. An increasingly popular approach is to place sensors with wireless communication capability on as many machines as is economically feasible [113]. When a centralized access point sends a sync signal, all sensors which detect faulty behavior of their machine then transmit data to the access point.

An important challenge lies in the fact that each sensor only transmits when it detects a fault. For applications with a relatively small number of sensors, it would be reasonable to allocate distinct subcarriers to each sensor, making the identification trivial since detecting a signal in a subcarrier would indicate which sensor is active. However, it will not be the case when the—potentially very large—number of sensors that transmit within a given frame can vary dramatically. It is therefore highly

inefficient to allocate orthogonal resources for each sensor's transmission especially when sensors are inactive with a high probability. Doing so would lead to high resource requirements, most of which is not utilized. As a consequence, this scenario—known as random access—requires non-orthogonal resource allocation (e.g., the same time-slots or subcarriers are utilized by multiple sensors), often known as non-orthogonal multiple access (NOMA) [25].

Another, less often accounted for, feature is that sensors on the same or nearby machines often observe physical variables, such as the temperature, that are similar. In particular, the observations by such sensors are statistically correlated [56].

In order to transmit data, each sensor first transmits a pilot sequence which provides a means of identifying active sensors (and hence faulty machines) and estimating channels between the sensors and the access point. There is presently a large body of work on sensor identification and channel estimation in random access for *uncorrelated* sensors. For example, a Bayesian estimation framework has been developed for the user identification problem in [114], which attempts to identify the active subset of sensors based on the posterior distribution informed by observations at the access point.

Due to the nature of random access, active sensors typically form a sparse subset of all sensors. As a consequence, the problem of channel estimation has recently been attacked using compressed sensing, with algorithms based on LASSO and OMP [115]–[117].

The problem of *joint* sensor identification and channel estimation in random access based on NOMA has recently seen significant attention. In [67], channel estimation is performed via a low complexity, but accurate, variant of BP known as GAMP, with sensor identification obtained through an *ad hoc* thresholding scheme. In [118], a GAMP-based sensor identification and channel estimation algorithm has been proposed for multi carrier communication systems, with activity probabilities estimated via expectation-maximization. In [119], expectation propagation is used to address the joint active user detection and channel estimation problem in a multi-user setup with inter-symbol interference introduced by a faster-than-Nyquist signaling. While the correlation between the symbols is considered, the activity of the users is assumed to be statistically independent. In [54], [99], [120], a systematic approach for the joint

identification and channel estimation problem exploited a general framework, known as the group-sparse model, where sensor activity—common to all subcarriers—is treated as a latent variable. A GAMP-type method, known as HGAMP, was then applied by exploiting the GS-HGAMP algorithm in [95] tailored for the group-sparse model.

This existing work on joint identification and channel estimation has largely focused on generic random access systems. For sensor networks tailored to industrial fault detection, the probability a fault occurs is also often dependent on physical variables, such as the temperature, for which a noisy estimate may also be obtained at the access point. Indeed, when the temperature deviates from standard operating levels, the probability of a machine faults or degradation can increase. Such a scenario arises in the context of semiconductor manufacturing [121]. As the access point can locally measure the ambient temperature, this provides additional information, which can potentially improve the performance of algorithms for sensor identification and channel estimation.

In this chapter, we develop algorithms for sensor identification and channel estimation in narrowband communication systems in presence of fault probability, which depend on physical variables (such as the temperature, which we will focus on in the remainder of the chapter) and may be statistically correlated. The first step is to introduce a statistical model relating observations at the access point (i.e., the ambient temperature and received signal) to the channel, activity of each sensor, and the probability each machine is faulty.

Based on our new model, we derive an identification and channel estimation algorithm by exploiting GAMP. In particular, the model falls into the framework of HGAMP [95]. The algorithm is obtained by developing a LBP algorithm for the model, and then applying GAMP for the variables associated with the communication channel.

A key feature of the algorithm is that it explicitly accounts for uncertainty and correlation in the probability sensors are active, as opposed to existing approaches where the activity probability is fixed and sensor transmissions are uncorrelated. In addition, our model accounts for the impact of physical variables (such as temperature) on the probability of a fault. We model the probability of a fault conditioned on temperature observations at the access point via the beta distribution, a highly flexible family of models. As such, we call the algorithm GHomA-HGAMP.

Numerical results demonstrate that GHomA-HGAMP outperforms existing algorithms based on GAMP [91] and GS-HGAMP [95], [120]. In particular, GHomA-HGAMP outperforms these approaches by up to 5dB in terms of the NMSE for channel estimation and for sensor identification, the UER is approximately four times lower when the expected activity probability is 0.35. Finally, unlike existing GAMP and GS-HGAMP algorithms, GHomA-HGAMP yields a posteriori estimates of activity probabilities. These estimates provide insight into the physical variables (e.g., temperature) at each machine that may be useful to detect degradation without including measurements within the data transmission.

I.1 Main contributions

The main contributions in this work are summarized as follows:

1. We develop a new framework for wireless sensor networks for fault detection, which incorporates knowledge of physical variables (e.g., temperature).
2. We introduce a new statistical model for device activity that allows for correlated sensors, which generalizes the group-sparse model and allows for features such as error-prone sensors.
3. We develop an algorithm within the HGAMP framework to estimate the channel and identify active sensors.
4. We show via Monte Carlo simulations that significant performance improvements can be obtained over existing algorithms in terms of NMSE and UER.

I.2 Organization

This chapter is organized as follows. In Sec. D.II, we introduce the system model. We then derive in Sec. D.III new loopy belief propagation and HGAMP algorithms tailored to our model. We assess and discuss the performance of the new GHomA-HGAMP in Sec. D.IV. Finally, we conclude and provide insights for future work in Sec. D.V.

II System Model and Problem Formulation

We consider a GFRA scenario based on Sec. B.II for a cellular wireless network with N single-antennas UEs and 1 K-antennas AP.

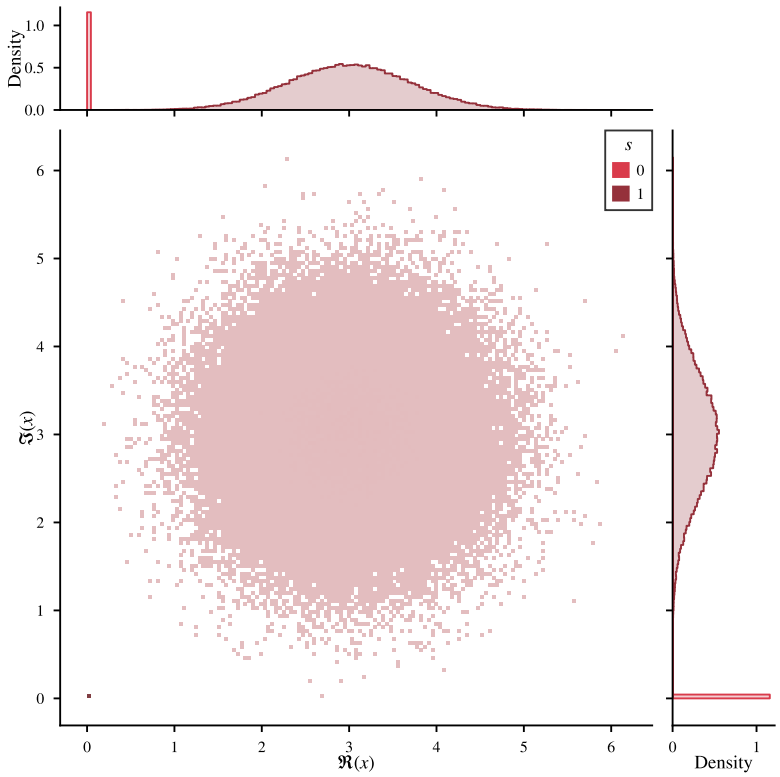


Figure D.1: Histogram of a Bernoulli-complex gaussian random variable.

II.1 Received signal

Given the UE n is active, we assume that the fading coefficients h_{nk} between this UE and the k-th access point's antenna is Gaussian distributed with mean μ_{h} and variance τ_{h}. When a UE is inactive, the fading coefficient is set to zero with probability one. The random signal received by the AP reads

$$\mathbf{Y} = \boldsymbol{P}\mathbf{H} + \mathbf{W} = \mathbf{Z} + \mathbf{W} \tag{D.1}$$

where we remember from Eq. B.10 that $\mathbf{Z} = \mathbf{XH}$ in the particular case $\mathbf{X} = \boldsymbol{P}$. The channel random coefficients are conditionnaly identically and independently distributed (i.i.d.) given

$$\forall (n,k) \in [N] \times [K],\ \mathsf{h}_{nk} \mid \mathsf{s}_n = s \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } s = 0 \\ \mathsf{CNorm}(\mu_{\mathsf{h}}, \tau_{\mathsf{h}}) & \text{if } s = 1 \end{cases} \tag{D.2}$$

and the noise is additive white gaussian with i.i.d. coefficients

$$\forall (m,k) \in [M] \times [K],\ \mathsf{w}_{mk} \sim \mathsf{CNorm}(\mu_{\mathsf{w}}, \tau_{\mathsf{w}})\,. \tag{D.3}$$

The corresponding pdfs admit the factorizations

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) = \prod_{n=1}^{N}\prod_{k=1}^{K} f_{\mathsf{h}_{nk}|\mathsf{s}_n}(h_{nk} \mid s_n) \tag{D.4}$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} \delta(h_{nk})^{1-s_n}\mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})^{s_n} \tag{D.5}$$

and

$$f_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{Y} \mid \boldsymbol{Z}) = \prod_{m=1}^{M}\prod_{k=1}^{K} f_{\mathsf{y}_{mk}|\mathsf{z}_{mk}}(y_{mk} \mid z_{mk}) \tag{D.6}$$

$$= \prod_{m=1}^{M}\prod_{k=1}^{K} \mathcal{CN}(y_{mk}; z_{mk} + \mu_{\mathsf{w}}, \tau_{\mathsf{w}}). \tag{D.7}$$

II.2 Group homogeneous activity pattern

II.2.a Model

As discussed in Sec. B.II.3.a, there is a need for modeling correlated activity patterns which do not consists in (independent) group sparsity nor multivariate bernoulli distributed activity pattern. We thus introduce a new activity pattern model, named group homogeneous activity (GHomA) pattern that is now described.

As for the independent GS model, the GHomA model assumes the UEs spread over $G \in \mathbb{N}_*$ independent groups where group $g \in [G]$ contains the UEs indexed by $\mathfrak{G}_g \subset [G]$. For a group g, the activity probability of the state of the UE $n \in \mathfrak{G}_g$ is assumed to be Bernoulli distributed as

$$\mathsf{s}_n \mid \mathsf{q}_g = q_g \sim \mathsf{Bern}(q_g) \tag{D.8}$$

where q_g is a common random variable describing the activity probability of the UEs. We then have the Markov chains

$$\forall g \in [G],\ \mathsf{q}_g \to \mathbf{s}_g. \tag{D.9}$$

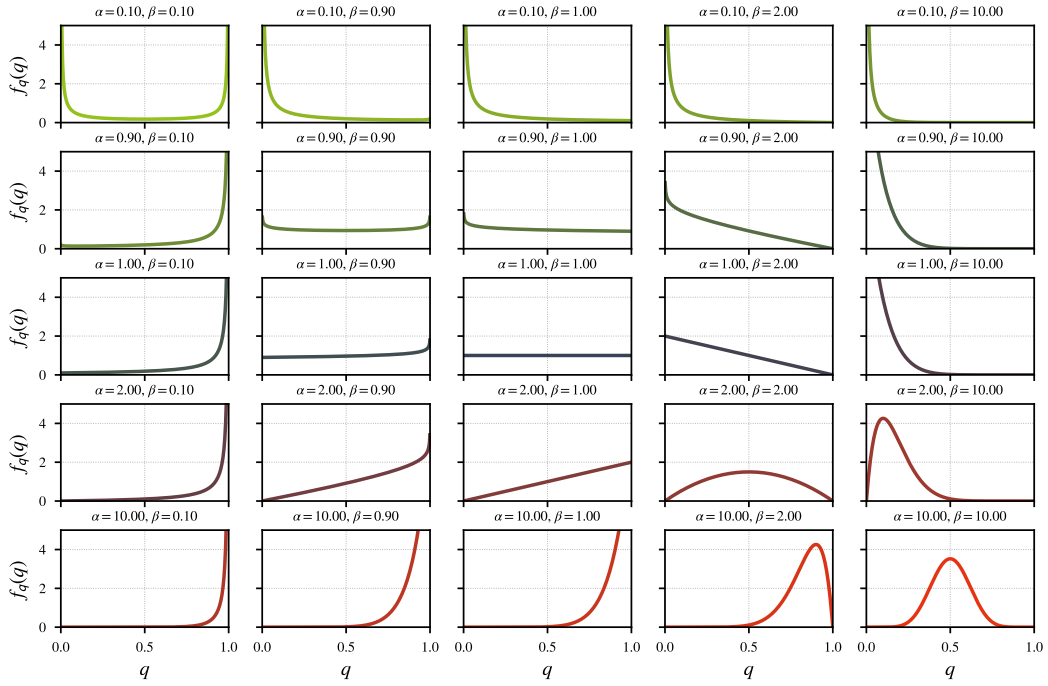


Figure D.2: Examples of probability density functions for the Beta distribution.

We also assume that the entries of the random vector of the activity probabilities $\mathbf{q} = [\mathsf{q}_g]_{g\in[G]}^{\mathsf{T}}$ are mutually independent, and so are the groups, and beta distributed as

$$\forall g \in [G],\ \mathsf{q}_g \sim \mathsf{Beta}(\alpha_g, \beta_g) \qquad \text{with } (\alpha_g, \beta_g) \in \mathbb{R}^2_{+,*}. \tag{D.10}$$

The general form of the beta pdf is

$$f_{\mathsf{q}_g}(q) = \mathcal{B}(q; \alpha_g, \beta_g) = \frac{q^{\alpha_g - 1}(1-q)^{\beta_g - 1}}{\mathrm{B}(\alpha_g, \beta_g)} \tag{D.11}$$

and has the following mean and variance

$$\mathbb{E}[\mathsf{q}_g] = \frac{\alpha}{\alpha + \beta}, \tag{D.12}$$

$$\mathbb{V}[\mathsf{q}_g] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{D.13}$$

The choice of the family of beta distributions is motivated by the following remarks.

- It is a continuous distribution.
- The support is $[0, 1]$, which is ideal for modeling an activity probability.
- With only two parameters, it is possible to describe many types of activity profiles with of some them depicted in Fig. D.2.
  - When $(\alpha_g, \beta_g) \in]0, 1[^2$, the beta pdf is $\cup$-shaped, concentrating the activity probability about 0 and 1.
  - When $(\alpha_g, \beta_g) \in]1, +\infty[^2$, the beta pdf is unimodal and concentrates the activity probability about its mode.
  - When $\alpha_g = \beta_g = 1$ the pdf is that of the uniform distribution $\mathsf{Unif}([0, 1])$.
- When $(\alpha_g, \beta_g) \in]0, 1[^2$ and $\alpha_g + \beta_g = 1$, the expected activity probability is $\mathbb{E}[\mathsf{q}_g] = \alpha_g$.

The choice of the parameters $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \dots & \alpha_G \end{bmatrix}^\mathsf{T}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \dots & \beta_G \end{bmatrix}^\mathsf{T}$ are assumed obtained based on some system considerations.

II.2.b Activity correlation

The correlation between any two states is

$$\forall (n, n') \in [N]^2, \ \mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \frac{\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] - \mathbb{E}[\mathsf{s}_n]\,\mathbb{E}[\mathsf{s}_{n'}]}{\sqrt{\mathbb{V}[\mathsf{s}_n]\,\mathbb{V}[\mathsf{s}_{n'}]}}. \tag{D.14}$$

For $n = n'$, it is obvious that $\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_n] = 1$ and hence we assume in the sequel that $n \neq n'$. Denote by g_n (resp. $g_{n'}$) the index of the group to which belongs s_n (resp. $\mathsf{s}_{n'}$). The expectation, variance and cross-expectation are derived in the following.

Expectation We use the law of total expectation to compute the expected state of UE n:

$$\begin{aligned} \mathbb{E}[\mathsf{s}_n] &= \mathbb{E}_{\mathsf{q}_{g_n}}[\mathbb{E}_{\mathsf{s}_n}[\mathsf{s}_n \mid \mathsf{q}_{g_n}]] && \text{(D.15)} \\ &= \mathbb{E}_{\mathsf{q}_{g_n}}[\mathsf{q}_{g_n}] && \text{(D.16)} \\ &= \frac{\alpha_{g_n}}{\alpha_{g_n} + \beta_{g_n}}. && \text{(D.17)} \end{aligned}$$

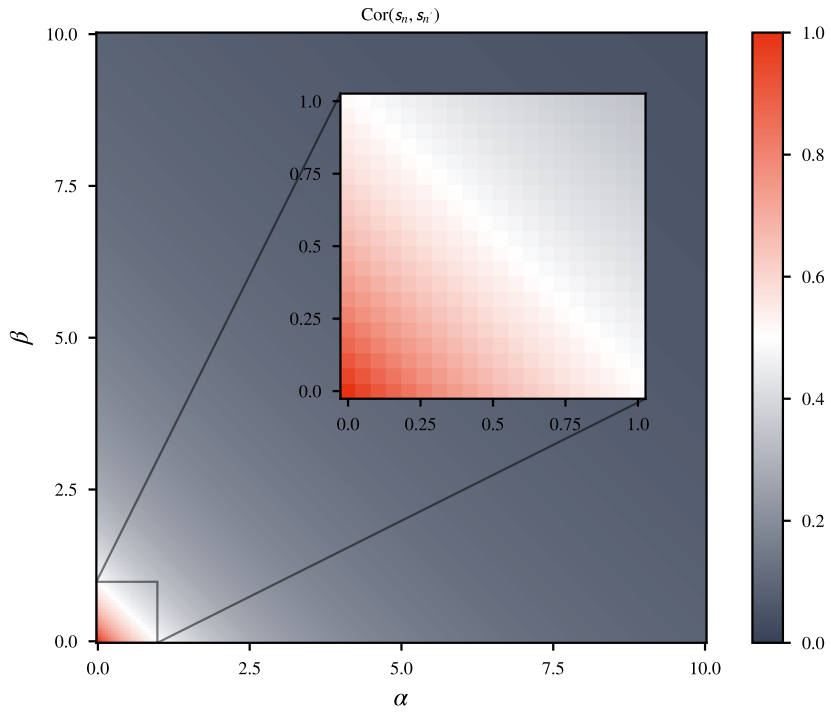


Figure D.3: Correlation between two random activity states s_n and $\mathsf{s}_{n'}$ of the same group as stated in Eq. D.32.

Variance Building on Eq. D.17, we can derive the state variance as:

$$\mathbb{V}[\mathsf{s}_n] = \mathbb{E}\left[\mathsf{s}_n^2\right] - \mathbb{E}[\mathsf{s}_n]^2 \tag{D.18}$$

$$= \mathbb{E}_{\mathsf{q}_{g_n}}\left[\mathbb{E}_{\mathsf{s}_n|\mathsf{q}_{g_n}}\left[\mathsf{s}_n^2 \mid \mathsf{q}_{g_n}\right]\right] - \left(\frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\right)^2 \tag{D.19}$$

$$= \mathbb{E}_{\mathsf{q}_{g_n}}[\mathsf{q}_{g_n}] - \left(\frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\right)^2 \tag{D.20}$$

$$= \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}} - \left(\frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\right)^2 \tag{D.21}$$

$$= \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\left(1 - \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\right) \tag{D.22}$$

$$= \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\frac{\beta_{g_n}}{\alpha_{g_n}+\beta_{g_n}} \tag{D.23}$$

Cross-expectation To compute the cross-expectation between two states s_n and $\mathsf{s}_{n'}$ for $n \neq n'$ (otherwise the correlation is 1), once again we use

the law of total expectations

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \mathbb{E}_{\mathbf{q}}\left[\mathbb{E}_{\mathsf{s}_n,\mathsf{s}_{n'}|\mathbf{q}}[\mathsf{s}_n \mathsf{s}_{n'} \mid \mathbf{q}]\right] \tag{D.24}$$

where the conditionning is done w.r.t. the random vector $\mathbf{q}$. Since the states only depends on the group they belong to, we write

$$\begin{aligned}\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] &= \mathbb{E}_{\mathsf{q}_{g_n},\mathsf{q}_{g_{n'}}}\left[\mathbb{E}_{\mathsf{s}_n|\mathsf{q}_n}[\mathsf{s}_n \mid \mathsf{q}_{g_n}]\,\mathbb{E}_{\mathsf{s}_{n'}|\mathsf{q}_{n'}}\left[\mathsf{s}_{n'} \mid \mathsf{q}_{g_{n'}}\right]\right] &\text{(D.25)}\\ &= \mathbb{E}\left[\mathsf{q}_{g_n}\mathsf{q}_{g_{n'}}\right]. &\text{(D.26)}\end{aligned}$$

If $g_n \neq g_{n'}$, the groups are independent and so we can separate the expectation into

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \mathbb{E}\left[\mathsf{q}_{g_{n'}}\right]\mathbb{E}\left[\mathsf{q}_{g_{n'}}\right] \tag{D.27}$$

and write otherwise

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \mathbb{E}\left[\mathsf{q}_g^2\right] \tag{D.28}$$

which, after using Eqs. (D.12), (D.13), (D.17) and (D.23), summarizes as

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \begin{cases} \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\frac{\alpha_{g_{n'}}}{\alpha_{g_{n'}}+\beta_{g_{n'}}} & \text{if } g_n \neq g_{n'} \\ \frac{\alpha_g\beta_g}{(\alpha_g+\beta_g)^2(\alpha_g+\beta_g+1)} + \left(\frac{\alpha_g}{\alpha_g+\beta_g}\right)^2 & \text{if } g_n = g_{n'} = g \end{cases}. \tag{D.29}$$

Correlation The correlation finally reads

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \frac{\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] - \frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\frac{\alpha_{g_{n'}}}{\alpha_{g_{n'}}+\beta_{g_{n'}}}}{\sqrt{\frac{\alpha_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\frac{\beta_{g_n}}{\alpha_{g_n}+\beta_{g_n}}\frac{\alpha_{g_{n'}}}{\alpha_{g_{n'}}+\beta_{g_{n'}}}\frac{\beta_{g_{n'}}}{\alpha_{g_{n'}}+\beta_{g_{n'}}}}} \tag{D.30}$$

$$= \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } g_n \neq g_{n'} \\ \frac{\frac{\alpha_g\beta_g}{(\alpha_g+\beta_g)^2(\alpha_g+\beta_g+1)}}{\frac{\alpha_g}{\alpha_g+\beta_g}\frac{\beta_g}{\alpha_g+\beta_g}} & \text{if } g_n = g_{n'} = g \end{cases} \tag{D.31}$$

$$= \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } g_n \neq g_{n'} \\ \frac{1}{\alpha_g+\beta_g+1} & \text{if } g_n = g_{n'} = g \end{cases}. \tag{D.32}$$

When $(\alpha_g, \beta_g) \to (0,0)$, it is clear from Eq. D.32 and Fig. D.3 that $\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] \to 1$ i.e. the correlation is maximal. The explanation lies in the interpretation of α_g and β_g as concentrations about 1 and 0 when $(\alpha_g, \beta_g) \in]0,1[^2$; when they simulaneously tends to 0, the infinite branches of the beta pdf approach their vertical asymptote ($\cup \xrightarrow{(\alpha,\beta)\to(0,0)} \sqcup$), allowing to interprete α_g as the probability of a UE's state to be active and β_g the probability to be inactive (see Fig. D.2). Hence, the UEs may either be all active or inactive together.

On the contrary, when $\alpha_g \to +\infty$, the pdf concentrates about its mode m ($\beta_g > 1$) or approaches the right asymptote ($0 < \beta_g < 1$). Symetrically, when $\beta_g \to +\infty$, the pdf concentrates about its mode ($\alpha_g > 1$) or approaches the left asymptote ($0 < \alpha_g < 1$). The consequence is that the activity probability tends to a unique value $\bar{q}$ and the states of the UEs belonging to the same group may be seen as i.i.d. $\mathsf{Bern}(\bar{q})$.

These two remarks lead to the following result that the GHomA model generalizes the independent GS one.

II.3 Active user detection and channel estimation

The focus of the remainder of this chapter is to develop algorithms to identify active UEs and estimate the channel coefficients, namely the joint AUDaCE problem. It is stated in its general form in Sec. B.IV and is now reformulated to account for the GHomA model of Sec. D.II.2. The joint MMSE estimator is

$$\begin{aligned} \boldsymbol{s}^\star, \boldsymbol{H}^\star &= \mathbb{E}[[\boldsymbol{s}, \boldsymbol{H}] \mid \boldsymbol{Y}] \\ &= \int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} [\boldsymbol{s}, \boldsymbol{H}] f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{H}. \end{aligned} \tag{D.33}$$

where the joint posterior pdf is

$$f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) = \frac{\int_{[0,1]^G} f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \int_{[0,1]^G} f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{H}}. \tag{D.34}$$

Different from Eq. B.55, Eq. D.34 includes the additional marginalization w.r.t. the activity probability random vector $\mathbf{q}$ allowing to explicitly introduce the GHomA model. The complexity of this estimator remains prohibitive to be computed as is so that, building on Chap. C, we introduce

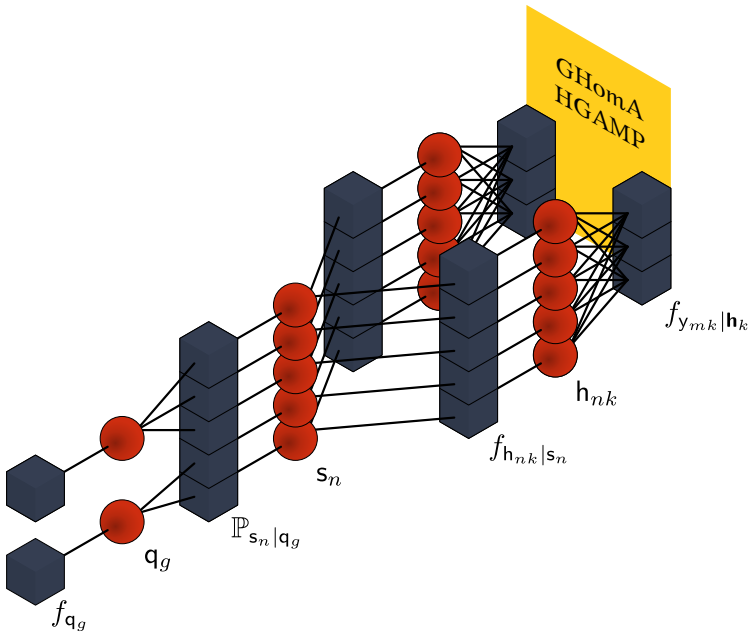


Figure D.4: Underlying factor graph of GHomA-HGAMP.

a LBP and HGAMP algorithms in Sec. D.III.

III Algorithms for the AUDaCE problem with GHomA

In this section, we derive algorithms for the MMSE-based AUDaCE problem (D.33). We first start by writing the LBP algorithm for the system model derived in Sec. D.II. We then develop an algorithm within the HGAMP framework called GHomA-HGAMP from LBP to reduce the complexity of the algorithm with limited loss in performance.

III.1 Loopy belief propagation approach

From Sec. C.II, LBP is an algorithm which aims at solving inference problems in a systematic fashion by exploiting the structure of the joint density of the system variables. The idea is to factorize this joint density into multiple subfactors, each depending on a subset of the system variables. It is then possible to build the factor graph corresponding to the factorization that will be used to build the instance of LBP

From Sec. D.II, the system variables are the entries of $\mathbf{q}, \mathbf{s}, \mathbf{H}$ and $\mathbf{Y}$.

Noting that they form the following Markov chain

$$\mathbf{q} \to \mathbf{s} \to \mathbf{H} \to \mathbf{Y}, \tag{D.35}$$

we can write their joint density as

$$f_{\mathbf{Y},\mathbf{H},\mathbf{s},\mathbf{q}}(\boldsymbol{Y},\boldsymbol{H},\boldsymbol{s},\boldsymbol{q}) = f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \tag{D.36}$$

where each term can be factorized as

$$f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) = \prod_{k=1}^{K} \prod_{m=1}^{M} f_{\mathsf{y}_{mk}|\mathbf{x}_n}(y_{mk} \mid \boldsymbol{x}_n), \tag{D.37a}$$

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) = \prod_{k=1}^{K} \prod_{n=1}^{N} f_{\mathsf{x}_{nk}|\mathsf{s}_n}(x_{nk} \mid s_n), \tag{D.37b}$$

$$\mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) = \prod_{g=1}^{G} \prod_{n \in \mathfrak{G}_g} \mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g}(s_n \mid q_g), \tag{D.37c}$$

$$f_{\mathbf{q}}(\boldsymbol{q}) = \prod_{g=1}^{G} f_{\mathsf{q}_g}(q_g), \tag{D.37d}$$

which follow from the description in Sec. D.II. The factor graph is detailed in Fig. D.4 where the factor nodes, the variable nodes and the edges between them are deduced from (D.36) and (D.37) using the rules in Sec. C.II.

LBP provides a means of approximating the solution to Eq. D.33. In order to develop the algorithm for the model in Sec. D.II, observe that

$$f_{\mathbf{H}|\mathbf{Y}}(\boldsymbol{H} \mid \boldsymbol{Y}) = \frac{\sum_{\boldsymbol{s}} \int_{[0,1]^G} f_{\mathbf{Y},\mathbf{H},\mathbf{s},\mathbf{q}}(\boldsymbol{Y},\boldsymbol{H},\boldsymbol{s},\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}{f_{\mathbf{Y}}(\boldsymbol{Y})}, \tag{D.38a}$$

$$\mathbb{P}_{\mathbf{s}|\mathbf{Y}}(\boldsymbol{s} \mid \boldsymbol{Y}) = \frac{\int_{\mathbb{C}^{N \times K}} \int_{[0,1]^G} f_{\mathbf{Y},\mathbf{H},\mathbf{s},\mathbf{q}}(\boldsymbol{Y},\boldsymbol{H},\boldsymbol{s},\boldsymbol{q}) \mathrm{d}\boldsymbol{q} \mathrm{d}\boldsymbol{H}}{f_{\mathbf{Y}}(\boldsymbol{Y})}, \tag{D.38b}$$

$$f_{\mathbf{q}|\mathbf{Y}}(\boldsymbol{q} \mid \boldsymbol{Y}) = \frac{\int \sum_{\boldsymbol{s}} f_{\mathbf{Y},\mathbf{H},\mathbf{s},\mathbf{q}}(\boldsymbol{Y},\boldsymbol{H},\boldsymbol{s},\boldsymbol{q}) \mathrm{d}\boldsymbol{H}}{f_{\mathbf{Y}}(\boldsymbol{Y})}. \tag{D.38c}$$

The messages exchanged by LBP tailored to Eqs. (D.33), (D.36) and (D.37)

are described in Tab. D.1 and lead to the beliefs

$$\underset{\mathsf{x}_{nk}}{\mathfrak{B}}(x) = \underset{f_{\mathsf{x}_{nk}|\mathsf{s}_n \to \mathsf{x}_n k}}{\mathfrak{M}}(x) \prod_{m\in[M]} \underset{f_{\mathsf{y}_{mk}|\mathbf{x}_k \to \mathsf{x}_n k}}{\mathfrak{M}}(x), \tag{D.39a}$$

$$\underset{\mathsf{s}_n}{\mathfrak{B}}(s) = \underset{f_{\mathsf{x}_{nk}|\mathsf{s}_n \to \mathsf{s}_n}}{\mathfrak{M}}(s) \underset{f_{\mathsf{s}_n|\mathsf{q}_{g_n} \to \mathsf{s}_n}}{\mathfrak{M}}(s) \tag{D.39b}$$

$$\underset{\mathsf{q}_g}{\mathfrak{B}}(q) = \underset{f_{\mathsf{q}_g \to \mathsf{q}_g}}{\mathfrak{M}}(q) \prod_{n\in\mathfrak{G}_g} \underset{f_{\mathsf{s}_n|\mathsf{q}_g \to \mathsf{q}_g}}{\mathfrak{M}}(q) \tag{D.39c}$$

and LBP estimates

$$\hat{h}^i_{nk} = \mathbb{E}\left[\mathsf{h}_{nk}; \underset{\mathsf{h}_{nk}}{\mathfrak{B}}\right], \tag{D.40a}$$

$$\hat{s}^i_n = \mathbb{1}\left(\log \underset{\mathsf{s}_n}{\mathfrak{B}}(1) > \log \underset{\mathsf{s}_n}{\mathfrak{B}}(0)\right). \tag{D.40b}$$

$$\hat{q}^i_g = \mathbb{E}\left[\mathsf{q}_g; \underset{\mathsf{q}_g}{\mathfrak{B}}\right]. \tag{D.40c}$$

at iteration i.

Remark This instance of LBP gives the opportunity to estimate the *a posteriori* group activity probabilities with $\{\hat{q}^i_g\}_{g\in[G]}$ as a byproduct of the joint AUDaCE. However, the focus will not be given to the quality of these estimates and so will be left for future work.

III.2 GHomA-HGAMP algorithm

LBP still suffers from the intractability of the computation of the messages exchanged in the dense part of the factor graph. Indeed, one can see in Tab. D.1 that the integrals required to compute the messages $\underset{f_{\mathsf{y}_{mk}|\mathbf{x}_k \to \mathsf{x}_{nk}}}{\mathfrak{M}}$ are high dimensional, which means that LBP still has a very high complexity.

From Sec. C.III.2, this issue can be addressed by GAMP from which we recall the gaussian approximation

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n \leftarrow \mathsf{x}_{nk}}}{\mathfrak{M}}(h) \simeq \mathcal{CN}(h; \hat{r}^i_{nk}, \hat{\tau}^{i,r}_{nk}) \tag{D.41}$$

where $\hat{r}^i_{nk}$ and $\hat{\tau}^{i,r}_{nk}$ are mean and variance variables iteratively updated by GAMP. This approximation is made possible by using the underlying linear mixing structure ***P*H** under a large system limit assumptions i.e. when both M and N are very large and the coefficients of the preamble

Factor	Variable	Factor $\rightarrow$ Variable Factor $\leftarrow$ Variable		
$f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}$	h_{nk}	$\mathfrak{M}_{f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}\rightarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\displaystyle\int_{\mathbb{C}^{N-1}} f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}(y_{mk}\mid \boldsymbol{h}_{:k})\left[\prod_{n'\in[N]\setminus\{n\}} \mathfrak{M}_{f_{\mathsf{y}_{m'k}\mid\mathbf{h}_k}\leftarrow\mathsf{h}_{n'k}}(h_{n'k})\mathrm{d}h_{n'k}\right]$
		$\mathfrak{M}_{f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}\leftarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\rightarrow\mathsf{h}_{nk}}(h_{nk})\prod_{m'\in[M]\setminus\{m\}} \mathfrak{M}_{f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}\rightarrow\mathsf{h}_{nk}}(h_{nk})$
$f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}$	h_{nk}	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\rightarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\displaystyle\sum_{s=0}^{1} f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}(h_{nk}\mid s)\,\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\leftarrow\mathsf{s}_n}(s)$
		$\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\leftarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\displaystyle\prod_{m=1}^{M} \mathfrak{M}_{f_{\mathsf{y}_{mk}\mid\mathbf{h}_k}\rightarrow\mathsf{h}_{nk}}(h_{nk})$
$f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}$	s_n	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\rightarrow\mathsf{s}_n}(s_n)$	$\propto$	$\displaystyle\int_{\mathbb{C}} f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}(h\mid s_n)\,\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\leftarrow\mathsf{h}_{nk}}(h)\mathrm{d}h$
		$\mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\leftarrow\mathsf{s}_n}(s_n)$	$\propto$	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\rightarrow\mathsf{s}_n}(s_n)\prod_{k'\in[K]\setminus\{k\}} \mathfrak{M}_{f_{\mathsf{h}_{nk'}\mid\mathsf{s}_n}\rightarrow\mathsf{s}_n}(s_n)$
$\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}$	s_n	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\rightarrow\mathsf{s}_n}(s_n)$	$\propto$	$\displaystyle\int_{0}^{1} \mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}(s_n\mid q)\,\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\leftarrow\mathsf{q}_g}(q)\mathrm{d}q$
		$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\leftarrow\mathsf{s}_n}(s_n)$	$\propto$	$\displaystyle\prod_{k=1}^{K} \mathfrak{M}_{f_{\mathsf{h}_{nk}\mid\mathsf{s}_n}\rightarrow\mathsf{s}_n}(s_n)$
$\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}$	q_g	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\rightarrow\mathsf{q}_g}(q_g)$	$\propto$	$\displaystyle\sum_{s} \mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}(s\mid q_g)\,\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\leftarrow\mathsf{s}_n}(s)$
		$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\leftarrow\mathsf{q}_g}(q_g)$	$\propto$	$\mathfrak{M}_{f_{\mathsf{q}_n}\rightarrow\mathsf{q}_g}(q_g)\prod_{n'\in\mathfrak{G}_g\setminus\{n\}} \mathfrak{M}_{\mathbb{P}_{\mathsf{s}_{n'}\mid\mathsf{q}_g}\rightarrow\mathsf{q}_g}(q_g)$
f_{q_n}	q_g	$\mathfrak{M}_{f_{\mathsf{q}_n}\rightarrow\mathsf{q}_g}(q_g)$	$\propto$	$f_{\mathsf{q}_g}(q_g)$
		$\mathfrak{M}_{f_{\mathsf{q}_n}\leftarrow\mathsf{q}_g}(q_g)$	$\propto$	$\displaystyle\prod_{n\in\mathfrak{G}_g} \mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\mid\mathsf{q}_g}\rightarrow\mathsf{q}_g}(q_g)$

Table D.1: Belief propagation messages for the factor graph induced by the joint density (D.36) and the factorizations (D.37).

matrix $\boldsymbol{P}$ scale as $O(1/N)$.

Following the approach in [95], [120], this approximation is then propagated successively in the messages of Tab. D.1, leading to the derivation of an HGAMP algorithm, namely GHomA-HGAMP where the estimates $\hat{x}^i_{nk}$, $\hat{s}^i_n$ and $\hat{q}^i_g$ are computed from D.40. The full derivation given in Sec. G.IV leads to the complete GHomA-HGAMP algorithm given in Algo. D.1.

Interpretations and connections of the algorithm's steps with the factor graph are made in the following lines

- *Lines* 8 – 19*:* The updates of the mean variables $\hat{p}^i_{mk}$, $\hat{z}^i_{mk}$, $\hat{u}^i_{mk}$, $\hat{r}^i_{mk}$ and their associated variances may be seen as messages sent by the factor nodes associated to $f_{\mathsf{y}_{mk}\mid\mathbf{z}_{mk}}$ to the variable nodes h_{nk}.

Algorithm D.1 GHomA-HGAMP

Description: GHomA-HGAMP consists into two parts. The red lines corresponds to the updates of the GAMP variables for estimating the channel and the blue lines corresponds to the updates of the pattern variables with BP. Estimates of the system variables are given in yellow.

1 **input**: $\boldsymbol{Y}, \boldsymbol{P}, \mu_{\mathsf{h}}, \tau_{\mathsf{h}}, \tau_{\mathsf{w}}, I_{\max}$
2 **init**:
3 $i = 0$
4 $\forall (n,k) \in [N] \times [K]\ \hat{h}^i_{nk} = \mu_{\mathsf{h}},\ \hat{\tau}^i_{\mathsf{h},nk} = \tau_{\mathsf{h}}$
5 $\forall (m,k) \in [N] \times [K]\ \hat{u}^i_{mk} = 0$
6 **end**
7 **for** $i \in [I_{\max}]$ **do**:
8 **for** $(m,k) \in [M] \times [K]$ **do**:
9 $\hat{\tau}^i_{\mathsf{p},mk} = \sum_{n=1}^{N} |p_{mn}|^2 \hat{\tau}^{i-1}_{\mathsf{h},nk}$
10 $\hat{p}^i_{mk} = \sum_{n=1}^{N} p_{mn} \hat{h}^{i-1}_{nk} - \hat{\tau}^i_{\mathsf{p},mk} \hat{u}^{i-1}_{mk}$
11 $\hat{\tau}^i_{\mathsf{z},mk} = \tau_{\mathsf{w}} \hat{\tau}^i_{\mathsf{p},mk} / (\hat{\tau}^i_{\mathsf{p},mk} + \tau_{\mathsf{w}})$
12 $\hat{z}^i_{mk} = \hat{p}^i_{mk} + \hat{\tau}^i_{\mathsf{p},mk} (y_{mk} - \hat{p}^i_{mk}) / (\hat{\tau}^i_{\mathsf{p},mk} + \tau_{\mathsf{w}})$
13 $\hat{\tau}^i_{\mathsf{u},mk} = (1 - \hat{\tau}^i_{\mathsf{z},mk}) / (\hat{\tau}^i_{\mathsf{p},mk})^2$
14 $\hat{u}^i_{mk} = (\hat{z}^i_{mk} - \hat{p}^i_{mk}) / \hat{\tau}^i_{\mathsf{p},mk}$
15 **end**
16 **for** $(n,k) \in [N] \times [K]$ **do**:
17 $\hat{\tau}^i_{\mathsf{r},nk} = (\sum_{m=1}^{M} |p_{mn}|^2 \hat{\tau}^i_{\mathsf{u},mk})^{-1}$
18 $\hat{r}^i_{nk} = \hat{h}^{i-1}_{nk} + \hat{\tau}^i_{\mathsf{r},nk} \sum_{m=1}^{M} p_{mn} \hat{u}^i_{mk}$
19 **end**
20 **for** $g \in [G]$ **do**:
21 **for** $n \in \mathfrak{G}_g$ **do**:
22 $\phi_{0,n} = \prod_{k=1}^{K} \mathcal{CN}(0; \hat{r}^i_{nk}, \hat{\tau}^i_{\mathsf{r},nk})$
23 $\phi_{1,n} = \prod_{k=1}^{K} \mathcal{CN}(0; \hat{r}^i_{nk} - \mu_{\mathsf{h}}, \hat{\tau}^i_{\mathsf{r},nk} + \tau_{\mathsf{h}})$
24 $\hat{q}^i_{g,n} = \frac{\int_{[0,1]} q f_{\mathsf{q}_g}(q) \prod_{n' \in \mathfrak{G}_g \setminus \{n\}} [(1-q)\phi_{0,n'} + q\phi_{1,n'}] \mathrm{d}q}{\int_{[0,1]} f_{\mathsf{q}_g}(q) \prod_{n' \in \mathfrak{G}_g \setminus \{n\}} [(1-q)\phi_{0,n'} + q\phi_{1,n'}] \mathrm{d}q}$
25 $\mathsf{LLR}_n = \log\left(\frac{\hat{q}^i_{g,n}}{(1-\hat{q}^i_{g,n})} \frac{\phi_{1,n}}{\phi_{0,n}}\right)$
26 $\hat{s}^i_n = \mathbb{1}(\mathsf{LLR}_n > 0)$
27 $\gamma_n = (1 + \exp(-\mathsf{LLR}_n))^{-1}$
28 **end**
29 $\hat{q}^i_g = \frac{\int_{[0,1]} q f_{\mathsf{q}_g}(q) \prod_{n' \in \mathfrak{G}_g} [(1-q)\phi_{0,n'} + q\phi_{1,n'}] \mathrm{d}q}{\int_{[0,1]} f_{\mathsf{q}_g}(q) \prod_{n' \in \mathfrak{G}_g} [(1-q)\phi_{0,n'} + q\phi_{1,n'}] \mathrm{d}q}$
30 **end**
31 **for** $(n,k) \in [N] \times [K]$ **do**:
32 $\kappa_{nk} = (1/\tau_{\mathsf{h}} + 1/\hat{\tau}^i_{\mathsf{r},nk})^{-1}$
33 $\nu_{nk} = \mu_{\mathsf{h}}/\tau_{\mathsf{h}} + \hat{r}^i_{nk}/\hat{\tau}^i_{\mathsf{r},nk}$
34 $\hat{h}^i_{nk} = \gamma_n \kappa_{nk} \nu_{nk}$
35 $\hat{\tau}^i_{\mathsf{h},nk} = \gamma_n(\kappa_{nk} + |\kappa_{nk}\nu_{nk}|^2) - \left|\hat{h}^i_{nk}\right|^2$
36 **end**
37 **end**

- *Lines* 20 – 30*:* These lines correspond to the portion of the factor graph where BP is applied. First, $\phi_{0,n}$ and $\phi_{1,n}$ represent likelihoods that the variables h_{nk} are active and inactive given $\boldsymbol{Y}$ and $\hat{q}^i_{g,n}$ estimates the group activity probability without the information of UE n, i.e. without the state information provided by $\phi_{0,n}$ and $\phi_{1,n}$. $\hat{s}^i_n$ then estimates the states of the UEs computed based on the log-likelood ratios. Finally, $\hat{q}^i_g$ is the local estimate of the group activity probability, accounting for the state information of all the UEs in the group.
- *Lines* 31 – 37*:* Estimates of the channel coefficient $\hat{h}^i_{nk}$ are updated here, which corresponds to the layer of variable nodes h_{nk}. Note that the *final* estimates $\hat{h}^i_{nk}$ will be set to 0 if the corresponding $\hat{s}^i_n$ is 0 and left untouched otherwise.

Method	Complexity
GAMP (modified) [91]	$O(MNK + NK)$
GS-HGAMP [95]	$O(MNK + NK)$
GHomA-HGAMP (Monte-Carlo)	$O(MNK + NK + NS)$
GHomA-HGAMP (moments)	$O(MNK + NK + \sum_{g=1}^{G} U_g^2)$

Table D.2: Complexity comparisons

To the best of our knowledge, Algo. D.1 is the first HGAMP algorithm relying on the activity probability vector $\boldsymbol{q}$. By viewing $\boldsymbol{q}$ as latent variables, it is possible to incorporate correlation in the activity of the sensors.

III.3 Complexity analysis

We briefly address the complexity of the proposed GHomA-HGAMP algorithm. It is clear that lines 7 to 19 requires $O(MNK)$ floating operations (multiplications and divisions) and lines 31 to 37 $O(NK)$ operations, which overall gives an order of $O(MNK)$ operations.

The complexity bottleneck appears from lines 20 to 29 with the computation of the estimates $\hat{q}_{g,n}^i$ and $\hat{q}_g^i$, which requires the evaluation of high-dimensional integrals.

One could use Monte-Carlo integration techniques to approximate the 4 integrals (2 for the numerators and 2 for the denominators). To this end, consider that S samples $\{q_{g,j}\}_{j\in[S]}$ are drawn (offline) for each of the G beta distributions. Then,

$$\hat{q}_{g,n}^i \approx \frac{\sum_{j=1}^{S} q_{g,j} \prod_{n'\in\mathfrak{G}_g\setminus\{n\}} [(1-q_{g,j})\phi_{0,n'} + q_{g,j}\phi_{1,n'}]}{\sum_{j=1}^{S} \prod_{n'\in\mathfrak{G}_g\setminus\{n\}} [(1-q_{g,j})\phi_{0,n'} + q_{g,j}\phi_{1,n'}]} \tag{D.42}$$

$$\hat{q}_{g}^i \approx \frac{\sum_{j=1}^{S} q_{g,j} \prod_{n'\in\mathfrak{G}_g} [(1-q_{g,j})\phi_{0,n'} + q_{g,j}\phi_{1,n'}]}{\sum_{j=1}^{S} \prod_{n'\in\mathfrak{G}_g} [(1-q_{g,j})\phi_{0,n'} + q_{g,j}\phi_{1,n'}]} \tag{D.43}$$

where each fraction requires $O(SU_g)$ operations.

Another way of seing this calculation resorts to the following computational trick to reduce the complexity of the integration. Observe that the integrands are polynomials in the mute variable q weighted by the density $f_{\mathfrak{q}_g}(q)$. After developing each polynomial, it is thus possible to

write

$$\hat{q}_g^i = \frac{\sum_{j=0}^{U_g+1} \pi_j \mathbb{E}\left[\mathsf{q}_g^j\right]}{\sum_{j=0}^{U_g} \varpi_j \mathbb{E}\left[\mathsf{q}_g^j\right]} \qquad \text{and} \qquad \hat{q}_{g,n}^i = \frac{\sum_{j=0}^{U_g} \pi_{n,j} \mathbb{E}\left[\mathsf{q}_g^j\right]}{\sum_{j=0}^{U_g-1} \varpi_{n,j} \mathbb{E}\left[\mathsf{q}_g^j\right]} \tag{D.44}$$

where the polynomial coefficients are obtained from

$$P_g(q) = q \prod_{n \in \mathfrak{G}_g} [(1-q)\psi_{0,n} + q\psi_{1,n}] \qquad = \sum_{j=0}^{S+1} \pi_j q^j \tag{D.45a}$$

$$Q_g(q) = \prod_{n \in \mathfrak{G}_g} [(1-q)\psi_{0,n} + q\psi_{1,n}] \qquad = \sum_{j=0}^{U_g} \varpi_j q^j \tag{D.45b}$$

$$P_{g,n}(q) = q \prod_{n' \in \mathfrak{G}_g \setminus \{n\}} [(1-q)\psi_{0,n'} + q\psi_{1,n'}] \qquad = \sum_{j=0}^{U_g} \pi_{n,j} q^j \tag{D.45c}$$

$$Q_{g,n}(q) = \prod_{n' \in \mathfrak{G}_g \setminus \{n\}} [(1-q)\psi_{0,n'} + q\psi_{1,n'}] \qquad = \sum_{j=0}^{U_g-1} \varpi_{n,j} q^j \tag{D.45d}$$

Since the polynomials' coefficients $\{\pi_j\}$, $\{\varpi_j\}$, $\{\pi_{n,j}\}$, $\{\varpi_{n,j}\}$ depends on the intermediate variables $\{\psi_{0,n}\}$ and $\{\psi_{1,n}\}$, they cannot be computed offline. Each set of coefficients are computed using a recursive algorithm[1] with at most $O(U_g^2)$ operations from the corresponding roots of the polynomials. Since the computation of the moments $\mathbb{E}\left[\mathsf{q}_g^j\right]$ may be computed offline and so are assumed to be of constant complexity $O(1)$, the estimates can be computed in $O(U_g^2)$ operations.

Hence, using Monte-Carlo integration, the computational complexity of lines 20 to 30 is

$$O(NK + \sum_{g=1}^{G} SU_g) = O(NK + S\sum_{g=1}^{G} U_g) \qquad = O(NK + SN) \tag{D.46}$$

and that of the moment-based method grants

$$O(NK + \sum_{g=1}^{G} U_g^2) = O(NK + \sum_{g=1}^{G} U_g^2). \tag{D.47}$$

[1]For a polynomial $P(X) = \sum_{i=0}^{S} a_i X^i = \prod_{i=1}^{S}(X - r_i)$, identify, for $I \in [S]$, the coefficients of each side of the recursion $P_{I+1}(X) = P_I(X)(X - r_{I+1})$ where $P_I(X) = \prod_{i=1}^{I}(X - r_j)$.

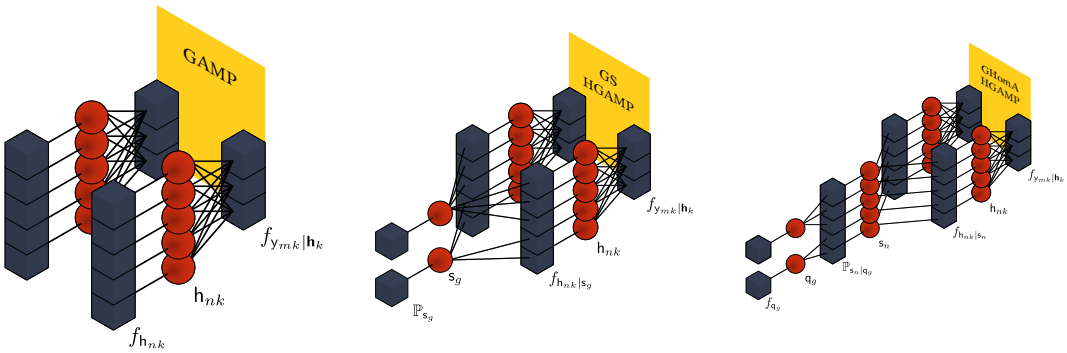


Figure D.5: Example of the underlying factor graphs of the modified GAMP, GS-HGAMP and GHomA-HGAMP

If we compare the Monte-Carlo and this moment-based method, the estimation costs have similar expressions but very different scale. It is well known that Monte-Carlo approximations converges with $\sqrt{S}$, which might slow down the estimation process, resulting in a required large sample size S. It is then likely that $S \gg U_g$ and so $O(NK+SN) \gg O(NK+\sum_{g=1}^{G} U_g^2)$ which therefore motivates the use of the moment-based method. It is worth noting that U_g would remain relatively small in the context of large-scale industrial plant where it is the number G of machines that would increase.

Finally, summing over each block leads to an overall complexity of

$$O(NMK + NK + NS) \qquad \text{(Monte-Carlo)} \qquad \text{(D.48a)}$$

$$O(NMK + NK + \sum_{g=1}^{G} U_g^2) \qquad \text{(moments-based)} \qquad \text{(D.48b)}$$

where the blue term is for GAMP and the red terms are for BP.

A complexity comparison with respect to GAMP and GS-HGAMP is provided in Tab. D.2.

IV Numerical results

IV.1 Framework

In particular, the numerical study compares the following algorithms whose the corresponding factor graphs are shown in Fig. D.5:

Modified GAMP This algorithm corresponds to Sec. C.III.2 but was slightly modified so that it can also perform activity detection. To do so, we have considered a GAMP algorithm with Bernoulli-Gaussian prior for the channel coefficients (see e.g. [118]) so that

$$f_{\mathsf{H}}(\boldsymbol{H}) = \prod_{n\in[N]} \prod_{k\in[K]} (1-q_n)\delta(h_{nk}) + q_n \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}}) \tag{D.49}$$

where the weights $\{q_n\}_{n\in[N]} \in [0,1]^N$ control the sparsity rate. For the purpose of the simulations, those weights will be identified to the parameters $\{\alpha_g\}_g \in [G]$ of the beta distributions, assuming that they verify

$$\alpha_g + \beta_g = 1 \tag{D.50}$$

which allows to inteprete α_g as the average activity probability of UE $n \in \mathfrak{G}_g$. Similar to Eq. APX.103, we then define log-likelihood ratios for this GAMP as

$$\mathsf{LLR}_n = \log \alpha_g + \log \phi_{0,n} - \log \beta_g - \log \phi_{1,n} \tag{D.51}$$

that will be used for activity detection. Note that this modified GAMP does not account for any activity correlation.

GS-HGAMP It was introduced in [95] and addresses the AUDaCE problem with the assumption of group activity, i.e. that the UEs belonging to the same group are either all active or all inactive. Hence this GS-HGAMP will perform AUDaCE by detecting the UEs by groups and estimating the channel coefficients accordingly. This algorithm thus assumes maximum correlation regarding the states of the UEs. For the simulations, the activity probability of the group g will be identified to

$$q_g = \alpha_g \tag{D.52}$$

GHomA-HGAMP It is the algorithm introduced in this chapter. It is briefly reminded that the HGAMP algorithm tailored to the GHomA model introduces correlation at the level of the activity probabilities and not directly at the level of the states. Also, the activity probability of UEs belonging to the same group is assumed to follow a beta distribution

$$q_g \sim \mathsf{Beta}(\alpha_g, 1-\alpha_g) \tag{D.53}$$

Algorithm D.2 Transmission Protocol

1. **step** 0 (Downlink): Sync signal sent by the access point to indicate the beginning of a frame.
2. **step** 1 (Local at Devices): The sensor on each machine detects whether or not the machine is faulty.
3. **step** 2 (Uplink): Sensors on faulty machines transmit pilot sequences.
4. **step** 3 (Local at Access Point): The access point locally measures the room temperature, then performs identification of active sensors and estimates of channel coefficients.
5. **step** 4 (Uplink): Sensors on faulty machines transmit data.

for the simulations, leading to the state correlation Eq. D.32.

IV.2 IIoT scenario

A typical scenario involving GHomA is IIoT. Consider an industrial plant consisting of G machines (= groups), each monitored via U sensors (= UEs) equipped with a single antenna and utilizing a single subcarrier. The total number of sensors is then $N = GU$. The sensors seek to transmit information about the state of the machines they are watching, e.g. faulty behavior, to an AP equipped with K antennas.

For many industrial processes, the machine states are *temperature dependent*: when the local temperature of a machine is outside a particular range, a machine is very likely to experience a faulty behavior. The probability of a fault thus depends on the temperature of the machine. When a fault occurs, the sensors monitoring the machine wake up and must inform the AP of the faulty state as soon as possible. This scenario arises, for example, in semiconductor manufacturing [121]. Other physical variables, such as pressure or light intensity, may also have a similar impact on faults. For the purpose of exposition, we focus on temperature although our framework can also be applied to these alternative variables.

A typical communication protocol between the sensors and the AP is given in Algo. D.2. In **Step 0**, the access point broadcasts a sync signal to all sensors, which indicates the beginning of a frame. At this time, each sensor detects whether or not the corresponding machine is faulty as detailed in **Step 1**. In **Step 2**, each active sensor n transmits its preamble $\boldsymbol{p}_n$ using a GFRA procedure. In **Step 3**, the access point observes the signal Eq. D.1 and performs active user detection in **Step 4**, as well as

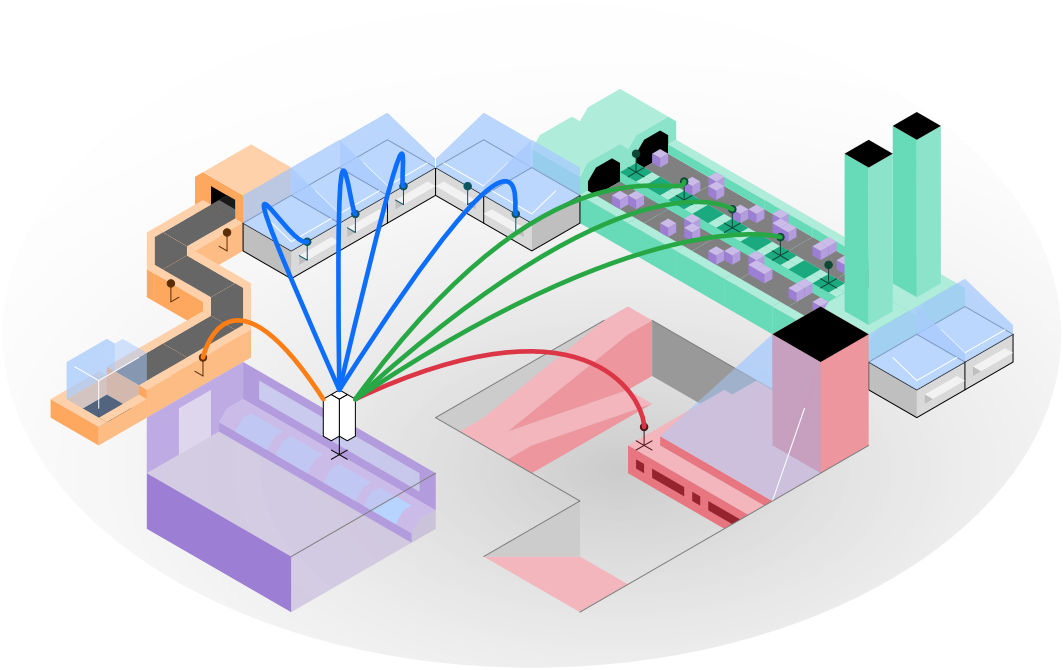


Figure D.6: Industrial IoT with group homogeneous activity.

channel estimation, namely AUDaCE.

Based on the AUDaCE outcomes, the AP is able to

1. identify faulty machines;
2. immediately send back control signaling to the sensors in a reliable way based on the channel estimation assuming reciprocity between UL and DL channels.

Let $\mathsf{T}_g \in]-273.15, +\infty[$ be the temperature of machine $g \in [G]$ in degrees Celsius. We assume that probability that group g is faulty, denoted by q_g, is defined by the function $\Theta_g : \mathsf{T}_g \mapsto \mathsf{q}_g$. Hence, each sensors on this machine will share the same activity probability, inducing a GHomA correlation in their activity as in Sec. D.II.2.b. In practice, the relationship between T_g and q_g is obtained via empirical tests at the design phase of machine g.

On the other hand, the AP does not have direct access to measurements of the temperature within machine g. Instead, as detailed in **Step 3**, the access point can only measure the room temperature locally, denoted by T_0. The (possibly random) temperature, T_0, measured at the access point and the temperature, T_g, at group g are not in general the

Parameter	Description	Value
I_{MC}	# of Monte-Carlo iterations	1000
$I_{\max}$	# of GAMP or HGAMP iterations	100
G	# of machines	64
U	# of sensors per machine	4
N	# of sensors	256
M	preamble length	128
K	# of antennas	2
$\boldsymbol{P}$	Normalized gaussian	N/A
T_0	Temperature at the AP	$\{26.5, 27.5, 28.5, 29.5\}$°C
$\alpha(T_0)$	Beta distribution's parameter	$\{0.15, 0.25, 0.35, 0.45\}$
$\beta(T_0)$		$\{0.85, 0.75, 0.65, 0.55\}$

Table D.3: Simulation settings

same, but are statistically dependent. Hence, we may write

$$\mathsf{q}_g = \Theta_g(\mathsf{T}_g(\mathsf{T}_0)) \tag{D.54}$$

and

$$\mathsf{q}_g|\mathsf{T}_0 = T_0 \sim \mathrm{Beta}(\alpha_g(T_0), \beta_g(T_0)), \tag{D.55}$$

with parameters $\alpha_g(T_0), \beta_g(T_0)$, which depend on the temperature $\mathsf{T}_0 = T_0$ observed by the access point. The motivations for modeling the activity probabilities were given in Sec. D.II.2.a. Note that the precise functional form of $\alpha_g(T_0), \beta_g(T_0)$ depends on the function Θ_g and the statistical dependence between T_0 and $\mathsf{T}_g(\mathsf{T}_0)$, which is established during the design phase of the plant.

IV.3 Settings

We assess the performances of GAMP, GS-HGAMP and GHomA-HGAMP through extensive Monte-Carlo simulations. As a baseline, GS-HGAMP is evaluated with a machine-based and sensor-based modes consisting in the following:

1. the machine-based mode performs active sensor detection taking into account the underlying group structure induced by the machines;
2. the sensor-based mode performs individual detection ignoring the group

structure; i.e., each sensor is assumed to be independent from the other sensors on the same machine. In practice, this means that the number of groups B is assumed to be equal to the number of sensors N (and so $S = 1$).

Performance is assessed in terms of the NMSE and UER (see Sec. B.IV.2), approximated by:

$$\mathsf{NMSE}[\hat{\mathbf{X}}, \mathbf{X}]_{\mathrm{dB}} \approx 10 \log_{10} \left(\frac{1}{I_{\mathrm{MC}}} \sum_{i=1}^{I_{\mathrm{MC}}} \frac{\left\| \hat{\boldsymbol{X}}_i - \boldsymbol{X}_i \right\|_2^2}{\|\boldsymbol{X}_i\|_2^2} \right), \tag{D.56}$$

$$\mathsf{UER}[\hat{\mathbf{s}}, \mathbf{s}] \approx \frac{1}{I_{\mathrm{MC}}} \sum_{i=1}^{I_{\mathrm{MC}}} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(\hat{s}_n^i \neq s_n). \tag{D.57}$$

In particular, the NMSE is a good indicator to assess the ability of the AP to reliably send back the control data to the sensors while the UER provides a means to investigate the impact of the communication system on the ability of the AP to correctly identify faults.

The corresponding results are given in Figs. D.7 to D.10. These figures have been obtained using the settings described in Tab. D.3 where the parameters of the beta distribution are computed using

$$\alpha(T_0) = \frac{|T_0 - 25|}{10} \qquad \text{and} \qquad \beta(T_0) = 1 - \alpha(T_0) \tag{D.58}$$

meaning that the more T_0 deviates from 25°C, the larger the value of $\alpha(T_0)$, the larger the fault probability, and the number of active sensors during the transmission time slot.

IV.4 Discussion

As noted in Chap. C, convergence guarantees of LBP and HGAMP are not necessarily available. With the aid of damped updates for the GAMP and HGAMP algorithms, the results obtained in Figs. D.7 and D.10 were stable without observing divergence after extensive simulations.

IV.4.a UER and NMSE

Observe from Fig. D.7 that GHomA-HGAMP outperforms the modified GAMP and GS-HGAMP. The same order is observed for the UER on Fig. D.10. As the average activity probability $\alpha(T_0)$ increases, the perfor-

mance differences between the algorithms also increase. In more detail:

- Figures D.7 to D.10 (a), $\alpha(T_0) = 0.15$: GAMP, GS-HGAMP (sensor-based) and GHomA-HGAMP have a similar NMSE. In terms of FAR, MDR and UER, GHomA-HGAMP and the modified GAMP have nearly the same performance.
- Figures D.7 to D.10 (b), $\alpha(T_0) = 0.25$: GHomA-HGAMP and the modified GAMP have the lowest NMSE, whereas GS-HGAMP has a significant performance degradation with a 22dB gap. The UER for GHomA-HGAMP and the the modified GAMP are comparable.
- Figures D.7 to D.10 (c), $\alpha(T_0) = 0.35$: each algorithm has a reduction of approximately 10dB at a SNR of 20dB in the NMSE for the channel estimate compared with the case of $\alpha(T_0) = 0.25$. The biggest loss is observed for the modified GAMP with a decrease of around 14dB leaving GHomA-HGAMP to have the best performance. For the UER, two orders of magnitude have been lost for GHomA-HGAMP and the modified GAMP with about a factor 2 to 3 difference between them.
- Figures D.7 to D.10 (d), $\alpha(T_0) = 0.45$: there is a significant degradation in the NMSE and UER for all algorithms due to, on average, a larger number of active devices.

From these observations, we see that GHomA-HGAMP outperforms existing methods for the model described in Sec. D.II. We also observe that despite incorporating some information about the positions of the sensors, GS-HGAMP has poor performance even compared with the modified GAMP. This is due to the fact that the individual behavior of each sensor cannot be accounted for.

However, the modified GAMP does not exploit the prior information of shared activity probability among groups, unlike GHomA-HGAMP. Hence, for larger average activity probability $\alpha(T_0)$, the probability to have multiple sensors of the same group to be active at the same time is also larger. This information is leveraged by GHomA-HGAMP in order to perform a joint detection for sensors belonging to the same group. In contrast, the modified GAMP does not use this information and performs independent detection for those sensors.

These explanations justify the UER of each algorithm and also the values of the NMSE. Indeed, detection errors have a direct consequence

on the channel estimates. When a sensor is detected to be inactive, the corresponding channel estimates will be set to zero. Since GS-HGAMP is subject to a large number of detection errors because of the block detection, a large number of channel estimates will be zero, significantly increasing the errors in channel estimates and, at the same time, degrading the NMSE. The same reasoning applies to the modified GAMP and GHomA-HGAMP, with a weaker impact of the NMSE.

Finally, the performance degradation observed at $\alpha(T_0) = 0.45$ is explained by the potentially large number of devices which might be active relatively to the length of the preambles. As a rule of thumb, an average of 45% of the sensors are active over each transmission meaning that some transmissions may have more sensors active than the length of the preambles. Note that considering larger values of M (i.e., longer preambles) will prevent such an issue.

V Conclusion

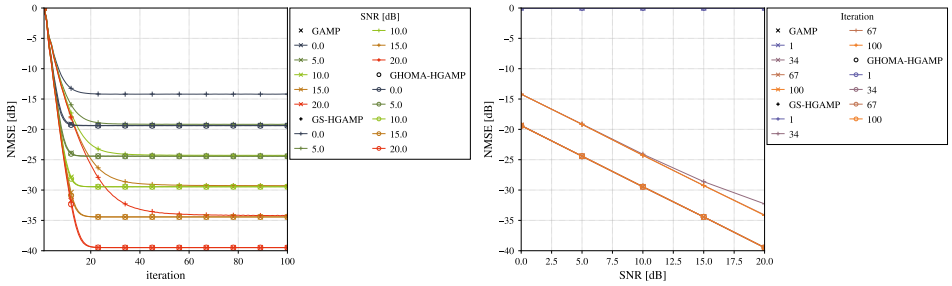
This chapter investigates the AUDaCE problem in the context of GFRA with under a GHomA pattern model. Such a pattern model is shown to generalize the independent group-sparsity activity with the introduction of common latent variables associated with homogeneous activity probilities for each group of UEs.

By means of Monte-Carlo simulations, it is shown that accounting for GHomA with a dedicated HGAMP algorithm leads to significant improvements in both the activity detection and channel estimation w.r.t. when a GS-HGAMP is used.

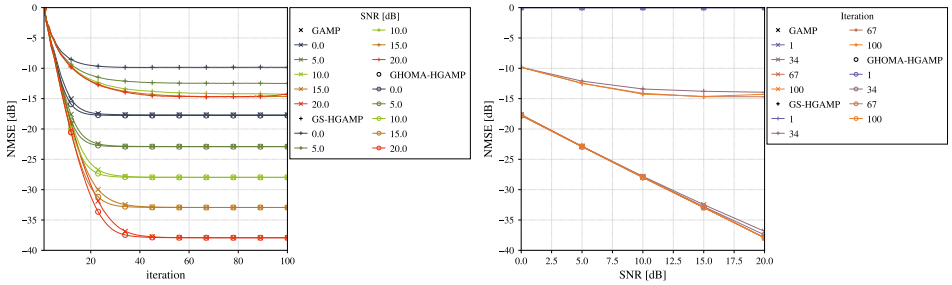
Motivated by these results, it is natural to push further this study by considering more general and more flexible models of the activity pattern. Indeed, although GHomA generalizes group-sparsity models, it is still limited to latent variables per group, not per UE. More precisely, the correlation between two activity states of the same group only depends on this group (see Eq. D.32).

Hence, being able to cope with more *heterogeneous* models is an interesting step forward that we are going to take in Chap. E.

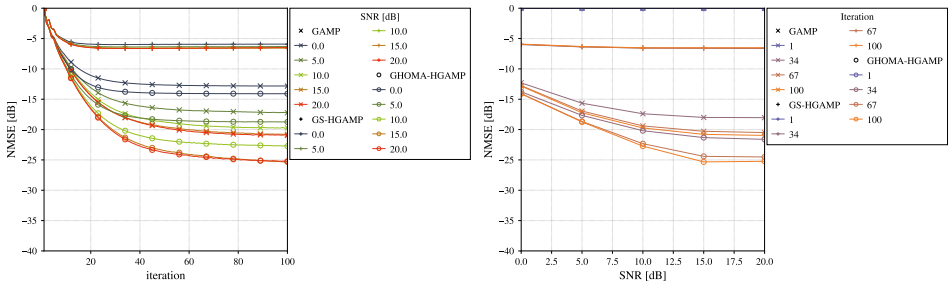
(a) $\alpha = 0.15$



(b) $\alpha = 0.25$



(c) $\alpha = 0.35$



(d) $\alpha = 0.45$

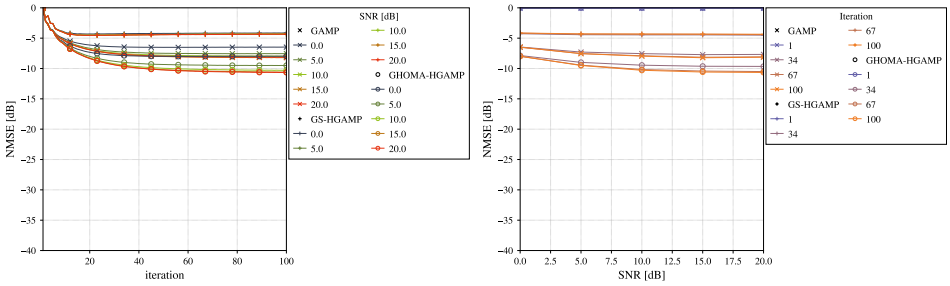
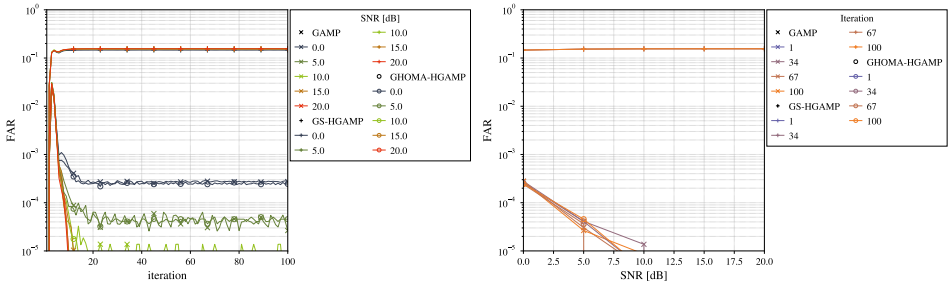
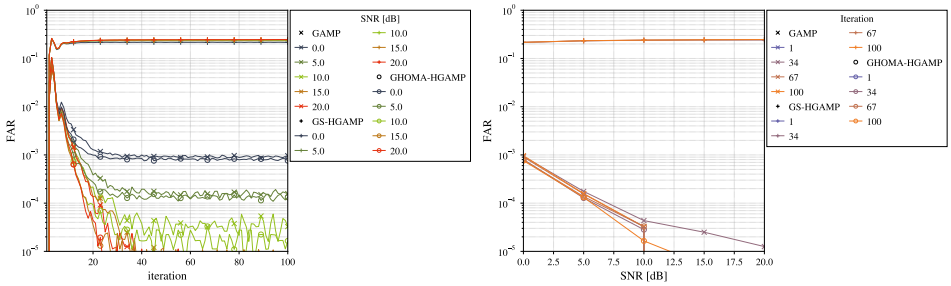


Figure D.7: NMSE for the channel estimation of GHomA-HGAMP against different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

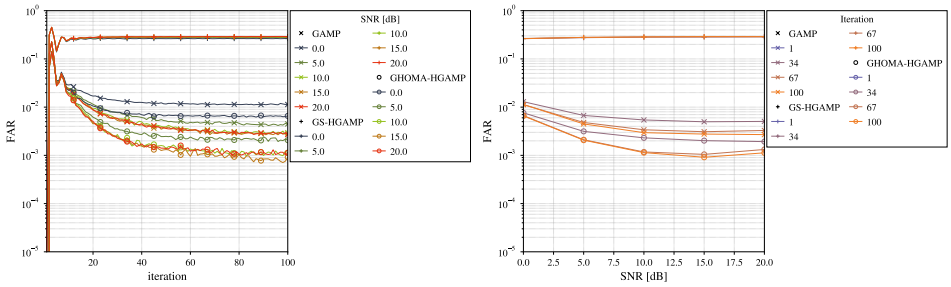
(a) $\alpha = 0.15$



(b) $\alpha = 0.25$



(c) $\alpha = 0.35$



(d) $\alpha = 0.45$

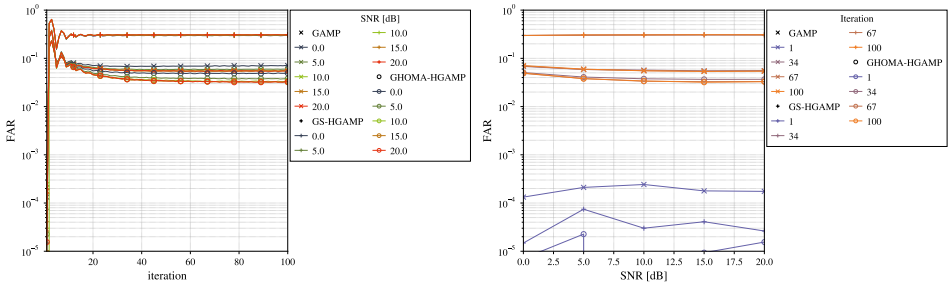
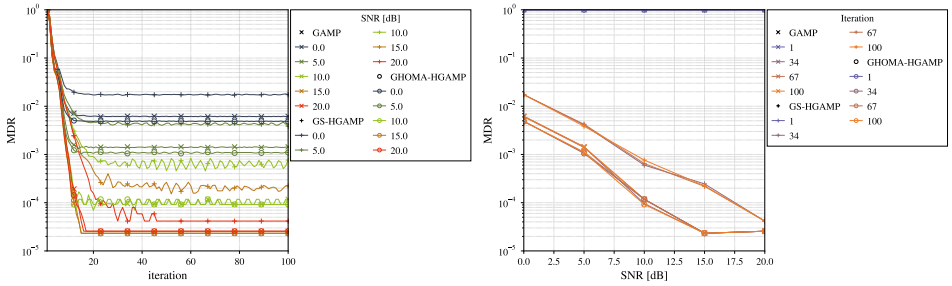
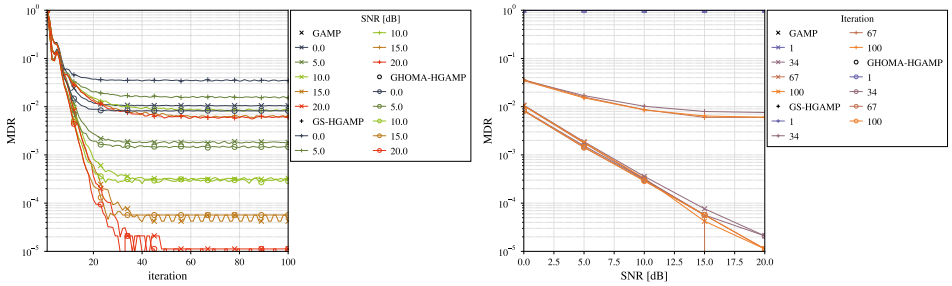


Figure D.8: FAR for the channel estimation of GHomA-HGAMP against different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

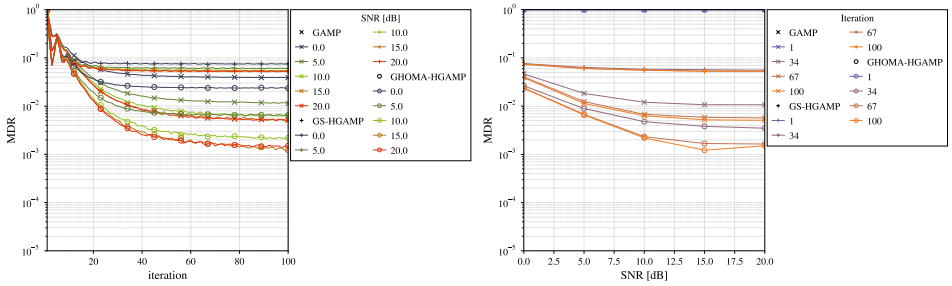
(a) $\alpha = 0.15$



(b) $\alpha = 0.25$



(c) $\alpha = 0.35$



(d) $\alpha = 0.45$

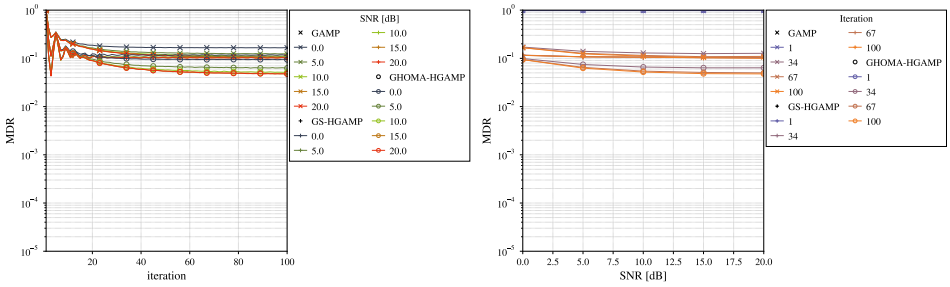
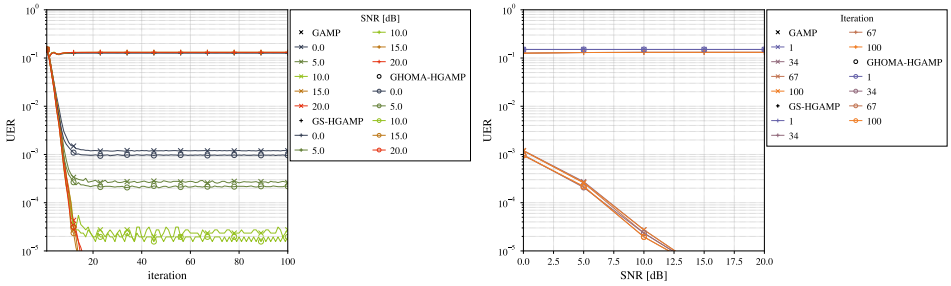
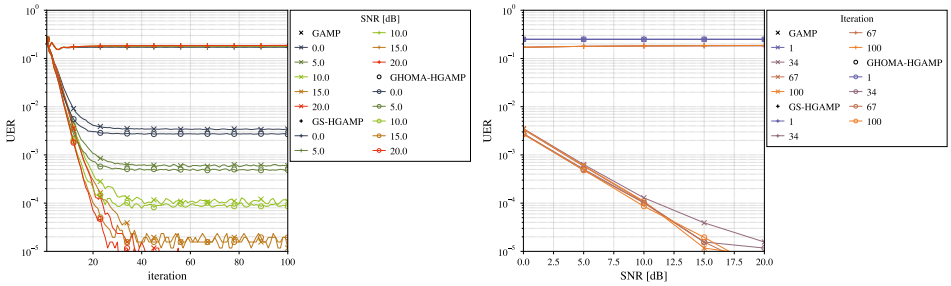


Figure D.9: MDR for the channel estimation of GHomA-HGAMP against different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

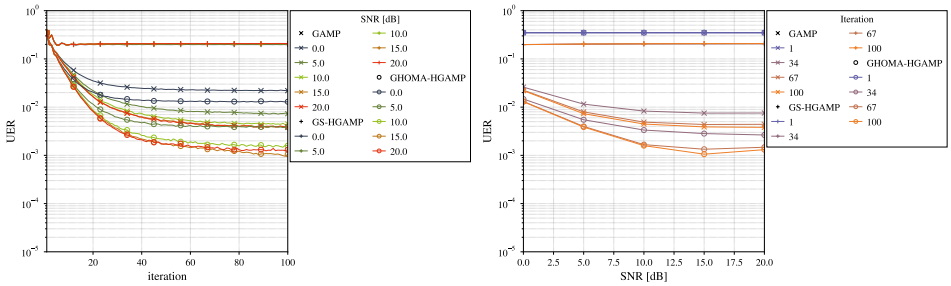
(a) $\alpha = 0.15$



(b) $\alpha = 0.25$



(c) $\alpha = 0.35$



(d) $\alpha = 0.45$

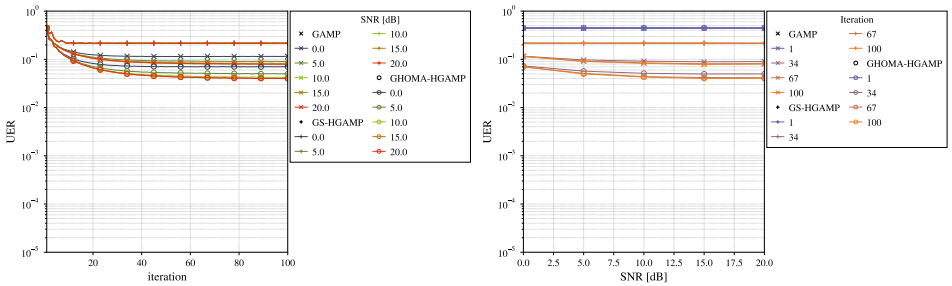


Figure D.10: UER for the channel estimation of GHomA-HGAMP against different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

E

HYBRID GENERALIZED APPROXIMATE MESSAGE PASSING FOR AUDACE WITH CORRELATED GROUP HETEROGENEOUS ACTIVITY

I INTRODUCTION

In Chap. D, AUDaCE for GFRA was studied under the GHomA model of Sec. D.II.2. This is suited for scenarios involving wireless networks where UEs are spread over multiple groups within a cell. Within a group, the UEs share the same probability to be active, assuming some sort of homogeneity between them. The validity of this assumption highly depends on the system considered. For instance, back to the IIoT scenario of Sec. D.IV.2, if each machine is equiped with sensors of the same quality, it may be reasonable to consider a GHomA pattern. However, at some point, a part of the sensors may break and will be replaced by others with the same applicative features. If the new sensors were fabricated with cheaper components, the homogeneity of the group would no longer hold since the sensors capability to detect the machine faults

would not be the same. Such *heterogeneity* in the detection would lead to heterogeneous activity states that cannot be properly captured by the GHomA model.

However, modeling statistically dependent states in a flexible way is a difficult task since the variables are discrete. In the literature, the theory of copula is known to have been successfully leveraged for developing models where complicated statistical dependence arises. Originally, copula-based models are used in the finance, economic or insurance fields [122]. In the field of telecommunications, copula are popular in radar analysis [123], interference modeling for IoT wireless networks [124] or for heterogeneous data recovery from distributed wireless sensor networks [125].

Hence we develop in this chapter a new model accounting for GHetA pattern. Different from Chap. D, this is achieved by applying the copula theory to the vector of latent activity probability random variables in order to propose a general model allowing for flexible design in the dependence structure of the UEs' activity pattern.

The joint AUDaCE problem is then formulated under this new prior information of heterogeneous and statistically dependent probabilities of activity. Similarly to Chap. D, a GHetA-HGAMP algorithm fitting this new model is developped.

It is then benchmarked against the other GAMP-based algorithms and GHomA-HGAMP with extensive Monte-Carlo simulations, in different activity correlation regimes. The results show the flexibility and the gains of using a tailored HGAMP to GHetA-based GFRA for AUDaCE compared to existing GAMP, GS-HGAMP and GHomA-HGAMP. Also, a numerical study of the robustness of GHetA-HGAMP to mismatched correlation is performed, spotlighting the importance of accurately knowing a true high activity correlation to improve the detection and estimation capability of the algorithm.

I.1 Contributions

The contributions of this chapter are summarized as follows:

1. The system model of Sec. D.II is generalized to consider group heterogeneous activity pattern leveraging the copula theory.
2. The corresponding activity correlation is semi-analytically derived.

3. An instance of LBP is derived to address the joint AUDaCE problem.
4. This instance is approximated within the framework of HGAMP by deriving a tailored GHetA-HGAMP algorithm.
5. GHetA-HGAMP is numerically evaluated and proved its superiority for AUDaCE in all the studied scenarios against GHomA-HGAMP, GS-HGAMP and GAMP.
6. It is also shown that GHetA-HGAMP is robust to mismatched correlation in the low activity correlation regime and less robust in the high correlation regime.

I.2 Organization

In Sec. E.II, the system model as well as the new GHetA activity pattern are described for this chapter. The next Sec. E.III tailors the HGAMP framework to the new GHetA-based AUDaCE problem. The performances are analyzed in Sec. E.IV and compared to others GAMP-based methods. The chapter is concluded in Sec. E.V.

II System model

We introduce in this section the system model based on Sec. B.II. The transmission part is similar to Sec. D.II. However, the activity model is changed to introduce heterogeneous patterns. This is where the theory of copula will be leveraged to flexibly describe the activity of UEs for correlated GFRA schemes.

II.1 Transmission modeling

The received signal is still

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \qquad \text{where} \qquad \mathbf{Z} = \boldsymbol{P}\mathbf{H} \tag{E.1}$$

where we remind that $\mathbf{Y} \in \mathbb{C}^{M\times K}$, $\boldsymbol{P} \in \mathbb{C}^{M\times N}$, $\mathbf{H} \in \mathbb{C}^{N\times K}$ and $\mathbf{W} \in \mathbb{C}^{M\times K}$. The same channel distribution and noise distributions are assumed i.e.

$$\forall (n,k) \in [N]\times[K],\ \mathsf{h}_{nk} \mid \mathsf{s}_n = s \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } s = 0 \\ \mathsf{CNorm}(\mu_{\mathsf{h}}, \tau_{\mathsf{h}}) & \text{if } s = 1 \end{cases} \tag{E.2}$$

and

$$\forall (m,k) \in [M] \times [K],\ \mathsf{w}_{mk} \sim \mathsf{CNorm}(\mu_{\mathsf{w}}, \tau_{\mathsf{w}})\,. \tag{E.3}$$

The resulting factorizations are

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) = \prod_{n=1}^{N}\prod_{k=1}^{K} f_{\mathsf{h}_{nk}|\mathsf{s}_n}(h_{nk} \mid s_n) \tag{E.4}$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} \delta(h_{nk})^{1-s_n}\mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})^{s_n} \tag{E.5}$$

and

$$f_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{Y} \mid \boldsymbol{Z}) = \prod_{m=1}^{N}\prod_{k=1}^{K} f_{\mathsf{y}_{mk}|\mathsf{z}_{mk}}(h_{nk} \mid z_{mk}) \tag{E.6}$$

$$= \prod_{m=1}^{N}\prod_{k=1}^{K} \mathcal{CN}(y_{mk}; z_{mk} + \mu_{\mathsf{w}}, \tau_{\mathsf{w}}). \tag{E.7}$$

II.2 Group heterogeneous activity pattern

We aim at generalizing further the GHomA activity pattern from Sec. D.II.2 to provide a new characterization of its joint pmf $\mathbb{P}_{\mathbf{s}}$. To do so, we first start by considering a similar pattern model

$$\forall n \in [N],\ \mathsf{s}_n \mid \mathsf{q}_n = q_n \sim \mathsf{Bern}(q_n) \tag{E.8}$$

where the random states are independently Bernoulli distributed given their activity probabilities $\mathbf{q} = [\mathsf{q}_n]_n^{\mathsf{T}}$ are equal to $\boldsymbol{q} = [q_n]_n^{\mathsf{T}} \in [0,1]^N$. The first difference from the GHomA pattern is that each state depends on its own activity probability.

Since our objective is to generalize GHomA, and so introduce a more general correlated activity pattern model, we are going to characterize the distribution of the random vector $\mathbf{q}$ in a way that allows flexibility in its possible correlation structures. *Copula* are a handy mathematical tool to achieve our goal that we briefly explain hereafter.

II.2.a Copula

A copula is a joint cumulative distribution function (cdf) of a random vector $\mathbf{u} = [\mathsf{u}_n]_{n\in[N]}^{\mathsf{T}}$ with uniform marginal cdfs on $[0,1]^N$. Now consider the random vector $\mathbf{q}$ and, for $n \in [N]$, denote by F_{q_n} the marginal cdf of q_n. We then have

$$\mathsf{q}_n = F_{\mathsf{q}_n}^{-1}(\mathsf{u}_n) \tag{E.9}$$

for a continuous uniform random variable $\mathsf{u}_n \sim \mathsf{Unif}(0,1)$. Denote the joint cdf of $\mathbf{u}$ by

$$F_{\mathbf{u}}(\boldsymbol{u}) : \begin{cases} [0,1]^N & \to [0,1] \\ \boldsymbol{u} & \mapsto \mathbb{P}(\mathsf{u}_1 \leq u_1, \ldots, \mathsf{u}_N \leq u_N) \end{cases}. \tag{E.10}$$

It is easy to see that $F_{\mathbf{u}}$ is a copula for the random vector $\mathbf{q}$ since

$$\begin{aligned} F_{\mathbf{u}}(\boldsymbol{u}) &= \mathbb{P}(\mathsf{u}_1 \leq u_1, \ldots, \mathsf{u}_N \leq u_N) && \text{(E.11)} \\ &= F(\mathsf{q}_1 \leq F_{\mathsf{q}_1}^{-1}(u_1), \ldots, \mathsf{q}_N \leq F_{\mathsf{q}_N}^{-1}(u_N),) && \text{(E.12)} \\ &= F_{\mathbf{q}}(F_{\mathsf{q}_1}^{-1}(u_1), \ldots, F_{\mathsf{q}_N}^{-1}(u_N)) && \text{(E.13)} \end{aligned}$$

or, differently said,

$$F_{\mathbf{q}}(\boldsymbol{q}) = F_{\mathbf{u}}(F_{\mathsf{q}_1}(q_1), \ldots, F_{\mathsf{q}_N}(u_N)) \tag{E.14}$$

which is a joint cdf on $\mathbf{q}$ with uniform marginals. This is in fact a consequence of Sklar's theorem which states that any multivariate cdf can be expressed in terms of its marginals and a copula Cop as

$$F_{\mathbf{q}}(\boldsymbol{q}) = \mathrm{Cop}(F_{\mathsf{q}_1}(q_1), \ldots, F_{\mathsf{q}_N}(q_N)). \tag{E.15}$$

From Eq. E.13, the copula Cop is $F_{\mathbf{u}}$. An interesting consequence of this theorem is that if $\mathbf{u}$ is a correlated random vector, so will be $\mathbf{q}$. One can also write the corresponding copula pdf as

$$f_{\mathbf{q}}(\boldsymbol{q}) = \mathrm{cop}(F_{\mathsf{q}_1}(q_1), \ldots, F_{\mathsf{q}_N}(q_N)) \prod_{n=1}^{N} f_{\mathsf{q}_n}(q_n). \tag{E.16}$$

where $\mathrm{cop}(\boldsymbol{u}) = \frac{\partial^N \mathrm{Cop}(\boldsymbol{u})}{\partial u_1 \ldots \partial u_N}$.

However, modeling a correlated random uniform vector cannot be achieved by a standard parametric model as it would be for, e.g. with a correlated gaussian random vector. Using once again Sklar's theorem allows to make the connection between $\mathbf{u}$ and a random vector $\mathbf{c}$ with a known statistical dependence structure. To summarize, the complete chain of transform from $\mathbf{c}$ to $\mathbf{q}$ is then

$$\mathbf{c} \to \mathbf{u} \to \mathbf{q} \tag{E.17}$$

where

1. $\mathbf{c}$ is a correlated random vector with known parametric distribution;
2. $\mathbf{u}$ is a correlated uniform random vector obtained as $\forall n \in [N], \mathsf{u}_n = F_{\mathsf{c}_n}(\mathsf{c}_n)$;
3. $\mathbf{q}$ is a correlated beta random vector obtained as $\forall n \in [N], \mathsf{u}_n = F^{-1}_{\mathsf{q}_n}(\mathsf{u}_n)$.

We denote the complete transform from $\mathbf{c}$ to $\mathbf{q}$ by

$$\boldsymbol{T}(\mathbf{c}) = \begin{bmatrix} T_1(\mathsf{c}_1) \\ \vdots \\ T_N(\mathsf{c}_N) \end{bmatrix} = \begin{bmatrix} (F^{-1}_{\mathsf{q}_1} \circ F_{\mathsf{c}_1})(\mathsf{c}_1) \\ \vdots \\ (F^{-1}_{\mathsf{q}_N} \circ F_{\mathsf{c}_N})(\mathsf{c}_N) \end{bmatrix} \tag{E.18}$$

and the resulting copula cdf and pdf are

$$F_{\mathbf{q}}(\boldsymbol{q}) = F_{\mathbf{c}}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) \tag{E.19}$$

$$f_{\mathbf{q}}(\boldsymbol{q}) = f_{\mathbf{c}}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) \prod_{n=1}^{N} \partial_{q_n} T_n^{-1}(q_n) \tag{E.20}$$

$$= f_{\mathbf{c}}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) \prod_{n=1}^{N} \partial_{q_n} (F^{-1}_{\mathsf{c}_n} \circ F_{\mathsf{q}_n})(q_n) \tag{E.21}$$

$$= f_{\mathbf{c}}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) \prod_{n=1}^{N} f_{\mathsf{q}_n}(q_n) \frac{1}{(f_{\mathsf{c}_n} \circ F^{-1}_{\mathsf{c}_n} \circ F_{\mathsf{q}_n})(q_n)} \tag{E.22}$$

$$= \mathrm{cop}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) \prod_{n=1}^{N} f_{\mathsf{q}_n}(q_n) \tag{E.23}$$

where

$$\mathrm{cop}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_1)) = \frac{f_{\mathbf{c}}(T_1^{-1}(q_1), \dots, T_N^{-1}(q_N))}{\prod_{n=1}^{N} (f_{\mathsf{c}_n}^{-1} \circ T_n^{-1})(q_n)}. \tag{E.24}$$

Algorithm E.1 Sampling from copula

input:

sample size N

correlated distribution $\mathsf{Dist}(\boldsymbol{\theta})$ on $\mathbb{R}^N$ (or $\mathbb{C}^N$)

the cdf $F_{\mathbf{c}}$ and the marginal cdfs $\{F_{\mathsf{c}_n}\}_{n\in[N]}$

inverse marginal cdfs $\left\{F^{-1}_{\mathsf{q}_n}\right\}_{n\in[N]}$

end

Generate S samples $\boldsymbol{c} \sim \mathsf{Dist}(\boldsymbol{\theta})$.

Form the vector $\boldsymbol{u} = [F_{\mathsf{c}_n}(c_n)]^{\mathsf{T}}_{n\in[N]}$

Form the vector $\boldsymbol{q} = [F^{-1}_{\mathsf{q}_n}(u_n)]^{\mathsf{T}}_{n\in[N]}$

return: $\boldsymbol{q}$

Based on Eq. E.17 and the remarks above, we see that generating correlated random vectors for simulation purposes is rather simple, provided that one has access to an efficient sampling of $\mathbf{c}$, its joint and marginal cdfs, and the inverse marginals cdfs of $\mathbf{q}$. The pseudo-code to sample $\mathbf{q}$ is given in Algo. E.1 and was used in the following example.

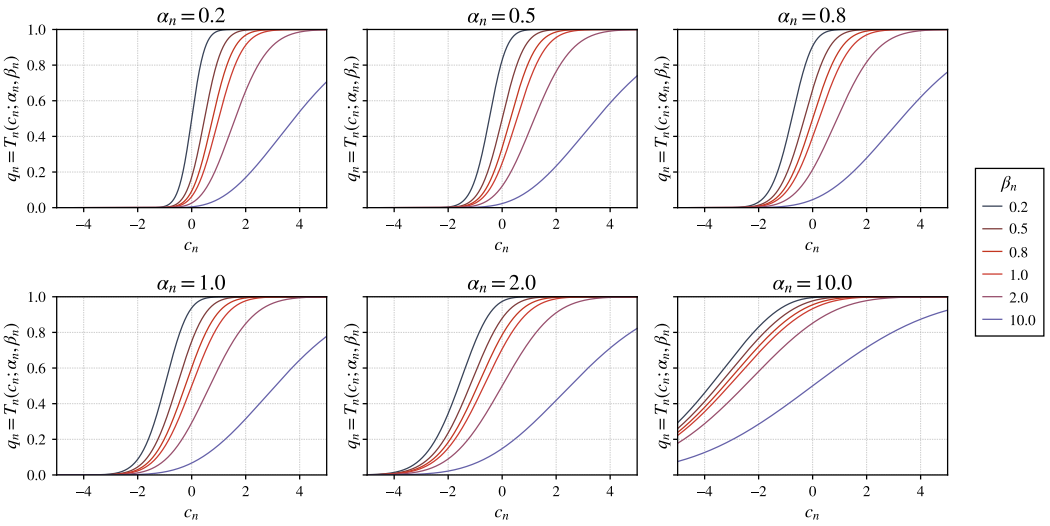
Example with gaussian copula We take the example of a correlated beta random vector with gaussian copula. We start with the random vector $\mathbf{c}$ that we assume to be correlated gaussian distributed as

$$\mathbf{c} \sim \mathsf{Norm}(\mathbf{0}_{N,1}, \boldsymbol{K}_{\mathbf{c}}) \qquad \text{where} \qquad \begin{cases} \boldsymbol{K}_{\boldsymbol{c}} \in [-1,1]^{N\times N} \\ \boldsymbol{K}_{\boldsymbol{c}} \succeq 0 \end{cases} . \tag{E.25}$$

Note that the covariance matrix $\boldsymbol{K}_{\mathbf{c}}$ is restricted to be a correlation matrix but can be semi-definite positive, i.e. possibly not invertible. The reason is to allow some particular structures of the correlation matrix. For instance, one can consider a correlation matrix of the form

$$\boldsymbol{K}_{\boldsymbol{c}} = \begin{bmatrix} \boldsymbol{K}_{\boldsymbol{c},1} & & \\ & \ddots & \\ & & \boldsymbol{K}_{\boldsymbol{c},G} \end{bmatrix} \in \mathbb{R}^{N\times N} \tag{E.26}$$

(a) Gaussian to beta



(b) Beta to gaussian

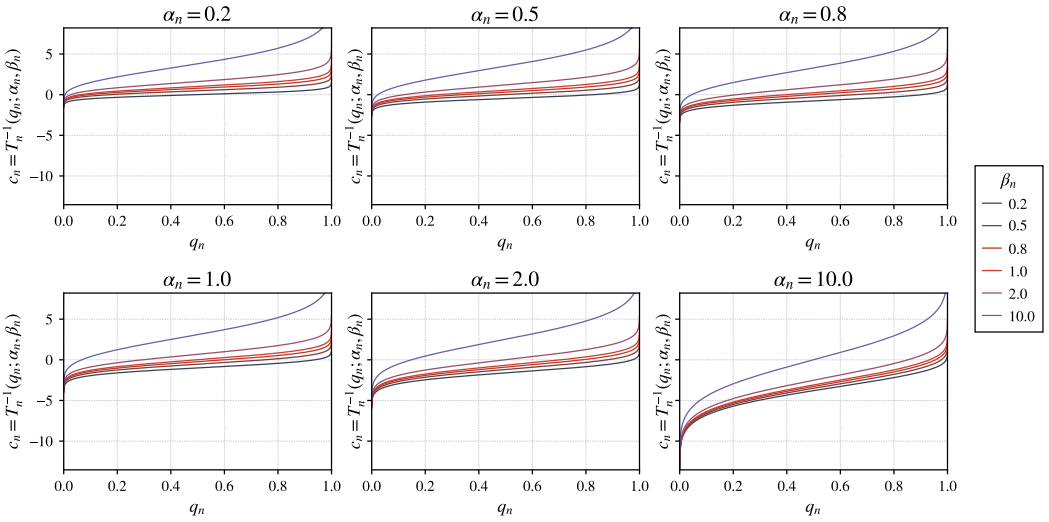


Figure E.1: Copula transform as in Eq. E.18.

which is block diagonal. The, non-necessarily identical, block matrices are

$$\forall g \in [G], \quad \boldsymbol{K}_{\boldsymbol{c},g} = \rho_g \mathbf{1}_{U_g \times U_g} + (1 - \rho_g) \boldsymbol{I}_{U_g}. \qquad \text{and} \qquad \rho_g > 0 \tag{E.27}$$

where $N = \sum_{g=1}^{G} U_g$. A consequence of such structure is that the marginals are standard normal distributions

$$F_{\mathsf{c}_n}(c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2}} \exp\left(-\frac{c^2}{2}\right) \mathrm{d}c. \tag{E.28}$$

Next, we assume that the activity probability vector $\mathbf{q}$ has beta marginal distributions such as

$$\forall n \in [N],\ \mathsf{q}_n \sim \mathsf{Beta}(\alpha_n, \beta_n) \tag{E.29}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameter vectors. We then obtain the correlated random vector $\mathbf{q}$ with the component-wise transform

$$\mathbf{q} = \boldsymbol{T}(\mathbf{c}) = \begin{bmatrix} F^{-1}_{\mathsf{q}_1}(F_{\mathsf{c}_1}(\mathsf{c}_1); \alpha_1, \beta_1) \\ \vdots \\ F^{-1}_{\mathsf{q}_N}(F_{\mathsf{c}_N}(\mathsf{c}_N); \alpha_N, \beta_N) \end{bmatrix} \tag{E.30}$$

This leads to the gaussian copula pdf

$$\mathrm{cop}(T_1^{-1}(q_1), \ldots, T_N^{-1}(q_1)) = \frac{\mathcal{N}(\boldsymbol{T}(\mathbf{c}); \mathbf{0}_{N\times 1}, \boldsymbol{K_c})}{\mathcal{N}(\boldsymbol{T}(\mathbf{c}); \mathbf{0}_{N\times 1}, \boldsymbol{I}_N)} \tag{E.31}$$

$$= \frac{1}{\sqrt{\det(\boldsymbol{K_c})}} \exp\left(-\boldsymbol{T}(\mathbf{c})^{\mathsf{T}}(\boldsymbol{K_c}^{-1} - \boldsymbol{I}_N)\boldsymbol{T}(\mathbf{c})\right) \tag{E.32}$$

The copula transform of Eq. E.30 is plotted in Fig. E.1 as well as the inverse transform. The complete process of generating correlated samples of $\mathbf{q}$ from samples of $\mathbf{c}$ is shown in Fig. E.2.

Note that when one of the correlation block $\boldsymbol{K}_{\mathbf{c},g}$ is equal to $\mathbf{1}_{U_g \times U_g}$, the corresponding random variables of $\mathbf{c}$ will be equal and so will be that of $\mathbf{u}$ and $\mathbf{q}$. Hence, choosing a correlation matrix of the form

$$\boldsymbol{K_c} = \begin{bmatrix} \mathbf{1}_{U_1\times U_1} & & \\ & \ddots & \\ & & \mathbf{1}_{U_G \times U_G} \end{bmatrix} \tag{E.33}$$

will lead to a model fully characterized by a single activity probability

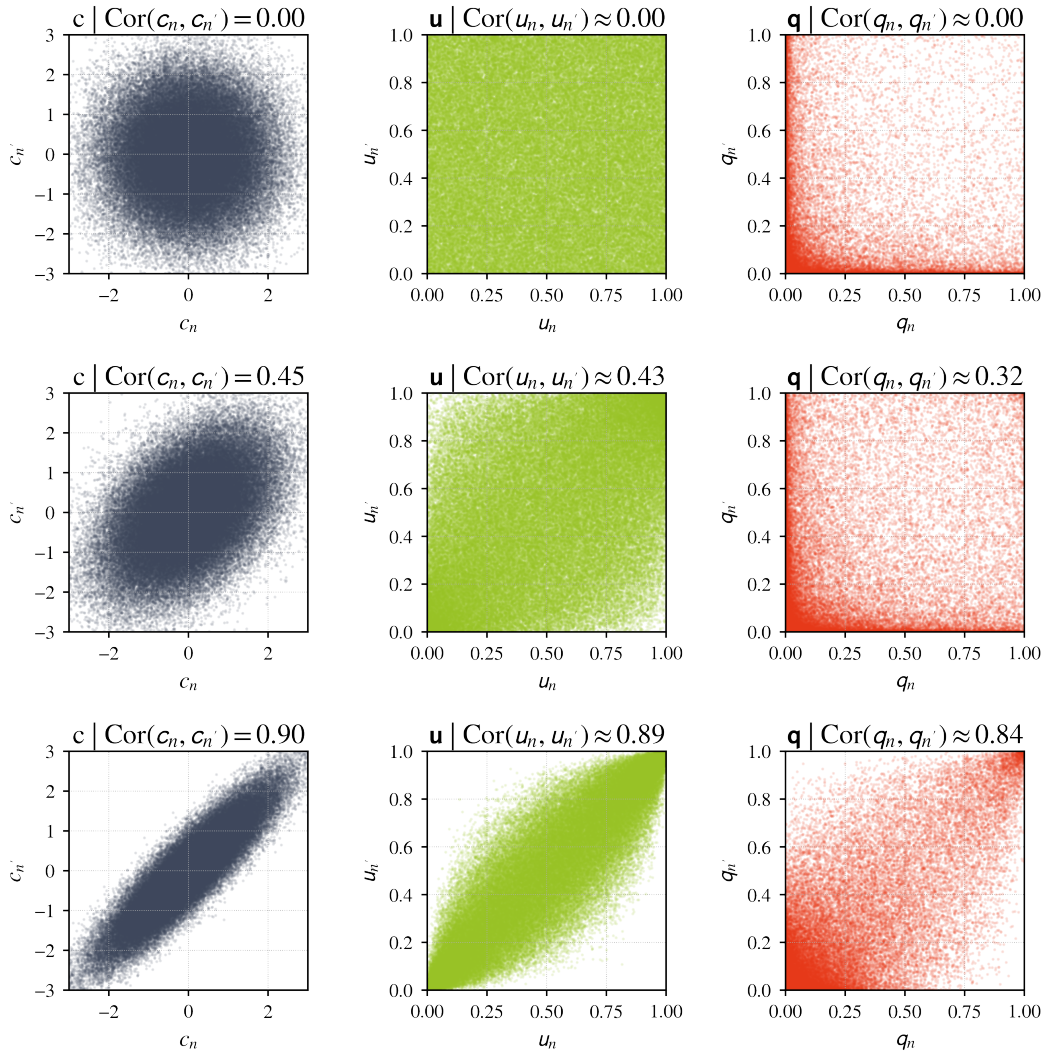


Figure E.2: Correlated beta random vector with $(\alpha, \beta) = (0.1, 0.9)$ using gaussian copula.

random variable for each group since

$$\forall g \in [G],\ \mathsf{q}_{\sum_{g'<g} U_{g'}+1} = \cdots = \mathsf{q}_{\sum_{g'\leq g} U_{g'}} = \mathsf{q}_g \tag{E.34}$$

and so GHetA generalizes GHomA.

II.2.b Correlation of the activity states

The correlation of the activity states with GHetA can now be formally derived, and similar to Sec. D.II.2.b, we write

$$\forall (n, n') \in [N]^2,\ n \neq n',\ \mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \frac{\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] - \mathbb{E}[\mathsf{s}_n]\,\mathbb{E}[\mathsf{s}_{n'}]}{\sqrt{\mathbb{V}[\mathsf{s}_n]\,\mathbb{V}[\mathsf{s}_{n'}]}}. \tag{E.35}$$

where we assume that $n \neq n'$, removing the trivial case $\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_n] = 1$. Keeping the beta model for the activity probability, the expectation and variance remains

$$\mathbb{E}[\mathsf{s}_n] = \frac{\alpha_n}{\alpha_n + \beta_n} \qquad \text{and} \qquad \mathbb{V}[\mathsf{s}_n] = \frac{\alpha_n}{\alpha_n + \beta_n} \frac{\beta_n}{\alpha_n + \beta_n}. \tag{E.36}$$

The expected cross-product is

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \mathbb{E}[\mathsf{q}_n \mathsf{q}_{n'}] \tag{E.37}$$

Since $\mathbf{q}$ is modeled according to the component-wise copula transform of Eq. E.18, a consequence is that if the underlying random variables c_n and $\mathsf{c}_{n'}$ were independent, so will be q_n and $\mathsf{q}_{n'}$ and then

$$\mathbb{E}[\mathsf{s}_n \mathsf{s}_{n'}] = \mathbb{E}[\mathsf{q}_n]\,\mathbb{E}[\mathsf{q}_{n'}] = \frac{\alpha_n}{\alpha_n + \beta_n} \frac{\alpha_{n'}}{\alpha_{n'} + \beta_{n'}} \tag{E.38}$$

However, if they are dependent, a closed-form expression of the expected cross-product cannot be obtained in general and one could numerically approximate the quantity

$$\mathbb{E}[\mathsf{q}_n \mathsf{q}_{n'}] = \iint_{[0,1]^2} q q' f_{\mathsf{q}_n, \mathsf{q}_{n'}}(q, q') \mathrm{d}q \mathrm{d}q' \tag{E.39}$$

$$= \iint_{[0,1]^2} q q' f_{\mathsf{q}_n}(q) f_{\mathsf{q}_{n'}}(q')\, \mathrm{cop}(T_n^{-1}(q), T_{n'}^{-1}(q')) \mathrm{d}q \mathrm{d}q' \tag{E.40}$$

by monte-carlo integration

$$\mathbb{E}[\mathsf{q}_n \mathsf{q}_{n'}] \approx \frac{1}{S} \sum_{s=1}^{S} q_{n,s} q_{n',s} \tag{E.41}$$

where the sample pairs $\{(q_{n,s}, q_{n',s})\}_{s \in [S]}$ are generated based on Algo. E.1.

The correlation between any two states is then

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } s_n \perp s_{n'} \\ \frac{\mathbb{E}[\mathsf{q}_n \mathsf{q}_{n'}] - \frac{\alpha_n}{\alpha_n+\beta_n}\frac{\alpha_{n'}}{\alpha_{n'}+\beta_{n'}}}{\sqrt{\frac{\alpha_n}{\alpha_n+\beta_n}\frac{\beta_n}{\alpha_n+\beta_n}\frac{\alpha_{n'}}{\alpha_{n'}+\beta_{n'}}\frac{\beta_{n'}}{\alpha_{n'}+\beta_{n'}}}} & \text{otherwise} \end{cases} \tag{E.42}$$

Using the identity

$$\mathbb{E}[\mathsf{q}_n \mathsf{q}_{n'}] = \mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}] \sqrt{\mathbb{V}[\mathsf{q}_n]\,\mathbb{V}[\mathsf{q}_{n'}]} + \mathbb{E}[\mathsf{q}_n]\,\mathbb{E}[\mathsf{q}_{n'}] \tag{E.43}$$

$$= \mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}] \sqrt{\frac{\alpha_n\beta_n}{(\alpha_n+\beta_n)^2(\alpha_n+\beta_n+1)}\frac{\alpha_n\beta_{n'}}{(\alpha_{n'}+\beta_{n'})^2(\alpha_{n'}+\beta_{n'}+1)}} + \frac{\alpha_n}{\alpha_n+\beta_n}\frac{\alpha_{n'}}{\alpha_{n'}+\beta_{n'}} \tag{E.44}$$

one can also formulates the states correlation as

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } s_n \perp s_{n'} \\ \frac{\mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}]}{\sqrt{(\alpha_n+\beta_n+1)(\alpha_{n'}+\beta_{n'}+1)}} & \text{otherwise} \end{cases}. \tag{E.45}$$

This last formulation is interesting for two reasons.

First the state correlation will necessarily be smaller than the activity probability correlation since the denominator is strictly greater than 1 since

$$\begin{cases} \alpha_n & > 0 \\ \beta_n & > 0 \\ \alpha_{n'} & > 0 \\ \beta_{n'} & > 0 \end{cases} \Rightarrow \begin{cases} \alpha_n + \beta_n + 1 > 1 \\ \alpha_{n'} + \beta_{n'} + 1 > 1 \end{cases} \Rightarrow \sqrt{(\alpha_n+\beta_n+1)(\alpha_{n'}+\beta_{n'}+1)} > 1. \tag{E.46}$$

In the particular case $\alpha_n = \alpha_{n'}$ and $\alpha_n + \beta_n = 1$, the correlation $\mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}]$ is halved.

Second, it explicitly generalizes the one in Eq. D.32. Indeed when q_n and $\mathsf{q}_{n'}$ are fully correlated, i.e. when $\mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}] = 1$, and $(\alpha_n, \beta_n) = (\alpha_{n'}, \beta_{n'}) = (\alpha_g, \beta_g)$ there exist underlying c_n and $\mathsf{c}_{n'}$ such as they are also fully correlated so that they belong to a common group g. Hence the

states correlation coefficient then equals that of GHomA since we get

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \frac{\mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}]}{\sqrt{(\alpha_n + \beta_n + 1)(\alpha_{n'} + \beta_{n'} + 1)}} \tag{E.47}$$

$$= \frac{1}{\alpha_g + \beta_g + 1} \tag{E.48}$$

$$= \textit{Eq. D.32} \tag{E.49}$$

which shows the generalization.

Example with gaussian copula An example of correlation matrix of the activity states under beta activity probabilities correlated with a gaussian copula is given in Fig. E.3. The same parameters are shared across all the activity probabilities for the marginal beta distributions so that

$$\forall n \in [N],\ (\alpha_n, \beta_n) = (\alpha, \beta). \tag{E.50}$$

For a pair n, n' with $n \neq n'$, each correlation point in Fig. E.3 is computed using Eq. E.42 and Eq. E.41 with $S = 10^5$ samples for each possible pairs $(\alpha, \beta) \in [0, 2]^2$.

Consider another simpler example where the correlation matrix is

$$\boldsymbol{K_c} = \begin{bmatrix} 1.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.9 & 0.9 \\ 0.0 & 0.0 & 0.0 & 0.9 & 1.0 & 0.9 \\ 0.0 & 0.0 & 0.0 & 0.9 & 0.9 & 1.0 \end{bmatrix} \tag{E.51}$$

suggesting that the activity states are split into 3 groups. The first group has 2 states with correlation 0.5, the second group has only 1 state and the last group has 3 states with correlation 0.9.

An important observation is that the estimated state correlation matrix can be very different of $\boldsymbol{K_c}$, even when the original correlation coefficient are high. This is especially true when the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have high heterogeneity in their components. Indeed, even if two states s_n and $\mathsf{s}_{n'}$ are the result of a copula transform with correlation coefficient $\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}]$ close to 1 but with very different beta parameters, say $(\alpha_n, \beta_n) = (0.5\alpha_{n'}, 0.5\beta_{n'})$, the marginal distribution of the states will be very diffent,

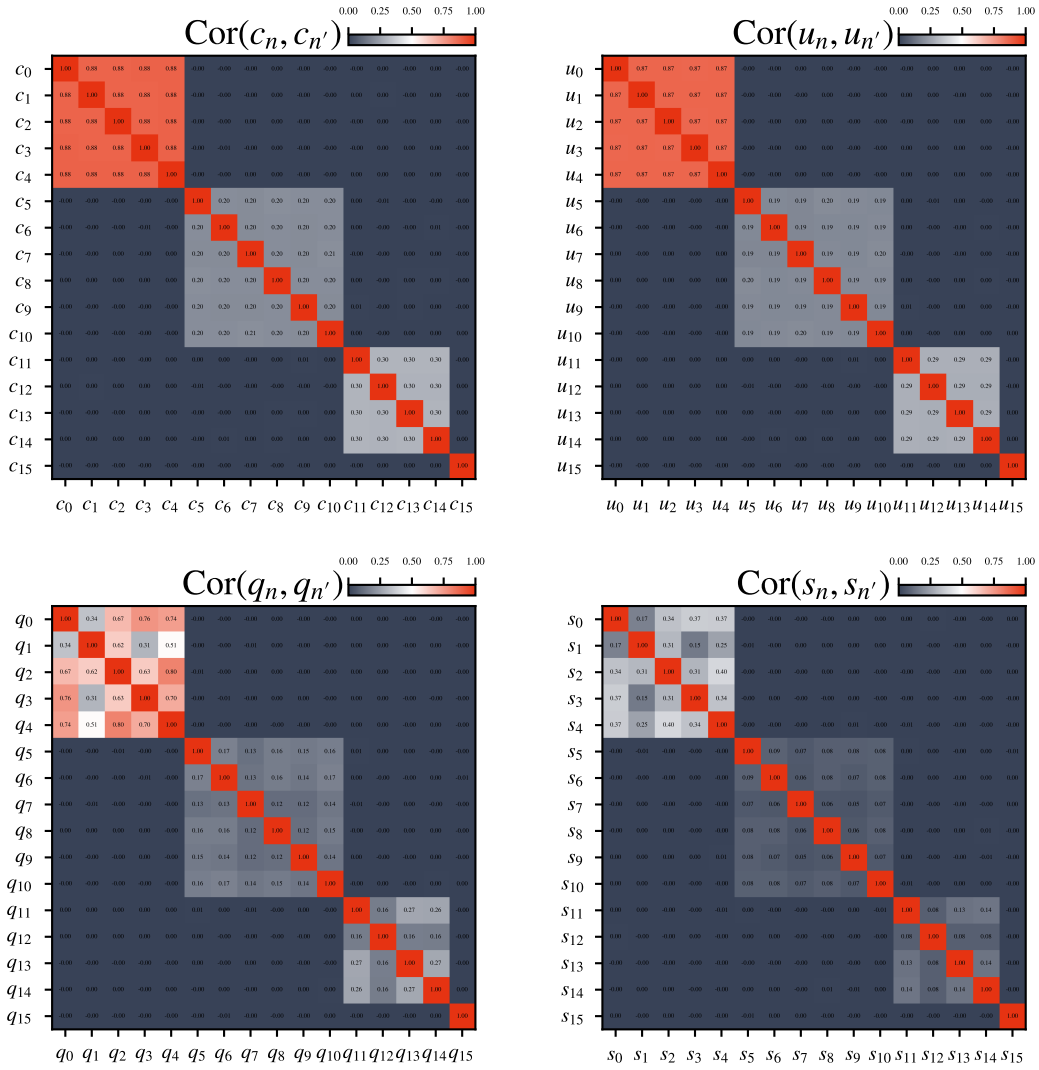


Figure E.3: Example of correlation matrices obtained through a copula transform from a correlated gaussian random vector **c** to a correlated state random vector **s**. Each correlation matrix exhibits a group structure with 4 blocks determined by the correlation matrix of the initial gaussian random vector. The correlation within each block is determined by the parameters of the successive transformations applied to each vector component.

decreasing the states correlation.

II.3 AUDaCE

The problem statement of joint AUDaCE with the transmission model of Sec. E.II.1 and the GHetA model of Sec. E.II.2 is the same as in Sec. D.II.3. The estimation of the channel matrix $\boldsymbol{H}$ and activity pattern $\boldsymbol{s}$ from an

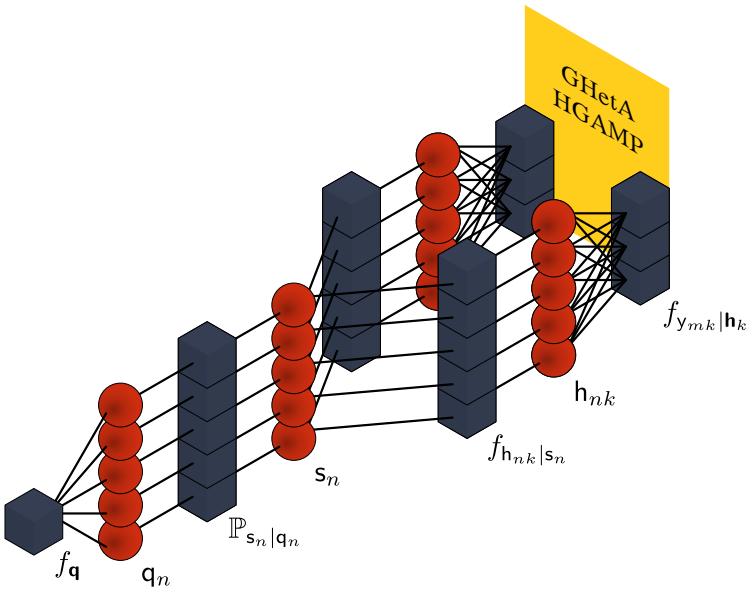


Figure E.4: Factor graph of the system presented in Sec. E.II with $N = 5$, $M = 3$ and $K = 2$. Different from the factor graph of GS-HGAMP and GHomA-HGAMP, this factor graph considers the correlation between the variables $\{q_n\}_{n\in[N]}$.

observed noisy signal $\boldsymbol{Y}$ is formulated as the joint MMSE estimator

$$[\hat{\boldsymbol{H}}, \hat{\boldsymbol{s}}] = \mathbb{E}[[\boldsymbol{H}, \boldsymbol{s}] \mid \mathbf{Y} = \boldsymbol{Y}]\,. \tag{E.52}$$

where the expectation expands to

$$\begin{aligned}&\mathbb{E}[[\boldsymbol{H}, \boldsymbol{s}] \mid \mathbf{Y} = \boldsymbol{Y}] \\ &= \frac{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \int_{[0,1]^N} [\boldsymbol{H}, \boldsymbol{s}] f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} \mathrm{d}\boldsymbol{H}}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \int_{[0,1]^G} f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q} \mathrm{d}\boldsymbol{H}}.\end{aligned} \tag{E.53}$$

We recall that the MMSE estimator is in general intractable so that we will derive, as in Sec. D.III, a LBP and corresponding HGAMP algorithms to address this issue and approximate the optimal estimates.

III Active user detection and channel estimation

III.1 Loopy belief propagation

The system's variables form the following Markov chain

$$\mathsf{q} \to \mathsf{s} \to \mathsf{H} \to \mathsf{Y} \tag{E.54}$$

which allows to write their posterior joint density as

$$f_{\mathsf{H},\mathsf{s},\mathsf{q}|\mathsf{Y}}(\boldsymbol{H}, \boldsymbol{s}, \boldsymbol{q} \mid \boldsymbol{Y}) = \frac{1}{f_{\mathsf{Y}}(\boldsymbol{Y})} f_{\mathsf{Y}|\mathsf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathsf{H}|\mathsf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) f_{\mathsf{s}|\mathsf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathsf{q}}(\boldsymbol{q}) \tag{E.55}$$

Based on Sec. E.II.1 and Sec. E.II.2, we can factorize the joint density's factors as

$$f_{\mathsf{Y}|\mathsf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) = \prod_{k\in[K]} \prod_{m\in[M]} f_{\mathsf{y}_{mk}|\mathsf{h}_k}(y_{mk} \mid \boldsymbol{h}_k) \tag{E.56}$$

$$f_{\mathsf{H}|\mathsf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) = \prod_{k\in[K]} \prod_{n\in[N]} f_{\mathsf{h}_{nk}|\mathsf{s}_n}(h_{nk} \mid s_n) \tag{E.57}$$

$$f_{\mathsf{s}|\mathsf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) = \prod_{n\in[N]} f_{\mathsf{s}_n|\mathsf{q}_n}(s_n \mid q_n) \tag{E.58}$$

and we keep, in the general case, $f_{\mathsf{q}}(\boldsymbol{q})$ unfactorized. This leads to the factor graph depicted in Fig. E.4. Note that an alternative factorization could be considered using the factorization of $f_{\mathsf{q}}(\boldsymbol{q})$ in Eq. E.23 which would change the factor graph by

- connecting the variable nodes $\{\mathsf{q}_n\}_{n\in[N]}$ respectively to the factor nodes $\{f_{\mathsf{q}_n}\}_{n\in[N]}$;
- connecting all the variable nodes $\{\mathsf{q}_n\}_{n\in[N]}$ to the factor node cop.

The messages of LBP based on Fig. E.4 are derived using the rules from Sec. C.II and are summarized in Tab. E.1. The posterior joint pdf can then be approximated by the product of the following beliefs:

$$\forall n \in [N], \underset{q_n}{\mathfrak{B}} = \underset{f_{\mathsf{q}} \to q_n \mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \to q_n}{\mathfrak{M}} \underset{}{\mathfrak{M}} \tag{E.59}$$

$$\forall s \in [N], \underset{s_n}{\mathfrak{B}} = \underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \to s_n \mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \to s_n}{\mathfrak{M}} \underset{}{\mathfrak{M}} \tag{E.60}$$

$$\forall (n,k) \in [N] \times [K], \underset{h_{nk}}{\mathfrak{B}} = \underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \to h_{nk}}{\mathfrak{M}} \prod_{m \in [M]} \underset{f_{\mathsf{y}_{mk}|\mathsf{h}_{:k}} \to h_{nk}}{\mathfrak{M}} \tag{E.61}$$

and then produce the estimates $\hat{\boldsymbol{q}}$, $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{H}}$ by computing them independently based on the beliefs:

$$\forall n \in [N], \ \hat{q}_n = \mathbb{E}\left[\mathsf{q}_n; \underset{q_n}{\mathfrak{B}}\right] \tag{E.62}$$

$$\forall n \in [N], \ \hat{s}_n = \mathbb{1}(\underset{s_n}{\mathfrak{B}}(0) < \underset{s_n}{\mathfrak{B}}(1)) \tag{E.63}$$

$$\forall (n,k) \in [N] \times [K], \ \hat{x}_{nk} = \mathbb{E}\left[\mathsf{h}_{nk}; \underset{h_{nk}}{\mathfrak{B}}\right] \tag{E.64}$$

III.2 GHetA-HGAMP algorithm

As for Sec. D.III.2, a major downside of LBP resides in the fact that it remains intractable since the messages that are exchanged are 1) densities and 2) numerous, especially in the dense part of the factor graph.

From Sec. C.III.2 and similar to Sec. D.III.2, the derivation of the HGAMP instance starts with the following approximation

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \leftarrow \mathsf{h}_{nk}}{\mathfrak{M}}(h_{nk}) \approx \mathcal{CN}(h_{nk}; r_{nk}, \tau_{r,nk}) \tag{E.65}$$

propagated in the LBP messages as it is described in the full derivation in Sec. G.V. If one compare the resulting GHetA-HGAMP algorithm Algo. E.2 to GHomA-HGAMP Algo. D.1, there are obvious similarities since the GAMP part is the same. We then focus on the differences that arise in the range of lines 21 – 29, i.e. the BP part for updating the activity-related variables.

1. Lines 22 and 23 remain the same with the computation of the activity likelihoods $\phi_{0,n}$ and $\phi_{1,n}$.
2. Line 24 consists in the estimation of a UE activity probability with

Factor	Variable	Factor → Variable / Factor ← Variable		
$f_{\mathsf{y}_{mk}\|\mathbf{h}_k}$	h_{nk}	$\mathfrak{M}_{f_{\mathsf{y}_{mk}\|\mathbf{h}_k}\to\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\int_{\mathbb{C}^{N-1}} f_{\mathsf{y}_{mk}\|\mathbf{h}_k}(y_{mk} \mid \boldsymbol{h}_{:k}) \left[\prod_{n'\in[N]\setminus\{n\}} \mathfrak{M}_{f_{\mathsf{y}_{m'k}\|\mathbf{h}_k}\leftarrow\mathsf{h}_{n'k}}(h_{n'k})\mathrm{d}h_{n'k}\right]$
		$\mathfrak{M}_{f_{\mathsf{y}_{mk}\|\mathbf{h}_k}\leftarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\to\mathsf{h}_{nk}}(h_{nk}) \prod_{m'\in[M]\setminus\{m\}} \mathfrak{M}_{f_{\mathsf{y}_{mk}\|\mathbf{h}_k}\to\mathsf{h}_{nk}}(h_{nk})$
$f_{\mathsf{h}_{nk}\|\mathsf{s}_n}$	h_{nk}	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\to\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\sum_{s=0}^{1} f_{\mathsf{h}_{nk}\|\mathsf{s}_n}(h_{nk} \mid s) \mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\leftarrow\mathsf{s}_n}(s)$
		$\mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\leftarrow\mathsf{h}_{nk}}(h_{nk})$	$\propto$	$\prod_{m=1}^{M} \mathfrak{M}_{f_{\mathsf{y}_{mk}\|\mathbf{h}_k}\to\mathsf{h}_{nk}}(h_{nk})$
$f_{\mathsf{h}_{nk}\|\mathsf{s}_n}$	s_n	$\mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\to\mathsf{s}_n}(s_n)$	$\propto$	$\int_{\mathbb{C}} f_{\mathsf{h}_{nk}\|\mathsf{s}_n}(h \mid s_n) \mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\leftarrow\mathsf{h}_{nk}}(h)\mathrm{d}h$
		$\mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\leftarrow\mathsf{s}_n}(s_n)$	$\propto$	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\to\mathsf{s}_n}(s_n) \prod_{k'\in[K]\setminus\{k\}} \mathfrak{M}_{f_{\mathsf{h}_{nk'}\|\mathsf{s}_n}\to\mathsf{s}_n}(s_n)$
$\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}$	s_n	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\to\mathsf{s}_n}(s_n)$	$\propto$	$\int_{0}^{1} \mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}(s_n \mid q) \mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\leftarrow\mathsf{q}_n}(q)\mathrm{d}q$
		$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\leftarrow\mathsf{s}_n}(s_n)$	$\propto$	$\prod_{k=1}^{K} \mathfrak{M}_{f_{\mathsf{h}_{nk}\|\mathsf{s}_n}\to\mathsf{s}_n}(s_n)$
$\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}$	q_n	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\to\mathsf{q}_n}(q_n)$	$\propto$	$\sum_{s=0}^{1} \mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}(s \mid q_n) \mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\leftarrow\mathsf{s}_n}(s)$
		$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\leftarrow\mathsf{q}_n}(q_n)$	$\propto$	$\mathfrak{M}_{f_{\mathbf{q}}\to\mathsf{q}_n}(q_n) \mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\to\mathsf{q}_n}(q_n)$
$f_{\mathbf{q}}$	q_n	$\mathfrak{M}_{f_{\mathbf{q}}\to\mathsf{q}_n}(q_n)$	$\propto$	$\int_{[0,1]^{N-1}} f_{\mathbf{q}}(q) \left[\prod_{n'\in[N]\setminus\{n\}} \mathfrak{M}_{f_{\mathbf{q}}\leftarrow\mathsf{q}_{n'}}(q_{n'})\mathrm{d}q_{n'}\right]$
		$\mathfrak{M}_{f_{\mathbf{q}}\leftarrow\mathsf{q}_n}(q_n)$	$\propto$	$\mathfrak{M}_{\mathbb{P}_{\mathsf{s}_n\|\mathsf{q}_n}\to\mathsf{q}_n}(q_n)$

Table E.1: Loopy belief propagation messages for the factor graph of Fig. E.4

the likelihood information corresponding to all the UEs except the nth one. It is different from GHomA-HGAMP since the latter estimates a group activity probability since all the UEs in a group share the same activity probability.

3. Line 25 still consists in the computation of individual likelihood-ratios that are used to perform state detection in line 26.

4. Line 27 is the update of an intermediate variable that will be used later in line 33 to estimate the channel coefficient. Since γ_n only depends on the corresponding individual log-likelihood ratio which itself depends on the individual estimate of the activity probability, the channel estimates can account for the information from all the UEs and not only the information from the UEs in a group. Each UE

Algorithm E.2 GHetA-HGAMP

Description: GHetA-HGAMP consists into two parts. The red lines corresponds to the updates of the GAMP variables for estimating the channel and the blue lines corresponds to the updates of the pattern variables with BP. Estimates of the system variables are colored in yellow.

input: $\boldsymbol{Y}, \boldsymbol{P}, \mu_{\mathsf{h}}, \tau_{\mathsf{h}}, \tau_{\mathsf{w}}, I_{\max}$
init:
  $i = 0$
  $\forall (n,k) \in [N] \times [K]\ \hat{h}^i_{nk} = \mu_{\mathsf{h}},\ \hat{\tau}^i_{\mathsf{h},nk} = \tau_{\mathsf{h}}$
  $\forall (m,k) \in [N] \times [K]\ \hat{u}^i_{mk} = 0$
end
for $i \in [I_{\max}]$ **do**:
  for $(m,k) \in [M] \times [K]$ **do**:
    $\hat{\tau}^i_{\mathsf{p},mk} = \sum_{n=1}^N |p_{mn}|^2 \hat{\tau}^{i-1}_{\mathsf{h},nk}$
    $\hat{p}^i_{mk} = \sum_{n=1}^N p_{mn} \hat{h}^{i-1}_{nk} - \hat{\tau}^i_{\mathsf{p},mk} \hat{u}^{i-1}_{mk}$
    $\hat{\tau}^i_{\mathsf{z},mk} = \tau_{\mathsf{w}} \hat{\tau}^i_{\mathsf{p},mk} / (\hat{\tau}^i_{\mathsf{p},mk} + \tau_{\mathsf{w}})$
    $\hat{z}^i_{mk} = \hat{p}^i_{mk} + \hat{\tau}^i_{\mathsf{p},mk} (y_{mk} - \hat{p}^i_{mk}) / (\hat{\tau}^i_{\mathsf{p},mk} + \tau_{\mathsf{w}})$
    $\hat{\tau}^i_{\mathsf{u},mk} = (1 - \hat{\tau}^i_{\mathsf{z},mk}) / (\hat{\tau}^i_{\mathsf{p},mk})^2$
    $\hat{u}^i_{mk} = (\hat{z}^i_{mk} - \hat{p}^i_{mk}) / \hat{\tau}^i_{\mathsf{p},mk}$
  end
  for $(n,k) \in [N] \times [K]$ **do**:
    $\hat{\tau}^i_{\mathsf{r},nk} = (\sum_{m=1}^M |p_{mn}|^2 \hat{\tau}^i_{\mathsf{u},mk})^{-1}$
    $\hat{r}^i_{nk} = \hat{h}^{i-1}_{nk} + \hat{\tau}^i_{\mathsf{r},nk} \sum_{m=1}^M p_{mn} \hat{u}^i_{mk}$
  end
  for $n \in [N]$ **do**:
    $\phi_{0,n} = \prod_{k=1}^K \mathcal{CN}(0; \hat{r}^i_{nk}, \hat{\tau}^i_{\mathsf{r},nk})$
    $\phi_{1,n} = \prod_{k=1}^K \mathcal{CN}(0; \hat{r}^i_{nk} - \mu_{\mathsf{h}}, \hat{\tau}^i_{\mathsf{r},nk} + \tau_{\mathsf{h}})$
    $\hat{q}^i_{n,n} = \frac{\int_{[0,1]^N} q_n f_{\mathbf{q}}(\boldsymbol{q}) \prod_{n' \in [N] \setminus \{n\}} [(1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}] \mathrm{d}q_n}{\int_{[0,1]^N} f_{\mathbf{q}}(\boldsymbol{q}) \prod_{n' \in [N] \setminus \{n\}} [(1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}] \mathrm{d}q_n}$
    $\mathsf{LLR}_n = \log\left(\frac{\hat{q}^i_{n,g}}{(1-\hat{q}^i_{n,g})} \frac{\phi_{1,n}}{\phi_{0,n}}\right)$
    $\hat{s}^i_n = \mathbb{1}(\mathsf{LLR}_n > 0)$
    $\gamma_n = (1 + \exp(-\mathsf{LLR}_n))^{-1}$
    $\hat{q}^i_n = \frac{\int_{[0,1]^N} q_n f_{\mathbf{q}}(\boldsymbol{q}) \prod_{n' \in [N]} [(1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}] \mathrm{d}q_n}{\int_{[0,1]^N} f_{\mathbf{q}}(\boldsymbol{q}) \prod_{n' \in [N]} [(1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}] \mathrm{d}q_n}$
  end
  for $(n,k) \in [N] \times [K]$ **do**:
    $\kappa_{nk} = (1/\tau_{\mathsf{h}} + 1/\hat{\tau}^i_{\mathsf{r},nk})^{-1}$
    $\nu_{nk} = \mu_{\mathsf{h}}/\tau_{\mathsf{h}} + \hat{r}^i_{nk}/\hat{\tau}^i_{\mathsf{r},nk}$
    $\hat{h}^i_{nk} = \gamma_n \kappa_{nk} \nu_{nk}$
    $\hat{\tau}^i_{\mathsf{h},nk} = \gamma_n(\kappa_{nk} + |\kappa_{nk}\nu_{nk}|^2) - \left|\hat{h}^i_{nk}\right|^2$
  end
end

activity probability benefiting from its own statistical distribution, the heterogeneity is incorporated into the channel estimates.

5. Line 28 is similar to line 25 but the information of the nth UE is included to perform an estimate of the activity probability.

III.3 Complexity analysis

A short analysis is given for the computational complexity. The complexity from the GAMP part remains the same and still require $O(NMK)$ operations. The main difference once again appears in the BP part where the computational bottleneck is risen by the estimates $\hat{q}^i_{n,n}$ and $\hat{q}^i_n$. Different from GHomA-HGAMP, each estimate requires the computation of a multidimensional integral where the integration is over the mute vector

variable $\boldsymbol{q}$. The joint pdf put aside, the integrand is a multivariate polynomial which constitutes the major difference, since the moment-based trick that was used before cannot be used anymore.

Hopefully, the Monte-Carlo method is still valid so that it will be used to approximate the 4 integrals, using S vector samples $\{\boldsymbol{q}_j\}_{j\in[S]}$ that can be sampled offline before. Note that the samples must be obtained using Algo. E.1, so that they are properly correlated which allow to approximate the estimates by

$$\hat{q}_{n,n} \approx \frac{\sum_{j\in[S]} q_{n,j} \left[\prod_{n'\in[N]\setminus\{n\}} \left((1-q_{n',j})\phi_{0,n'} + q_{n',j}\phi_{1,n'}\right)\right]}{\sum_{j\in[S]} \left[\prod_{n'\in[N]\setminus\{n\}} \left((1-q_{n',j})\phi_{0,n'} + q_{n',j}\phi_{1,n'}\right)\right]} \tag{E.66a}$$

$$\hat{q}_{n} \approx \frac{\sum_{j\in[S]} q_{n,j} \left[\prod_{n'\in[N]} \left((1-q_{n',j})\phi_{0,n'} + q_{n',j}\phi_{1,n'}\right)\right]}{\sum_{j\in[S]} \left[\prod_{n'\in[N]} \left((1-q_{n',j})\phi_{0,n'} + q_{n',j}\phi_{1,n'}\right)\right]} \tag{E.66b}$$

leading to a computational complexity of the order of $O(NS)$. As a consequence, the overall complexity of GHetA-HGAMP is

$$O(NMK + NK + NS) \qquad \text{(Monte-Carlo)} \tag{E.67}$$

where the blue term is for GAMP and the red terms are for BP.

IV Numerical results

In this section, the performance of the proposed GHetA-HGAMP algorithm is studied and compared to other GAMP-based algorithms.

IV.1 Framework

In particular, the numerical study compares GHetA-HGAMP to the algorithms from Sec. D.IV.1 whose the corresponding factor graphs are shown in Fig. E.5.

Modified GAMP This modified GAMP (see Sec. C.III.2) differs from the one in Sec. D.IV.1 where the weights $\{q_n\}_{n\in[N]} \in [0,1]^N$ are identified to the parameters $\{\alpha_n\}_{n\in[N]}$ of the beta distributions, still assuming that they

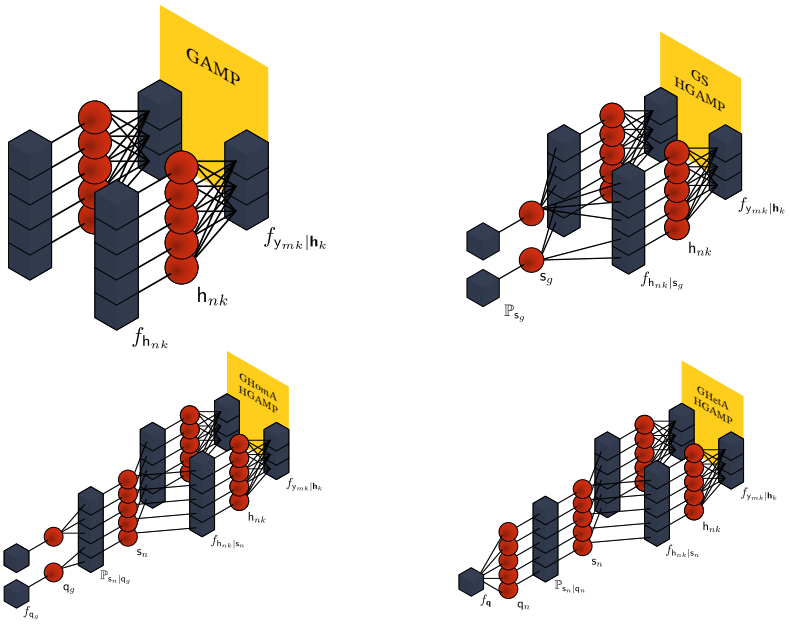


Figure E.5: Example of the underlying factor graphs of the modified GAMP, GS-HGAMP, GHomA-HGAMP and GHetA-HGAMP.

verify

$$\alpha_n + \beta_n = 1. \tag{E.68}$$

The log-likelihood ratios for this GAMP become

$$\mathsf{LLR}_n = \log \alpha_n + \log \phi_{0,n} - \log \beta_n - \log \phi_{1,n} \tag{E.69}$$

that will be used for activity detection. Again, this modified GAMP does not account for any activity correlation.

GS-HGAMP For the simulations, the activity probability of the group g will be identified to

$$q_g = \frac{1}{U_g} \sum_{n=\sum_{g'<g} U_{g'}+1}^{U_g} \alpha_n \tag{E.70}$$

which consists in computing the average of $\{\alpha_n\}_n$ as a proxy for the group activity probability

Parameter	Description	Value
I_{MC}	# of Monte-Carlo iterations	1000
$I_{\max}$	# of GAMP or HGAMP iterations	100
G	# of groups	64
U	# of UEs per group	4
N	# of UEs	256
M	preamble length	128
K	# of antennas	2
α	Beta distribution's parameters	0.35
β		$1 - 0.35 = 0.65$
ρ	correlation within each group with gaussian copula Sec. E.II.2	$\{0.25, 0.50, 0.75\}$

Table E.2: Simulation parameters for the comparative study of GHetA against modified GAMP, GS-HGAMP and GHomA-HGAMP

GHomA-HGAMP The activity probability of UEs belonging to the same group is assumed to be the same and defined as for GS-HGAMP

$$\tilde{\alpha}_g = \frac{1}{U_g} \sum_{n=\sum_{g'<g} U_{g'}+1}^{U_g} \alpha_n \tag{E.71}$$

for the simulations and leads to the state correlation of Eq. D.32.

GHetA-HGAMP It is the algorithm introduced in this chapter. Different from GAMP, GS-HGAMP and GHomA-HGAMP, it can deal with more general dependence and correlation structures, still introduced at the level of the activity probability that affects the state correlation as explained in Sec. E.II.2.

IV.2 Relative performances of GHetA

The first numerical study focuses on the relative performances of GHetA-HGAMP against modified GAMP, GS-HGAMP and GHomA-HGAMP. The performances of each algorithm (NMSE, FAR, MDR and UER) are assessed through Monte-Carlo simulations.

The simulation parameters are given in Tab. E.2 and the corresponding results have been drawn in Figs. E.6 to E.9.

Analysis The simulation model is that of GHetA, the important indicator is the correlation coefficient ρ that fully describes the correlation matrix

$\boldsymbol{K_c}$ of a gaussian copula. The latter is chosen to be block diagonal

$$\boldsymbol{K_c} = \boldsymbol{I}_G \otimes (\rho \mathbf{1}_{U\times U} + (1-\rho \boldsymbol{I}_U)) \tag{E.72}$$

When ρ is close to 0 the matrix is close to $\boldsymbol{I}_N$ and when ρ is close to 1 the matrix is block diagonal with 1-only blocks. Hence, three correlation regimes, low $\rho = 0.25$, moderate $\rho = 0.75$ and high $\rho = 0.75$, have been considered to assess the performances, for which we give a review hereafter.

- In all the correlation regimes, GS-HGAMP always performs the worst and GHetA-HGAMP always performs the best.
- In the low correlation regime, the activity of the UEs is close to be independent ($\rho = 0$) so that the modified GAMP algorithm is as good as GHetA-HGAMP over the whole SNR range. GHomA-HGAMP performs by a factor of ~ 2.5dB (cf. subfigure (a) in Fig. E.6) below GAMP and GHetA-HGAMP. The reason is that GHomA-HGAMP still relies on the underlying group structure of the UEs, which is responsible for a higher number of false alarms (subfigure (a) in Fig. E.8) and missed detections (subfigure (a) in Fig. E.9) for GHomA-HGAMP, leading to an overall worse UER (subfigure (a) in Fig. E.7). Indeed, when the activity correlation is low, the activity behavior of UEs within the same group is very unlikely to be the same.
- In the moderate correlation regime, all the algorithms, except GS-HGAMP, perform similarly for all the metrics (see subfigure (b) in Figs. E.6 to E.9)
- In the high correlation regime, the trend observed in the low correlation regime is reversed. Both GHomA and GHetA-HGAMP have the best performance, making a 4dB gap with the modified GAMP algorithm in subfigure (c) Fig. E.6. GAMP cannot leverage the correlation structure which is favorable to similar activity probabilities within each group of UEs to improve its detection capability (see subfigure (c) in Figs. E.7 to E.9)

The reasons justifying these performances are the same as those in Sec. D.IV.4.a. An additional remark is that GHetA-HGAMP is able to perform the best in all scenarios where GHomA-HGAMP and GAMP only perform best when the correlation regime respectively approaches the high

Parameter	Description	Value
I_{MC}	# of Monte-Carlo iterations	1000
$I_{\max}$	# of HGAMP iterations	100
G	# of groups	64
U	# of UEs per group	4
N	# of UEs	256
M	preamble length	128
K	# of antennas	2
α	Beta distribution's parameters	0.35
β		$1 - 0.35 = 0.65$
ρ	correlation within each group	$\{0.25, 0.75\}$
$\tilde{\rho}$	biased correlation given to GHetA	$\{0.0, 0.5, 1.0\}$

Table E.3: Simulation parameters for the robustness study of GHetA to biased correlation

and low correlation ones. The justification is that GHetA-HGAMP leverages the inner correlation structure when computing the activity probability estimates. The consequence is an improvement in the estimation of both the states and the channel coefficients since the log-likelihood ratios are improved.

IV.3 Robustness of GHetA to biased correlation

Another interesting aspect to study is then the robustness of GHetA-HGAMP to mismatched or biased correlation coefficient. Simulation parameters are described in Tab. E.3 and performance results are given in Figs. E.10 and E.12 for the NMSE and in Figs. E.11 and E.13 for the UER.

Analysis From Figs. E.10 and E.11, it can be seen that overestimating a rather weak correlation of the underlying correlated gaussian random vector seems to have little effects on the estimation and detection. The mismatched correlation affects the sampling process used for the Monte-Carlo integrations of Eq. E.66. If the true correlation is low, realizations of the components of the activity probability random vector $\mathbf{q}$ are likely to not be similar. However, if a mismatched high correlation is used, the sampled vectors will have similar values per group in the considered symmetric scenario. Then if the numbers of group is low, and so the number of UEs per group is high, the approximations will simply involve sums with little diversity in their terms; but in the considered scenario,

the size of the groups is small compared to the total number of UEs so that this loss in sampling diversity does not matters. Hence, systems with small groups where the activity states are likely to be weakly correlated do not require a precise estimate of the correlation for GHetA-HGAMP.

Although it can been seen in Figs. E.11 and E.12 that an accurate estimation of the correlation does not lead to significant gains in the cases $(0.25, \tilde{\rho})$ for $\tilde{\rho} \in \{0.25, 0.50, 0.75, 1\}$, from Figs. E.12 and E.13, underestimating a strong correlation however leads to degraded performance.

1. From Fig. E.13, one can conclude that the less biased is the correlation given to the algorithm, the better are the detection performances. An order of magnitude is gained between the cases $(0.75, 0)$ and $(0.75, 0.75)$ for the UER.
2. As a consequence, such additionnal errors account significantly in terms of NMSE with a gain of about 5dB in favor of the unbiased case compared to the biased case $(0.75, 0.25)$.
3. The convergence rate of GHetA-HGAMP is similar in all scenarios.

The explanation for this is very similar to that of the bad performance of GS-HGAMPs, i.e. that GHetAs-HGAMPs does not use the group structure information when the given biased correlation is close to zero, leading to bad estimates of the activity probability that affects the states detection and channel coefficients estimation.

V Conclusion

This chapter has studied the question raised in Sec. D.V regarding the extension of GHomA to GHetA activty pattern in the context of GFRA. To do so, the theory of copula has been shown to be helpful, allowing to consider very general dependence structures at the level of the latent activity probability random variables, one for each UE. Thanks to it, it has been explicitly shown that GHetA can generalize GHomA in some scenarios.

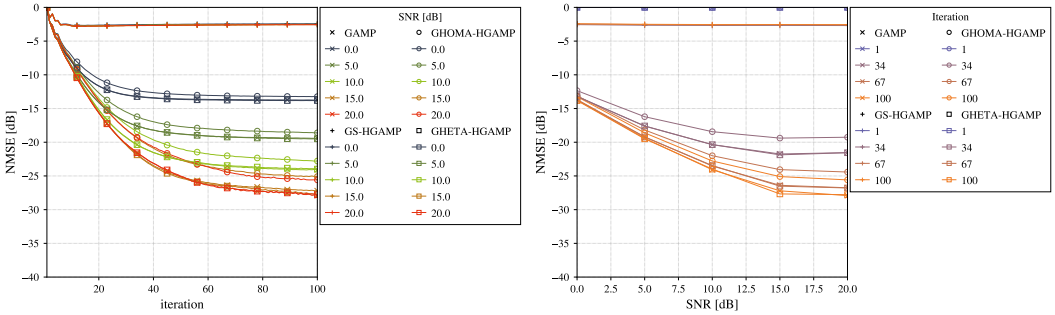
The AUDaCE problem has then been studied for GHetA where the same systematic LBP approach was first considered before being approximated thanks to HGAMP. The resulting GHetA-HGAMP algorithm has been numerically evaluated, showing significant improvements against

GHomA-HGAMP, GS-HGAMP and a modified version of GAMP. In particular, when GHomA-HGAMP and GAMP respectively perform well in the high and weak correlation regimes, GHetA-HGAMP perfoms well in both regimes, with an average channel estimation gain around 3dB and with halved FAR, MDR and UER in the high SNR regime. Furthermore, GHetA-HGAMP has been tested against mismatched correlation coefficient and proved to be robust when the true activity correlation is low. In the high correlation regime, it is however important to provide an accurate value of the correlation since a gain in the channel estimation of up to 5dB and an order of magnitude in the UER can be achieved w.r.t to an assumed mismatched zero correlation.

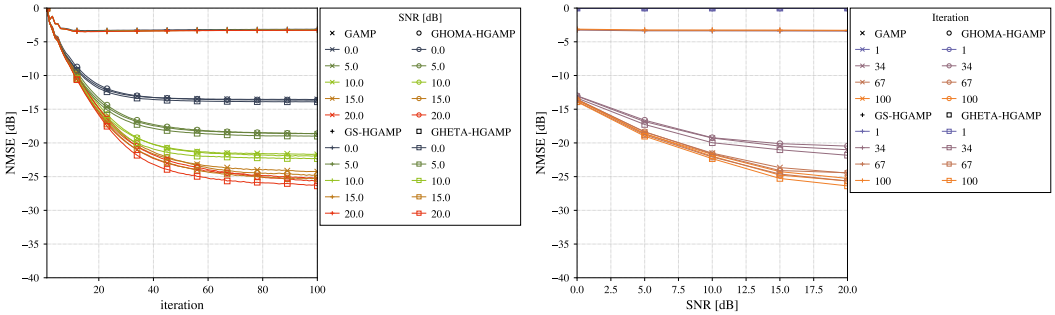
For future work, one can consider two different levels of improvement. The first is from the point of view of algorithmic complexity with the development of fast evaluation techniques for the computation of the BP updates in Algo. E.2 which can significantly slow down the iterative estimation process. The second aspect is the exploration of relevant statistical dependence structures induced by the copula so that they precisely fit RA scenarios in order to improve AUDaCE performances.

Perspectives on applying this work are given in Chap. F, with a focus on multi-carrier OFDM systems and grant-free data transmissions.

(a) $\rho = 0.25$



(b) $\rho = 0.50$



(c) $\rho = 0.75$

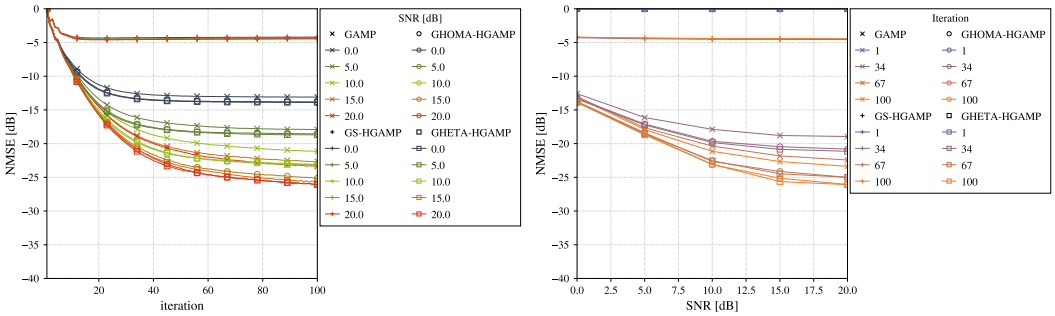
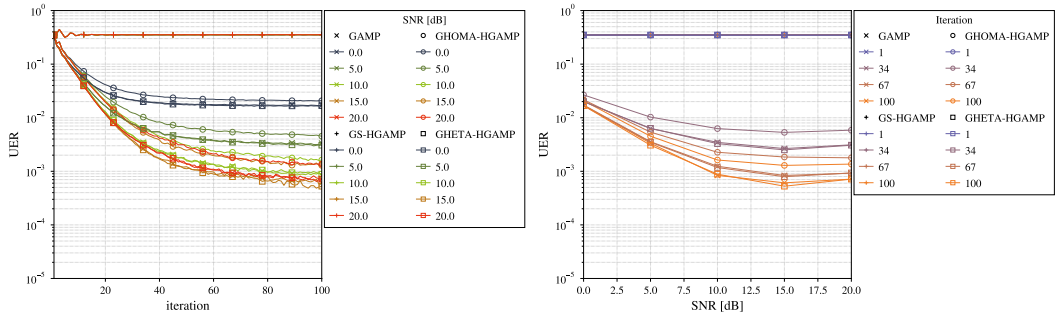
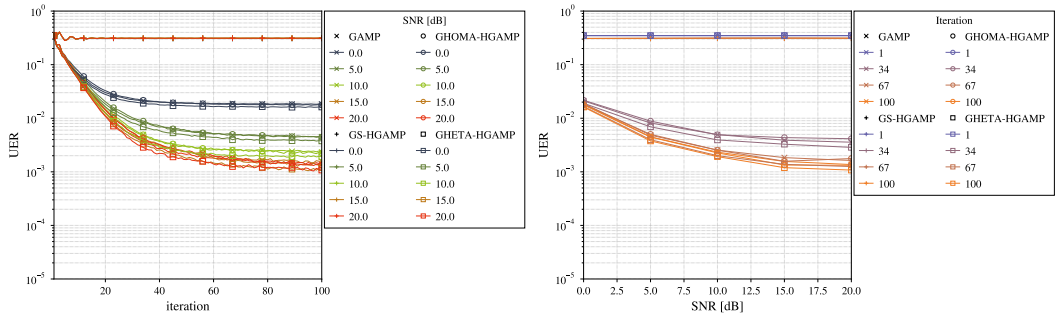


Figure E.6: NMSE for the channel estimation of GHetA-HGAMP against different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

(a) $\rho = 0.25$



(b) $\rho = 0.50$



(c) $\rho = 0.75$

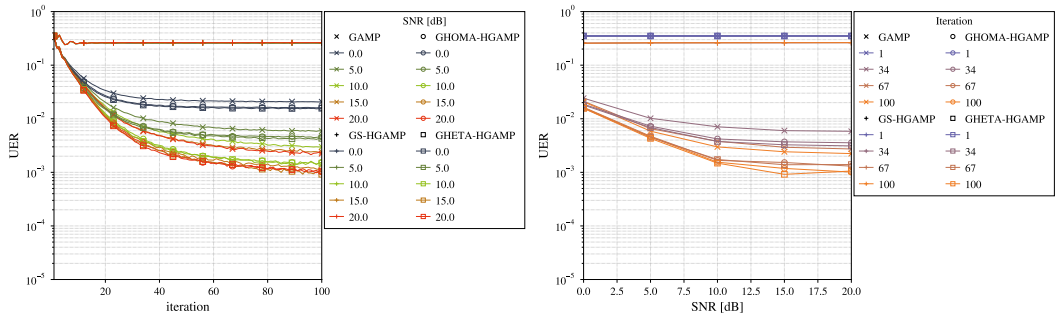
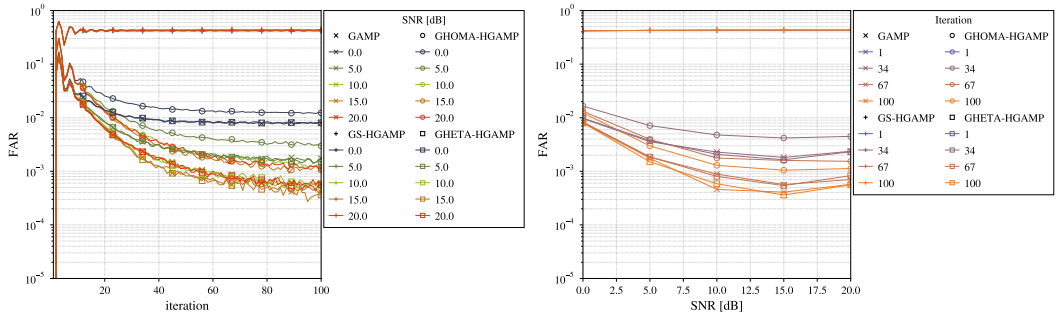
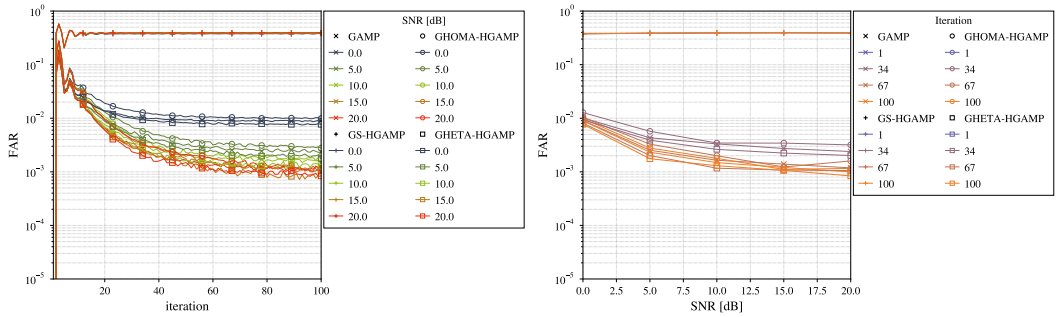


Figure E.7: UER for the activity detection of different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

(a) $\rho = 0.25$



(b) $\rho = 0.50$



(c) $\rho = 0.75$

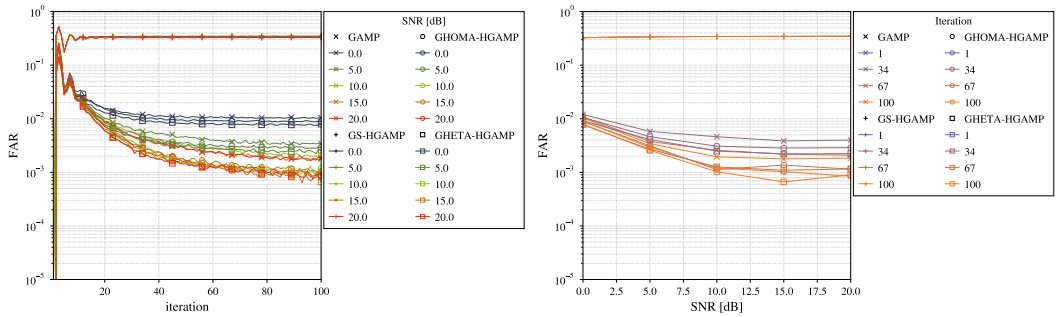
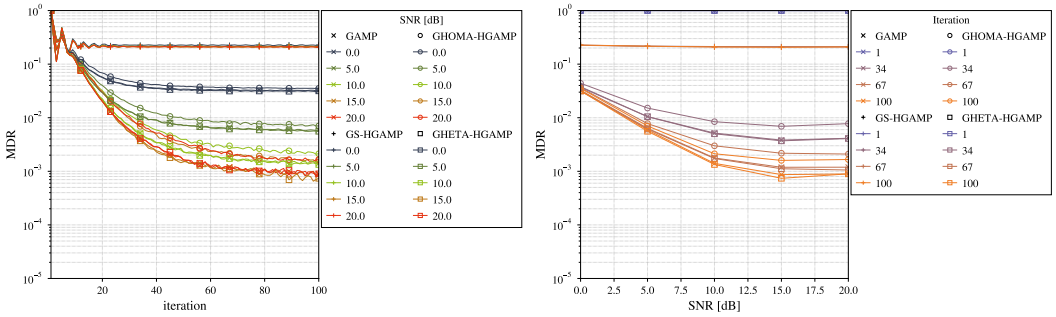
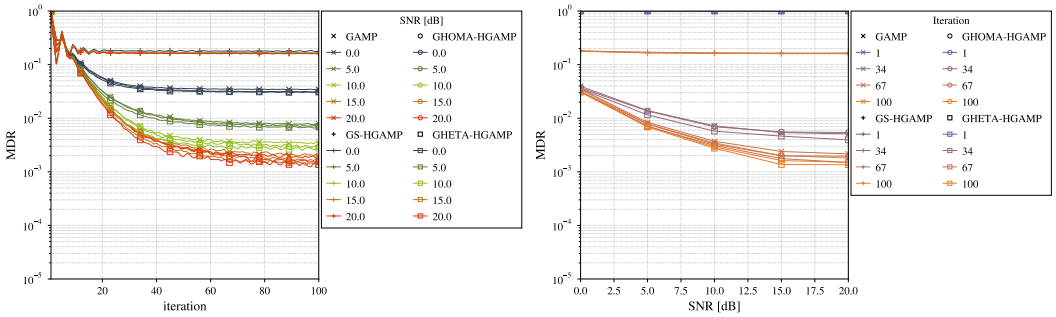


Figure E.8: FAR for the activity detection of different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

(a) $\rho = 0.25$



(b) $\rho = 0.50$



(c) $\rho = 0.75$

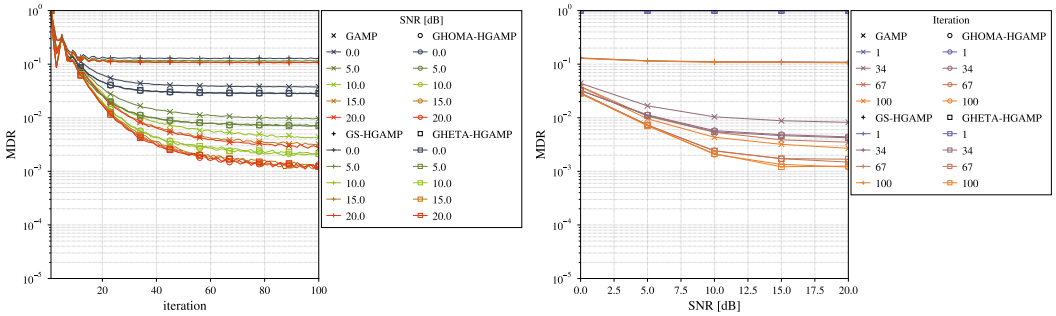
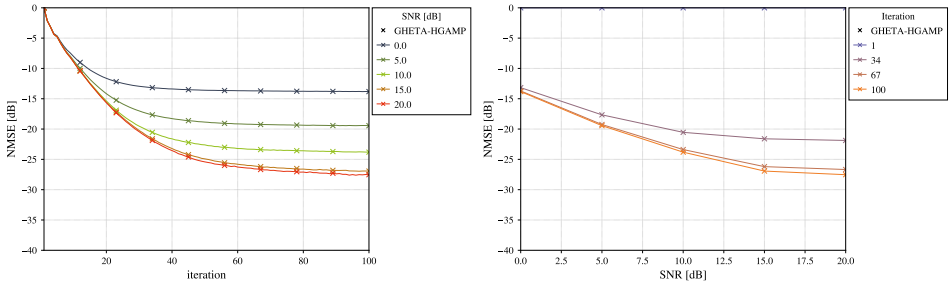
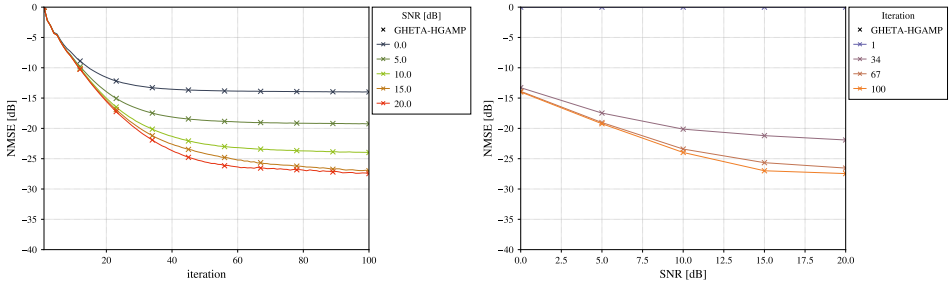


Figure E.9: MDR for the activity detection of different GAMP-based algorithms w.r.t. to the number of iterations and the SNR.

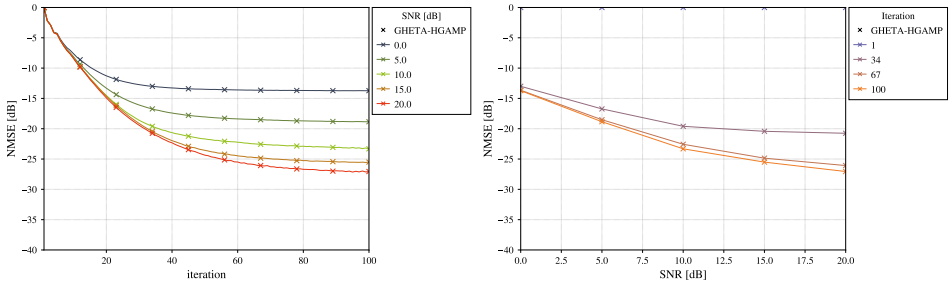
(a) $\rho = 0.25, \tilde{\rho} = 0.25$



(b) $\rho = 0.25, \tilde{\rho} = 0.50$



(c) $\rho = 0.25, \tilde{\rho} = 0.75$



(d) $\rho = 0.25, \tilde{\rho} = 1.00$

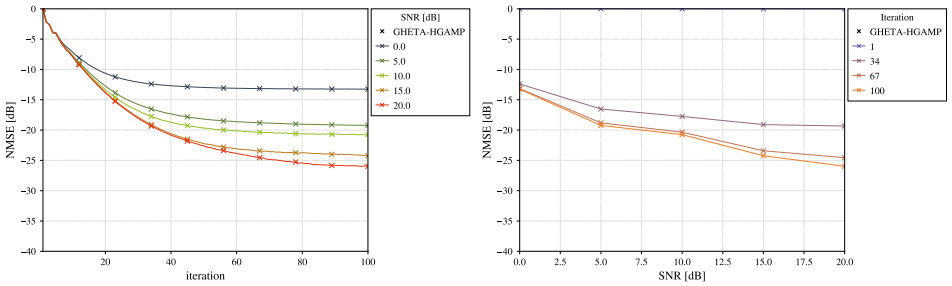
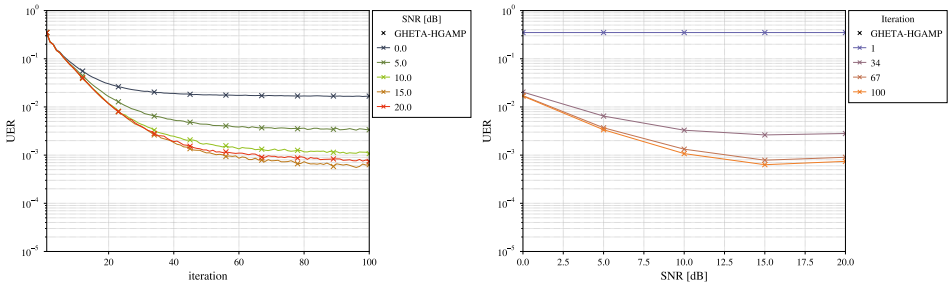
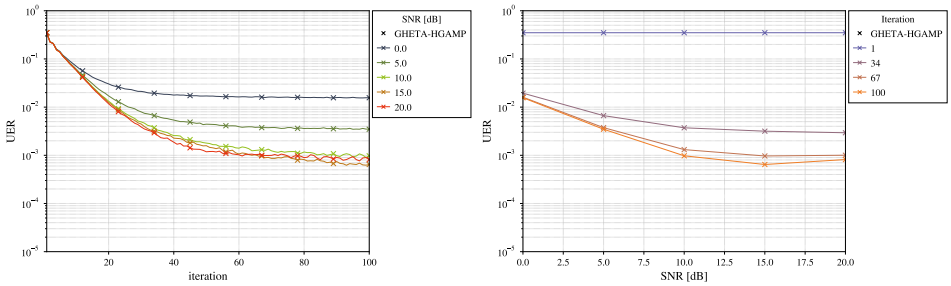


Figure E.10: NMSE for the channel estimation of GHetA-HGAMP with biased correlation.

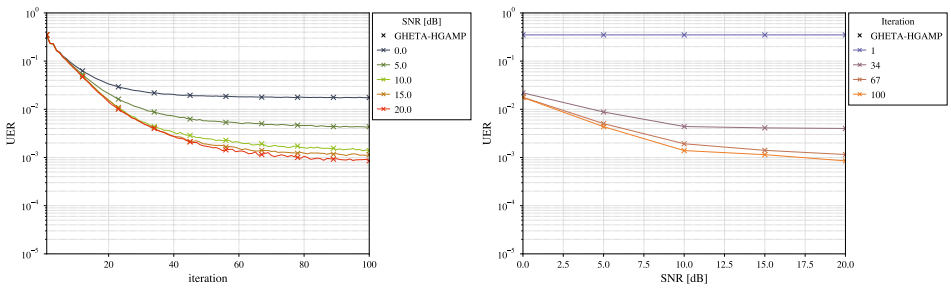
(a) $\rho = 0.25, \tilde{\rho} = 0.25$



(b) $\rho = 0.25, \tilde{\rho} = 0.50$



(c) $\rho = 0.25, \tilde{\rho} = 0.75$



(d) $\rho = 0.25, \tilde{\rho} = 1.00$

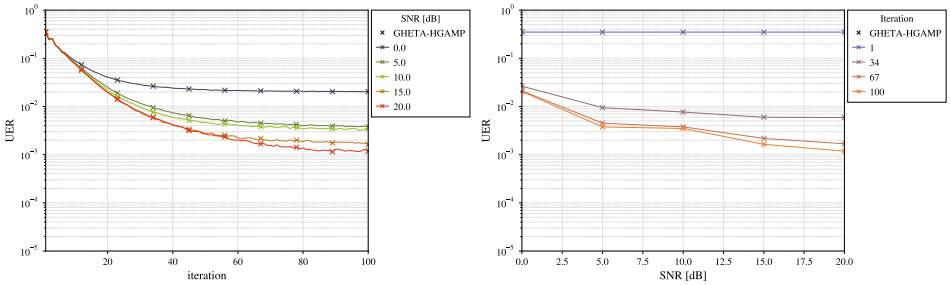
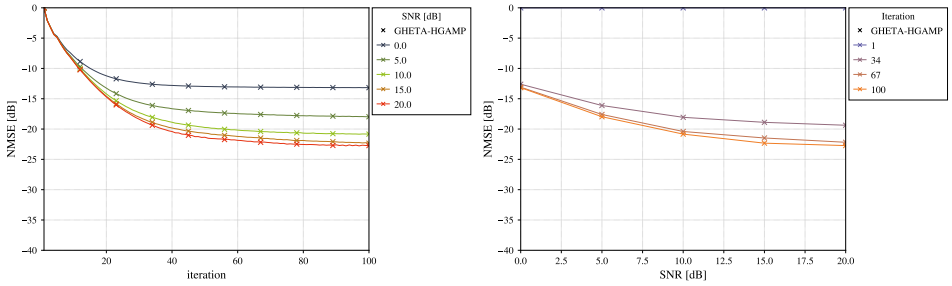
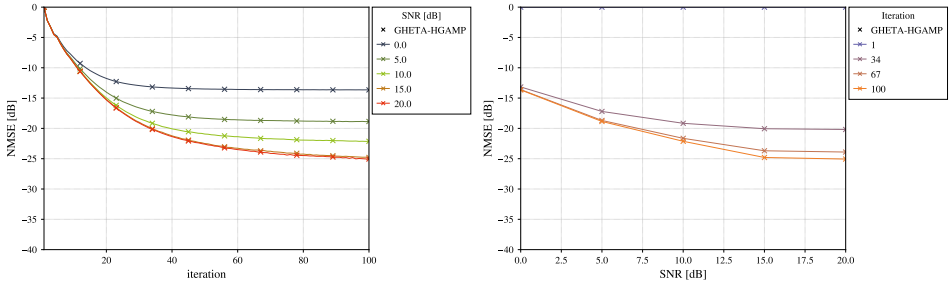


Figure E.11: UER for the activity detection of GHetA-HGAMP with biased correlation.

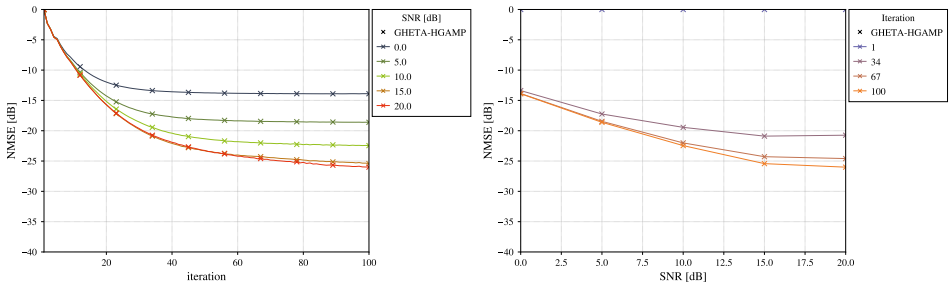
(a) $\rho = 0.75, \tilde{\rho} = 0.00$



(b) $\rho = 0.75, \tilde{\rho} = 0.25$



(c) $\rho = 0.75, \tilde{\rho} = 0.50$



(d) $\rho = 0.75, \tilde{\rho} = 0.75$

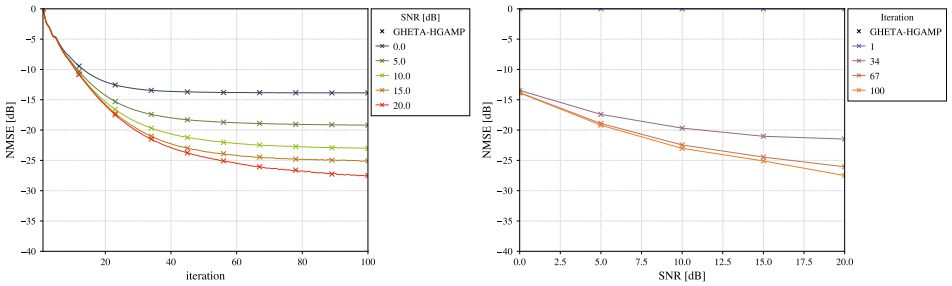
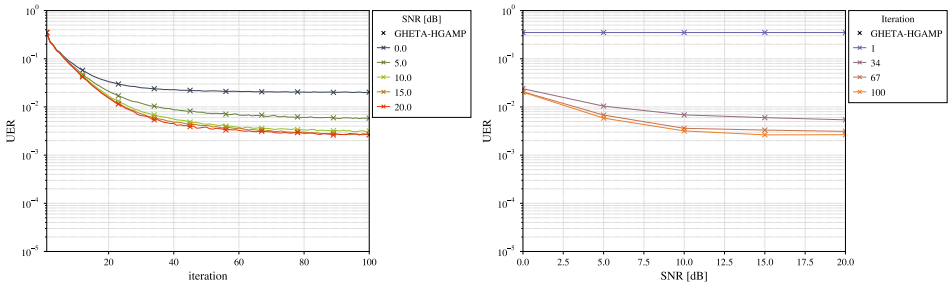
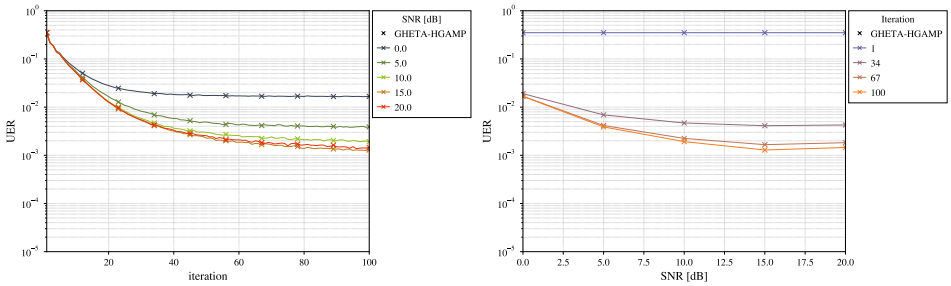


Figure E.12: NMSE for the channel estimation of GHetA-HGAMP with biased correlation.

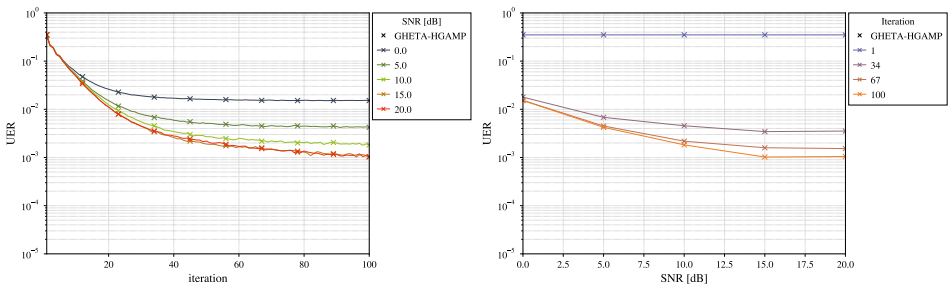
(a) $\rho = 0.75, \tilde{\rho} = 0.00$



(b) $\rho = 0.75, \tilde{\rho} = 0.25$



(c) $\rho = 0.75, \tilde{\rho} = 0.50$



(d) $\rho = 0.75, \tilde{\rho} = 0.75$

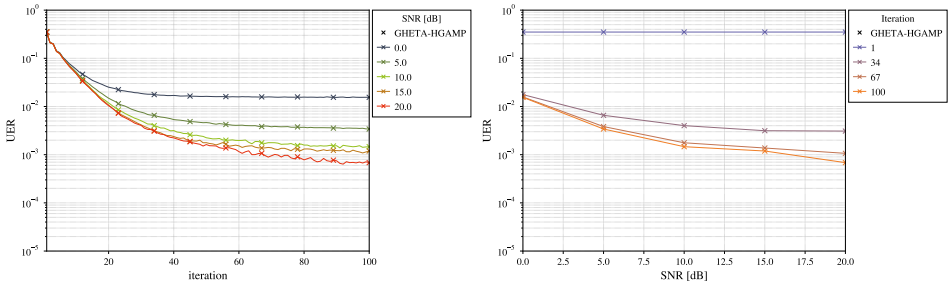


Figure E.13: UER for the activity detection of GHetA-HGAMP with biased correlation.

F

CONCLUSION

I THESIS OUTCOMES

This thesis has studied the problem of GFRA in the context of mMTC and uRLLC within 5G-NR. It was discussed the importance of the communication systems to leverage AUDaCE for enabling GFRA. Using CS-NOMA, it is possible to deal efficiently with the structure of the system variables involved in AUDaCE. In particular, it was shown the capability of the HGAMP bayesian framework to cope with underlying correlated activity of UEs.

First, a GHomA pattern model has been considered in Chap. D. It has been shown to generalize the independent group-sparsity activity with the introduction of latent group variables identified with the activity probabilities of each UEs within the groups. Based on this new model, a LBP has been derived before being approximated by an instance of HGAMP, namely GHomA-HGAMP. The latter has numerically demonstrated its interest for AUDaCE in some activity regimes against other GAMP-based algorithms that do not account for the correlation induced by GHomA.

Second, a GHetA pattern model generalizing the GHomA pattern model has been introduced. The generalization is carried by the copula theory so that general dependence structures between the UEs' activity

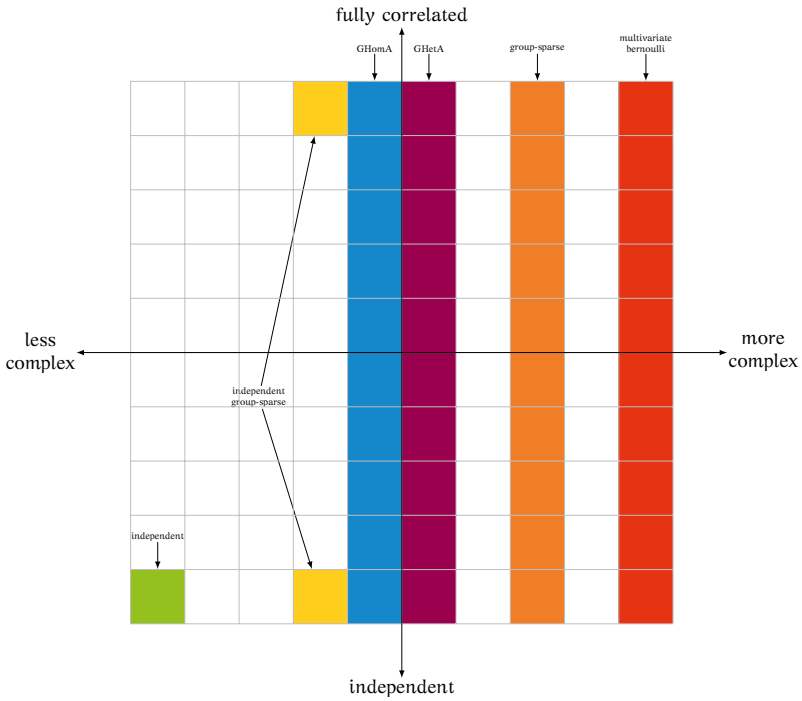


Figure F.1: Qualitative comparison of activity pattern models, updated with GHomA and GHetA. The location of GHomA on the complexity/-correlation axes is justified by the fact it only adds few latent variables, each being described with a simple distribution. For GHetA, latent variables are also added but they depend on a general copula structure for which the complexity highly depends on the underlying dependence structure. In the case where the dependence structure is that of a correlated gaussian random vector, it only requires a correlation matrix and the parameters of the targeted distribution (e.g. beta).

states can be built. With a similar approach to GHomA, instances of LBP and HGAMP have been developed, leading to GHetA-HGAMP. The algorithm has shown its capability to deal efficiently with AUDaCE under different scenarios of activivty correlation, whereas the algorithms GHomA-HGAMP and modified GAMP have proved to be efficient only in the cases of strong and weak activity correlation respectively.

Among other directions that could be considered, we introduce and give in Sections F.II and F.III some insights of interesting future works that could be done building on the work of this thesis.

II Extension to multi-carrier OFDM

In Chapters D and E, the UL transmissions were performed on a single carrier only. In order to extend the work of AUDaCE to an OFDM scheme where the transmissions are conveyed over $F \in \mathbb{N}_*$ frequencies, one can consider a system where, at each time slot, the UEs are allowed to transmit on subsets of this frequencies.

Formally, for the nth UE, denote by $\boldsymbol{b}_n \in \{0,1\}^F$ the binary vector so that

$$\forall f \in [F],\ b_{fn} = \begin{cases} 0 & \text{if the UE is not allowed to transmit on frequency } f \\ 1 & \text{otherwise} \end{cases} \tag{F.1}$$

and form the allocated subbands matrix $\boldsymbol{B}$ such that

$$\boldsymbol{B} = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_N] \in \{0,1\}^{F \times N}. \tag{F.2}$$

At each time slot, the UE n transmits on the frequency f if and only if

$$b_{fn} = 1 \qquad \text{and} \qquad \mathsf{s}_n = 1. \tag{F.3}$$

Assume that the number of antennas is 1 and let $\mathbf{H} \in \mathbb{C}^{N \times F}$ denote the random channel matrix between the AP and the N UEs over the F frequency subbands.

The channel distribution conditionned on the activity pattern $\mathbf{s} = \boldsymbol{s}$ is then

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}; \boldsymbol{B}) = \prod_{n \in [N]} \prod_{f \in [F]} f_{\mathsf{h}_{nf}|\mathsf{s}_n}(h_{nf} \mid s_n; b_{nf}) \tag{F.4}$$

assuming that the random channel coefficients are independent and where we have highlighted the dependency on the allocated subbands we the matrix $\boldsymbol{B}$. If we keep the assumption of complex gaussian i.i.d. random channel coefficients, we have

$$\forall (n,f) \in [N] \times [F],\ \mathsf{h}_{nk} \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } \mathsf{s}_n = 0 \text{ or } b_{nf} = 0 \\ \mathsf{CNorm}(\mu_{\mathsf{h}}, \tau_{\mathsf{h}}) & \text{if } \mathsf{s}_n = 1 \text{ and } b_{nf} = 1 \end{cases}. \tag{F.5}$$

Defining the new activity state $\tilde{\mathsf{s}}_{nf} = \mathsf{s}_n b_{nf}$, the channel coefficients distribution is characterized by the following pdf

$$\forall (n,f) \in [N] \times [F],\ \ f_{\mathsf{h}_{nf}|\mathsf{s}_n}(h_{nf} \mid \tilde{s}_{nf}) = \delta(h_{nf})^{1-\tilde{s}_{nf}} \mathcal{CN}(h_{nf}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})^{\tilde{s}_{nf}}. \quad \text{(F.6)}$$

The main differences between Eq. F.6 and the channel pdf used before in Eq. E.5 are

1. the factorization over the antennas is swaped for a factorization over the frequencies;
2. the activity states contain a deterministic component determined by the subbands allocation that may force the channel coefficients to be zero, no matter the UE states.

Based on these remarks, it is clear that the factor graph of Fig. E.4 can be reused as is, and so is GHetA-HGAMP. If one wants to consider a multi-antenna OFDM communication system, the factorization of the channel tensor $\mathcal{H} \in \mathbb{C}^{N \times K \times F}$ would read as

$$f_{\mathsf{H}|\mathsf{s}}(\mathcal{H} \mid \boldsymbol{s}; \boldsymbol{B}) = \prod_{n \in [N]} \prod_{k \in [K]} \prod_{f \in [F]} f_{\mathsf{h}_{nkf}|\mathsf{s}_n}(h_{nkf} \mid s_n; b_{fn}) \quad \text{(F.7)}$$

and the factor graph of Fig. E.4 would be augmented with the appropriate factor and variable nodes. Such a tensorization of the system model is not an issue to the use of a proper HGAMP algorithm that would in fact consist in parallel instances of GHetA-HGAMP along the antennas (or frequencies).

III Data transmission

An important aspect of the next generation services is the possibility to transmit small amount of data, skipping the traditional grant-based approaches. This idea is very similar to GFRA and corresponds to the scenario depicted in Fig. B.2c. As AUDaCE is a key enabler for GFRA, joint DrAUDaCE would be a key enabler for GF data transmission.

Different from Sec. B.II where the transmitted signal $\mathbf{X}$ was fully deterministic and described by the preamble matrix $\boldsymbol{P}$, it is now assumed that this matrix may be partially or entirely unknown.

Formally, one could consider the following structure for the random

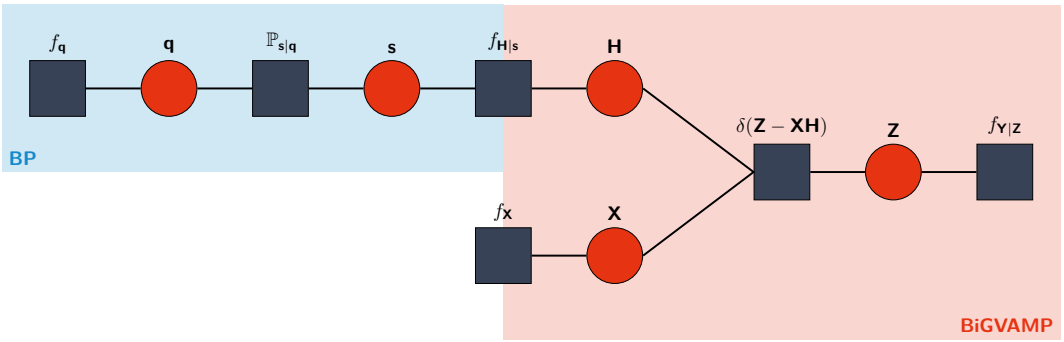


Figure F.2: Factor graph of Eq. F.11. The variable and factor nodes are taken in the vector/matrix form. The factor graph is composed of two parts: the right part includes the variable and factor nodes dedicated to the data recovery and the channel estimation while the left part includes the variable and factor nodes corresponding to the activity pattern. The data recovery and channel estimation is performed by BiGVAMP and the activity detection by BP.

signal matrix

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{P} \\ \hline \mathbf{D} \end{bmatrix} \in \mathbb{C}^{(M_p+M_d)\times N} \tag{F.8}$$

where the top part $\boldsymbol{P}$ of the matrix is deterministic (e.g. preambles) and the bottom part $\mathbf{D}$ contains the random data. When the random data are drawn from classical symbol constellations, DrAUDaCE shows connection with multi-user symbol data detection in non-coherent communications [126]–[128].

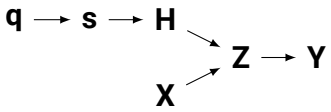
The lengths of the deterministic (M_p) and random (M_d) parts are chosen so that they fill $M = M_p + M_d$ OFDM symbols. Hence, the column $n \in [N]$ of $\mathbf{X}$ corresponds to the signal of the nth UE with deterministic (known) and random (unknown) components.

The rest of the transmission remains the same as in Sec. B.II

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \qquad \text{with} \qquad \mathbf{Z} = \mathbf{XH} \tag{F.9}$$

where all the system variables are random. If the channel matrix is known, DrAUDaCE reduces to joint data recovery and active user detection and one can consider [129]. One can expand the Markov chains Eqs. (D.35) and (E.54) in order to incorporate the new random signal

such that



(F.10)

The factorization of the joint pdf follows and reads

$$\begin{aligned} f_{\mathbf{q},\mathbf{s},\mathbf{H},\mathbf{X},\mathbf{Z},\mathbf{Y}}(\boldsymbol{q},\boldsymbol{s},\boldsymbol{H},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}) = & f_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{Y} \mid \boldsymbol{Z})\delta(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{H}) && \text{(system output)} \\ & f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s})\mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q})f_{\mathbf{q}}(\boldsymbol{q}) && \text{(channel \& activity)} \\ & f_{\mathbf{X}}(\boldsymbol{X}) && \text{(input signal)} \end{aligned} \tag{F.11}$$

where a GHetA pattern is assumed.

The DrAUDaCE problem then consists in the following bayesian estimator

$$(\boldsymbol{s}^\star, \boldsymbol{H}^\star, \boldsymbol{X}^\star) = \mathbb{E}[[\mathbf{s},\mathbf{H},\mathbf{X}] \mid \mathbf{Y} = \boldsymbol{Y}] \tag{F.12}$$

where the expectation relies on the joint factorized pdf above. Given Chapters D and E, it would be natural to develop a similar LBP algorithm with the factor graph represented (in a condensed form) in Fig. F.2 and then approximate it using the framework of EP and HGAMP.

However, a quick look at the factorization in Eq. F.11 and to its factor graph shows that the studied inference problem is more difficult to address because of the unknown $\mathbf{X}$. In Sec. C.III and Chapters D and E, the assumption that the measurement matrix $\boldsymbol{A}$ or signal matrix $\mathbf{X} = \boldsymbol{P}$ is known is crucial for the derivation of GAMP equations. This issue was first addressed in [105], [107] within the framework of bilinear GAMP (BiGAMP) where both the unknown signal and channel matrices are estimated by approximating the underlying LBP instance. This technique has successfully been applied to a communication scenario in [54], combining BiGAMP and HGAMP, but without considering any group structures or correlation structures that could be relevant for some scenarios.

Recently, this framework has been extended to bilinear GVAMP (BiGVAMP) [52], where the main difference lies in the fact that the algorithm is derived based on EP and that the variables are no longer scalars (entries of $\mathbf{X}$ and $\mathbf{H}$) but are vectors (rows of $\mathbf{X}$ and columns of $\mathbf{H}$). Unlike BiGAMP, BiGVAMP should benefit from VAMP which has been proved to be more

robust to potentially very ill-conditionned signal matrices [100]. To our knowledge, there is no attempt at the time of writing this manuscript to combine this more robust framework with BP for the activiy detection and to apply it in the context of grant-free data transmission.

Appendix

I Development environment

I.1 Typing and graphical content

This document has been typeset with LaTeX. Graphical contents which do not involve numerical simulations were generated using either TikZ [130] (e.g Figs. B.3 and C.6) or Inkscape [131] (e.g Figs. A.2, D.6 and E.5).

I.2 Simulations and related

Simulations were all implemented using the programming language Python [132]. Along Python's standard library, the scientific programming packages NumPy [133] and SciPy [134] are the core software bricks used to develop the code of the simulations. All the figures were generated using Matplotlib [135]. I am also thankful to ENS Lyon for letting my simulations run on the computing cluster of their Pole Scientifique de Modélisation Numérique [136].

II Miscellanious results

II.1 Gaussian product identity

Let $(z, \mu_a, \mu_b) \in \mathbb{C}^3$ and $(\tau_a, \tau_b) \in \mathbb{R}_{+*}^2$. Then the following identity holds:

$$\mathcal{CN}(z; \mu_a, \tau_a)\mathcal{CN}(z; \mu_b, \tau_b) = \mathcal{CN}(0; \mu_b - \mu_a, \tau_b + \tau_a)\mathcal{CN}(z; \mu_c, \tau_c) \tag{APX.1}$$

where

$$\tau_c = \frac{1}{\frac{1}{\tau_a} + \frac{1}{\tau_b}}, \qquad \mu_c = \tau_c \left(\frac{\mu_a}{\tau_a} + \frac{\mu_b}{\tau_b} \right). \tag{APX.2}$$

II.2 Gaussian quotient identity

Let $(z, \mu_a, \mu_b) \in \mathbb{C}^3$ and $(\tau_a, \tau_b) \in \mathbb{R}_{+*}^2$. Then the following identity holds:

$$\frac{\mathcal{CN}(z; \mu_a, \tau_a)}{\mathcal{CN}(z; \mu_b, \tau_b)} = \left(\frac{\tau_b}{\tau_b - \tau_a} \right)^2 \mathcal{CN}(0; \mu_b - \mu_a, \tau_b - \tau_a)\mathcal{CN}(z; \mu_c, \tau_c) \tag{APX.3}$$

where

$$\tau_c = \frac{1}{\frac{1}{\tau_a} - \frac{1}{\tau_b}}, \qquad \mu_c = \tau_c \left(\frac{\mu_a}{\tau_a} - \frac{\mu_b}{\tau_b} \right). \tag{APX.4}$$

III Proof of GAMP

This section provides a proof of GAMPs (see Sec. C.III.2) based on EPs following [92]. The proof is the same but written with more details.

III.1 Problem formulation

Precisely, the derivation of GAMPs is adapted to the following MMSE estimation problem

$$\boldsymbol{x}^\star = \mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}] = \frac{\int_{\mathbb{C}^N} \boldsymbol{x} \prod_{m\in[M]} f_{\mathbf{y}_m|\mathbf{x}}(y_m \mid \boldsymbol{x}) \prod_{n\in[N]} f_{\mathbf{x}_n}(x_n) \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{C}^N} \prod_{m\in[M]} f_{\mathbf{y}_m|\mathbf{x}}(y_m \mid \boldsymbol{x}) \prod_{n\in[N]} f_{\mathbf{x}_n}(x_n) \mathrm{d}\boldsymbol{x}} \tag{APX.5}$$

given that

$$\mathbf{y} = \mathbf{z} + \mathbf{w} \qquad \text{with} \qquad \mathbf{z} = \boldsymbol{A}\mathbf{x} \tag{APX.6}$$

where $\mathbf{x} \in \mathbb{C}^N, \mathbf{z} \in \mathbb{C}^M, \mathbf{w} \in \mathbb{C}^M, \mathbf{y} \in \mathbb{C}^M$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ for $M \leq N$. This MMSE estimator involves the factorization of the joint density

$$f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{m \in [M]} f_{\mathsf{y}_m|\mathbf{x}}(y_m \mid \boldsymbol{x}) \prod_{n \in [N]} f_{\mathsf{x}_n}(x_n) \tag{APX.7}$$

leading to the factor graph in Fig. C.6.

III.2 Derivation based on expectation propagation

We can write the EP messages using Sec. C.II.3 in the framework of VIEF

$$\underset{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}{\mathfrak{M}}(x) = \frac{\text{Proj}\left(f^{\text{out}}_{\mathsf{x}_n}(x)\right)}{\underset{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}{\mathfrak{M}}(x)} = \mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}) \tag{APX.8}$$

$$\underset{f_{\mathsf{x}_n} \leftarrow \mathsf{x}_n}{\mathfrak{M}}(x) = \prod_{m' \in [M]} \underset{f_{\mathsf{y}_{m'}|\mathsf{z}_{m'}} \to \mathsf{x}_n}{\mathfrak{M}}(x) = \mathcal{CN}(x; \mu_{f_{\mathsf{x}_n} \leftarrow \mathsf{x}_n}, \tau_{f_{\mathsf{x}_n} \leftarrow \mathsf{x}_n}) \tag{APX.9}$$

$$\underset{f_{\mathsf{x}_n} \to \mathsf{x}_n}{\mathfrak{M}}(x) = \frac{\text{Proj}\left(f^{\text{in}}_{\mathsf{x}_n}(x)\right)}{\underset{f_{\mathsf{x}_n} \leftarrow \mathsf{x}_n}{\mathfrak{M}}(x)} = \mathcal{CN}(x; \mu_{f_{\mathsf{x}_n} \to \mathsf{x}_n}, \tau_{f_{\mathsf{x}_n} \to \mathsf{x}_n}) \tag{APX.10}$$

$$\underset{f_{\mathsf{y}_m|\mathsf{z}_m} \leftarrow \mathsf{x}_n}{\mathfrak{M}}(x) = \underset{f_{\mathsf{x}_n} \to \mathsf{x}_n}{\mathfrak{M}}(x) \prod_{m' \in [M] \setminus \{m\}} \underset{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}{\mathfrak{M}}(x) = \mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m} \leftarrow \mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m} \leftarrow \mathsf{x}_n}) \tag{APX.11}$$

where

$$f^{\text{in}}_{\mathsf{x}_n}(x) = f_{\mathsf{x}_n}(x) \underset{f_{\mathsf{x}_n} \leftarrow \mathsf{x}_n}{\mathfrak{M}}(x), \tag{APX.12}$$

$$f^{\text{out}}_{\mathsf{x}_n}(x) = \underset{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_n}{\mathfrak{M}}(x) \int f_{\mathsf{y}_m|\mathsf{z}_m}(y_m \mid a_{mn}x + \sum_{n' \in [N] \setminus \{n\}} a_{mn'}x_{n'}) \prod_{n' \in [N] \setminus \{n\}} \underset{f_{\mathsf{y}_m|\mathsf{z}_m} \to \mathsf{x}_{n'}}{\mathfrak{M}}(x_{n'}) \mathrm{d}x_{n'}. \tag{APX.13}$$

Each of the message is approximated by a complex gaussian distribution and we look at how they are computed in the next paragraphs.

Projection of $f^{\text{out}}_{\mathsf{x}}$ Let $n \in [N]$ and $m \in [M]$. We first start by computing the projection of $f^{\text{out}}_{\mathsf{x}_n}$ on the family of complex gaussian distribution according to Sec. C.II.3. When $N \gg 1$, in virtue of the central limit theorem, one

have

$$\mathsf{z}_m = \sum_{n'} a_{mn'}\mathsf{x}_{n'} \sim \mathsf{CNorm}(\hat{p}_m, \tau_{\hat{p}_m}) \text{ with } \begin{cases} \hat{p}_m &= \sum_{n'\in[N]} a_{mn}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_{n'}} \\ \tau_{\hat{p}_m} &= \sum_{n'\in[N]} |a_{mn}|^2 \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_{n'}} \end{cases} \tag{APX.14}$$

and

$$\mathsf{z}_m \mid \mathsf{x}_n = x \sim \mathsf{CNorm}(a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}}) \text{ with } \begin{cases} \hat{p}_{mn} &= \sum_{n'\in[N]\setminus\{n\}} h_{mn'}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_{n'}} \\ \tau_{\hat{p}_{mn}} &= \sum_{n'\in[N]\setminus\{n\}} |a_{mn}|^2 \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_{n'}} \end{cases} \tag{APX.15}$$

Hence, the integral of $f^{\text{out}}_{\mathsf{x}_n}$ now reads as

$$f^{\text{out}}_{\mathsf{x}_n}(x) = \mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}) \int f_{\mathsf{y}_m|\mathsf{z}_m}(y_m \mid z)\mathcal{CN}(z; a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathrm{d}z \tag{APX.16}$$

$$= \int f_{\mathsf{y}_m|\mathsf{z}_m}(y_m \mid z)\mathcal{CN}(z; a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n})\mathrm{d}z \tag{APX.17}$$

where Eq. APX.15 has been used to approximate the density of $\mathsf{z}_m \mid \mathsf{x}_n = x$. We now focus on the product of gaussian pdfs. First, note that

$$\begin{aligned} &\mathcal{CN}(z; a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}) \\ &= \mathcal{CN}(a_{mn}x; z - \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}) && \text{(APX.18)} \\ &= \mathcal{CN}\left(x; \frac{z - \hat{p}_{mn}}{a_{mn}}, \frac{\tau_{\hat{p}_{mn}}}{|a_{mn}|^2}\right)\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}) && \text{(APX.19)} \end{aligned}$$

where we used elementary properties of the gaussian pdf. Then, using the gaussian product idendity

$$\mathcal{CN}(x; \mu_1, \tau_1)\mathcal{CN}(x; \mu_2, \tau_2) = \mathcal{CN}(0; \mu_1 - \mu_2, \tau_1 + \tau_2)\mathcal{CN}\left(x; \frac{\frac{\mu_1}{\tau_1} + \frac{\mu_2}{\tau_2}}{\frac{1}{\tau_1} + \frac{1}{\tau_2}}, \frac{1}{\frac{1}{\tau_1} + \frac{1}{\tau_2}}\right) \tag{APX.20}$$

we can write

$$\begin{aligned}&\mathcal{CN}(z; a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n})\\&= \mathcal{CN}\left(0; \frac{z-\hat{p}_{mn}}{a_{mn}} - \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \frac{\tau_{\hat{p}_{mn}}}{|a_{mn}|^2} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}\right)\mathcal{CN}\left(x; \mu_{\mathsf{x}_n,m}, \tau_{\mathsf{x}_n,m}\right)\end{aligned} \tag{APX.21}$$

where the new variance is

$$\tau_{\mathsf{x}_n,m} = \left(\frac{1}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}} + \frac{|a_{mn}|^2}{\tau_{\hat{p}_{mn}}}\right)^{-1} \tag{APX.22}$$

$$= \frac{\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}}{\tau_{\hat{p}_{mn}} + |a_{mn}|^2\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}} \tag{APX.23}$$

$$= \frac{\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}}{\tau_{\hat{p}_m}} \tag{APX.24}$$

and the new mean is

$$\mu_{\mathsf{x}_n,m} = \tau_{\mathsf{x}_n,m}\left(\frac{\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}} + \frac{a^*_{mn}(z-\hat{p}_{mn})}{\tau_{\hat{p}_{mn}}}\right) \tag{APX.25}$$

$$= \frac{\tau_{\hat{p}_{mn}}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}a^*_{mn}(z-\hat{p}_{mn})}{\tau_{\hat{p}_m}}. \tag{APX.26}$$

Using once again the properties of the gaussian pdf leads to

$$\begin{aligned}&\mathcal{CN}(z; a_{mn}x + \hat{p}_{mn}, \tau_{\hat{p}_{mn}})\mathcal{CN}(x; \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}}\mathsf{x}_n, \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}\mathsf{x}_n)\\&= \mathcal{CN}\left(0; z - \hat{p}_{mn} - a_{mn}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}, \tau_{\hat{p}_{mn}} + |a_{mn}|^2\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}\right)\mathcal{CN}\left(x; \mu_{\mathsf{x}_n,m}, \tau_{\mathsf{x}_n,m}\right)\end{aligned} \tag{APX.27}$$

$$= \mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\mathcal{CN}\left(x; \mu_{\mathsf{x}_n,m}, \tau_{\mathsf{x}_n,m}\right). \tag{APX.28}$$

The projection is then done by computing the mean and variance w.r.t. $f^{\text{out}}_{\mathsf{x}_n}(x)$:

$$
\begin{aligned}
&\mathbb{E}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right] \\
&= \frac{\int x f^{\text{out}}_{\mathsf{x}_n}(x)\mathrm{d}x}{\int f^{\text{out}}_{\mathsf{x}_n}(x)\mathrm{d}x} && \text{(APX.29)} \\
&= \frac{\iint x\mathcal{CN}\left(x; \mu_{\mathsf{x}_n,m}(z), \tau_{\mathsf{x}_n,m}\right) f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\mathrm{d}x\mathrm{d}z}{\iint \mathcal{CN}\left(x; \mu_{\mathsf{x}_n,m}(z), \tau_{\mathsf{x}_n,m}\right) f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\mathrm{d}x\mathrm{d}z} && \text{(APX.30)} \\
&= \int \mu_{\mathsf{x}_n,m}(z) f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\mathrm{d}z && \text{(APX.31)} \\
&= \mathbb{E}\left[\mu_{\mathsf{x}_n,m}(\mathsf{z}_m); f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right] && \text{(APX.32)} \\
&= \frac{\tau_{\hat{p}_{mn}}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}} \mathsf{x}_n + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} a^*_{mn}(\hat{z}_m - \hat{p}_{mn})}{\tau_{\hat{p}_m}} && \text{(APX.33)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{V}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right] \\
&= \mathbb{E}\left[|\mathsf{x}_n|^2; f^{\text{out}}_{\mathsf{x}_n}(x)\right] - \left|\mathbb{E}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right]\right|^2 && \text{(APX.34)} \\
&= \tau_{\mathsf{x}_n,m} + \mathbb{E}\left[|\mu_{\mathsf{x}_n,m}(\mathsf{z}_m)|^2; f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right] \\
&\qquad\qquad - \left|\mathbb{E}\left[\mu_{\mathsf{x}_n,m}(\mathsf{z}); f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right]\right|^2 && \text{(APX.35)} \\
&= \tau_{\mathsf{x}_n,m} + \mathbb{V}\left[\mu_{\mathsf{x}_n,m}(\mathsf{z}); f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right] && \text{(APX.36)} \\
&= \tau_{\mathsf{x}_n,m} + \mathbb{V}\left[\frac{\tau_{\hat{p}_{mn}}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n} a^*_{mn}(\mathsf{z}_m - \hat{p}_{mn})}{\tau_{\hat{p}_m}}; f_{\mathsf{y}_m|z}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right] \\
& && \text{(APX.37)} \\
&= \frac{\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}}{\tau_{\hat{p}_m}} + \frac{\tau^2_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}|a_{mn}|^2\tau_{\hat{z}_m}}{\tau^2_{\hat{p}_m}} && \text{(APX.38)} \\
&= \frac{\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}\tau_{\hat{p}_m} + \tau^2_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}|a_{mn}|^2\tau_{\hat{z}_m}}{\tau^2_{\hat{p}_m}} && \text{(APX.39)}
\end{aligned}
$$

with

$$
\begin{aligned}
\hat{z}_m &= \mathbb{E}\left[\mathsf{z}_m; f_{\mathsf{y}_m|\mathsf{z}_m}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right], && \text{(APX.40)} \\
\tau_{\hat{z}_m} &= \mathbb{V}\left[\mathsf{z}_m; f_{\mathsf{y}_m|\mathsf{z}_m}(y_m \mid z)\mathcal{CN}\left(z; \hat{p}_m, \tau_{\hat{p}_m}\right)\right]. && \text{(APX.41)}
\end{aligned}
$$

Computation of message Eq. APX.8 To compute the message Eq. APX.8, and so its mean and variance, we use the gaussian quotient identity

$$\frac{\mathcal{CN}(x;\mu_1,\tau_1)}{\mathcal{CN}(x;\mu_2,\tau_2)} = \left(\frac{\tau_2}{\tau_2-\tau_1}\right)^2 \frac{\mathcal{CN}\left(x;\frac{\frac{\mu_1}{\tau_1}-\frac{\mu_2}{\tau_2}}{\frac{1}{\tau_1}-\frac{1}{\tau_2}},\frac{1}{\frac{1}{\tau_1}-\frac{1}{\tau_2}}\right)}{\mathcal{CN}(0;\mu_2-\mu_1,\tau_2-\tau_1)}. \tag{APX.42}$$

It comes

$$\begin{aligned}
&\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow \mathsf{x}_n} \\
&= \left(\frac{1}{\mathbb{V}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right]} - \frac{1}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}\right)^{-1} && \text{(APX.43)}\\
&= \frac{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\mathbb{V}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right]}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} - \mathbb{V}\left[\mathsf{x}_n; f^{\text{out}}_{\mathsf{x}_n}(x)\right]} && \text{(APX.44)}\\
&= \frac{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\frac{\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}+\tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}{\mathsf{x}_n}^2|a_{mn}|^2}{\tau^2_{\hat{p}_m}}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} - \frac{\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}+\tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}{\mathsf{x}_n}^2|a_{mn}|^2}{\tau^2_{\hat{p}_m}}} && \text{(APX.45)}\\
&= \frac{\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}{\mathsf{x}_n}^2|a_{mn}|^2\tau_{\hat{z}_m}}{\tau^2_{\hat{p}_m} - \tau_{\hat{p}_m}\tau_{\hat{p}_{mn}} - \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2\tau_{\hat{z}_m}} && \text{(APX.46)}\\
&= \frac{\tau_{\hat{p}_m}(\tau_{\hat{p}_m} - \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2)\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}{\mathsf{x}_n}^2|a_{mn}|^2\tau_{\hat{z}_m}}{\tau^2_{\hat{p}_m} - \tau_{\hat{p}_m}(\tau_{\hat{p}_m} - |a_{mn}|^2\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}) - \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2\tau_{\hat{z}_m}} && \text{(APX.47)}\\
&= \frac{\tau^2_{\hat{p}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}}{\mathsf{x}_n}^2|a_{mn}|^2(\tau_{\hat{z}_m} - \tau_{\hat{p}_m})}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2(\tau_{\hat{p}_m} - \tau_{\hat{z}_m})} && \text{(APX.48)}\\
&= \frac{\tau^2_{\hat{p}_m} - \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2(\tau_{\hat{p}_m} - \tau_{\hat{z}_m})}{|a_{mn}|^2(\tau_{\hat{p}_m} - \tau_{\hat{z}_m})} && \text{(APX.49)}
\end{aligned}$$

for the variance and for the mean

$$
\begin{aligned}
&\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} \\
&= \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\left(\frac{\mathbb{E}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{out}}(x)\right]}{\mathbb{V}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{out}}(x)\right]} - \frac{\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}\right) && \text{(APX.50)} \\
&= \frac{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\mathbb{E}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{out}}(x)\right] - \mathbb{V}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{out}}(x)\right]\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} - \mathbb{V}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{out}}(x)\right]} && \text{(APX.51)} \\
&= \frac{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\frac{\tau_{\hat{p}_{mn}}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}+\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}a_{mn}^*(\hat{z}_m-\hat{p}_{mn})}{\tau_{\hat{p}_m}} - \frac{\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}+\tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}{}^2|a_{mn}|^2}{\tau_{\hat{p}_m}^2}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} - \frac{\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}+\tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}{}^2|a_{mn}|^2}{\tau_{\hat{p}_m}^2}} && \text{(APX.52)} \\
&= \frac{\tau_{\hat{p}_m}\left(\tau_{\hat{p}_{mn}}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} + \tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}a_{mn}^*(\hat{z}_m - \hat{p}_{mn})\right) - \left(\tau_{\hat{p}_m}\tau_{\hat{p}_{mn}} + \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2\right)\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{\tau_{\hat{p}_m}^2 - \tau_{\hat{p}_m}\tau_{\hat{p}_{mn}} - \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2} && \text{(APX.53)} \\
&= \frac{\tau_{\hat{p}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}a_{mn}^*(\hat{z}_m - \hat{p}_{mn}) - \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{\tau_{\hat{p}_m}^2 - \tau_{\hat{p}_m}(\tau_{\hat{p}_m} - |a_{mn}|^2\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}) - \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2} && \text{(APX.54)} \\
&= \frac{\tau_{\hat{p}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}a_{mn}^*(\hat{z}_m - \hat{p}_m + a_{mn}\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}) - \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}}{|a_{mn}|^2\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} - \tau_{\hat{z}_m}\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}|a_{mn}|^2} && \text{(APX.55)} \\
&= \frac{\tau_{\hat{p}_m}a_{mn}^*(\hat{z}_m - \hat{p}_m) - |a_{mn}|^2\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}(\tau_{\hat{p}_m} - \tau_{\hat{z}_m})}{|a_{mn}|^2 - \tau_{\hat{z}_m}|a_{mn}|^2}. && \text{(APX.56)}
\end{aligned}
$$

A key step of GAMP is that it neglects terms of order $O(|a_{mn}|^2)$ so that the following approximations can be made

$$\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} = \frac{1}{|a_{mn}|^2\tau_{\hat{v}_m}} + O(|a_{mn}|^2) \qquad \text{(APX.57)}$$

$$\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} = \frac{a_{mn}^*\hat{v}_m + |a_{mn}|^2\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n}\tau_{\hat{v}_m}}{|a_{mn}|^2\tau_{\hat{v}_m}} \qquad \text{(APX.58)}$$

with

$$\tau_{\hat{v}_m} = \frac{\tau_{\hat{p}_m} - \tau_{\hat{z}_m}}{\tau_{\hat{p}_m}^2}, \qquad \hat{v}_m = \frac{\hat{z}_m - \hat{p}_m}{\tau_{\hat{p}_m}}. \qquad \text{(APX.59)}$$

Message from x_n to f_{x_n} Remembering that

$$\underset{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}{\mathfrak{M}}(x) = \prod_{m'\in[M]} \mathcal{CN}(x;\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\to\mathsf{x}_n},\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\to\mathsf{x}_n}) = \mathcal{CN}(x;\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n},\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}) \tag{APX.60}$$

where, based on Eq. APX.20, the variance is

$$\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n} = \left(\sum_{m\in[M]} \frac{1}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\to\mathsf{x}_n}}\right)^{-1} \tag{APX.61}$$

$$= \left(\sum_{m\in[M]} |a_{mn}|^2 \tau_{\hat{v}_m}\right)^{-1} \tag{APX.62}$$

and the mean is

$$\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}, = \tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n} \sum_{m\in[M]} \frac{\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\to\mathsf{x}_n}}{\tau_{f_{\mathsf{y}_m|\mathsf{z}_m}\to\mathsf{x}_n}} \tag{APX.63}$$

$$= \tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n} \sum_{m\in[M]} \left(a_{mn}^* \hat{v}_m + |a_{mn}|^2 \mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\mathsf{x}_n} \tau_{\hat{v}_m}\right). \tag{APX.64}$$

Projection of $f_{\mathsf{x}}^{\text{in}}$ The projection of $f_{\mathsf{x}}^{\text{in}}$ on a complex gaussian distribution requires the computation of the mean and variance

$$\hat{x}_n = \mathbb{E}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{in}}(x)\right] = \mathbb{E}\left[\mathsf{x}_n; f_{\mathsf{x}_n}(x)\mathcal{CN}(x;\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n},,\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n})\right] \tag{APX.65}$$

$$\tau_{\hat{x}_n} = \mathbb{V}\left[\mathsf{x}_n; f_{\mathsf{x}_n}^{\text{in}}(x)\right] = \mathbb{V}\left[\mathsf{x}_n; f_{\mathsf{x}_n}(x)\mathcal{CN}(x;\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n},,\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n})\right] \tag{APX.66}$$

that will depend on the model.

Message from f_{x_n} to x_n Similar to the message from $f_{\mathsf{y}_m|\mathsf{z}_m}$ to x_n, we use Eq. APX.42 to compute the message variance

$$\tau_{f_{\mathsf{x}_n}\to\mathsf{x}_n} = \left(\frac{1}{\tau_{\hat{x}_n}} - \frac{1}{\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}}\right)^{-1} \tag{APX.67}$$

$$= \left(\frac{1}{\tau_{\hat{x}_n}} - |a_{mn}|^2 \tau_{\hat{v}_m}\right)^{-1} \tag{APX.68}$$

$$= \tau_{\hat{x}_n} + O(|a_{mn}|^2) \tag{APX.69}$$

and the message mean

$$\mu_{f_{\mathsf{x}_n}\to\mathsf{x}_n} = \tau_{f_{\mathsf{x}_n}\to\mathsf{x}_n}\left(\frac{\hat{x}_n}{\tau_{\hat{x}_n}} - \frac{\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}}{\tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}}\right), \tag{APX.70}$$

$$= \tau_{\hat{x}_n}\left(\frac{\hat{x}_n}{\tau_{\hat{x}_n}} - a_{mn}^*\hat{v}_m + |a_{mn}|^2\,\mu_{f_{\mathsf{y}_m|\mathsf{z}_m}\leftarrow\mathsf{x}_n}\tau_{\hat{v}_m}\right) \tag{APX.71}$$

$$= \hat{x}_n - \tau_{\hat{x}_n}a_{mn}^*\hat{v}_m + O(|a_{mn}|^2) \tag{APX.72}$$

removing higher order terms.

Closing the loop The loop is finally closed by injecting the precedent mean and variance in $\tau_{\hat{p}_m}$, $\hat{p}_m$ and $\mu_{f_{\mathsf{x}_n}}\mathsf{x}_n$ to obtain the final approximations

$$\tau_{\hat{p}_m} = \sum_{n\in[N]} |a_{mn}|^2\,\tau_{f_{\mathsf{x}_n}}\mathsf{x}_n \tag{APX.73}$$

$$= \sum_{n\in[N]} |a_{mn}|^2\,\tau_{\hat{x}_n} \tag{APX.74}$$

$$\hat{p}_m = \sum_{n\in[N]} a_{mn}\mu_{f_{\mathsf{x}_n}\to\mathsf{x}_n} \tag{APX.75}$$

$$= \sum_{n\in[N]} a_{mn}(\hat{x}_n - \tau_{\hat{x}_n}a_{mn}^*\hat{v}_m) \tag{APX.76}$$

$$= \sum_{n\in[N]} a_{mn}\hat{x}_n - \tau_{\hat{p}_m}\hat{v}_m \tag{APX.77}$$

$$\mu_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n} = \tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}\left(\sum_{m\in[M]}\left(a_{mn}^*\hat{v}_m + |a_{mn}|^2\,(\hat{x}_n - \tau_{\hat{x}_n}a_{mn}^*\hat{v}_m)\tau_{\hat{v}_m}\right)\right) \tag{APX.78}$$

$$= \hat{x}_n + \tau_{f_{\mathsf{x}_n}\leftarrow\mathsf{x}_n}\left(\sum_{m\in[M]} a_{mn}^*\hat{v}_m\right) \tag{APX.79}$$

The resulting algorithm simply consists in the exchange of messages that are means and variances, namely GAMP.

IV Derivation of GHomA-HGAMP

This section provides a derivation of an instance of HGAMPs within the GHomAs framework. The final algorithm is given in Sec. D.III.2.

In the sequel, we assume that $n \in [N], k \in [K], m \in [M]$ and $g \in [G]$. When the indices n and g appear in the same expression, it is assumed that $n \in \mathfrak{G}_g$.

IV.1 From the channel output to the activity probabilities

We first consider the propagation of the messages from $f_{\mathbf{Y}|\mathbf{H}}$ to $f_{\mathbf{q}}$ given the following chain

$$f_{\mathbf{q}} \leftarrow \mathbf{q} \leftarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}} \leftarrow \mathbf{s} \leftarrow f_{\mathbf{H}|\mathbf{s}} \leftarrow \mathbf{H} \leftarrow f_{\mathbf{Y}|\mathbf{H}} \tag{APX.80}$$

where we already dealt with the subchain $f_{\mathbf{H}|\mathbf{s}} \leftarrow \mathbf{H} \leftarrow f_{\mathbf{Y}|\mathbf{H}}$ using GAMP. The message of the next link $\mathbf{s} \leftarrow f_{\mathbf{H}|\mathbf{s}}$ becomes

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n \to s_n}}{\mathfrak{M}}(s_n) \approx \begin{cases} \int \delta(x)\mathcal{CN}(x; r_{nk}, \tau_{r,nk})\mathrm{d}x & \text{if } s_n = 0 \\ \int \mathcal{CN}(h; \mu_{\mathsf{h}}, \tau^x)\mathcal{CN}(x; r_{nk}, \tau_{r,nk})\mathrm{d}x & \text{if } s_n = 1 \end{cases} \tag{APX.81}$$

$$\approx \begin{cases} \mathcal{CN}(0; r_{nk}, \tau_{r,nk}) & \text{if } s_n = 0 \\ \mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases}. \tag{APX.82}$$

The next link $\mathbb{P}_{\mathbf{s}|\mathbf{q}} \leftarrow \mathbf{s}$ collect all the messages for each antenna $k' \in [K]$ given the rules of Sec. C.II.3:

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g \leftarrow s_n}}{\mathfrak{M}}(s_n) \approx \begin{cases} \prod_{k' \in [K]} \mathcal{CN}(0; r_{nk}, \tau_{r,nk}) & \text{if } s_n = 0 \\ \prod_{k' \in [K]} \mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases} \tag{APX.83}$$

$$\approx \begin{cases} \phi_{0,n} \triangleq \mathcal{CN}(0; \boldsymbol{r}_n, \mathrm{diag}(\boldsymbol{\tau}_{r,n})) & \text{if } s_n = 0 \\ \phi_{1,n} \triangleq \mathcal{CN}(0; \boldsymbol{r}_n - \mu_{\mathsf{h}}\mathbf{1}_{K\times 1}, \mathrm{diag}(\boldsymbol{\tau}_{r,n} + \tau_{\mathsf{h}}\mathbf{1}_{K\times 1})) & \text{if } s_n = 1 \end{cases} \tag{APX.84}$$

where $\boldsymbol{r}_n = [r_{nk'}]^{\mathsf{T}}_{k' \in [K]}$ and $\boldsymbol{\tau}_{r,n} = [\tau_{r,nk'}]^{\mathsf{T}}_{k' \in [K]}$. Then it comes

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g \to q_g}}{\mathfrak{M}}(q_g) \approx (1 - q_g)\phi_{0,n} + q_g\phi_{1,n} \tag{APX.85}$$

for the link $\mathbf{q} \leftarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}}$ after averaging over the state variable before being propagated unchanged

$$\underset{f_{\mathsf{q}_g} \leftarrow q_g}{\mathfrak{M}}(q_g) \approx \prod_{n \in \mathfrak{G}_g} \underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g} \rightarrow q_g}{\mathfrak{M}}(q_g) \tag{APX.86}$$

with the last link $f_{\mathbf{q}} \leftarrow \mathbf{q}$.

IV.2 From the activity probabilities to the channel output

The propagation is then conducted in the opposite direction from $f_{\mathbf{q}}$ to $f_{\mathbf{Y}|\mathbf{H}}$ reversing the chain of Eq. APX.80 as

$$f_{\mathbf{q}} \rightarrow \mathbf{q} \rightarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}} \rightarrow \mathbf{s} \rightarrow f_{\mathbf{H}|\mathbf{s}} \rightarrow \mathbf{H} \rightarrow f_{\mathbf{Y}|\mathbf{H}}. \tag{APX.87}$$

The first message corresponding to the link $f_{\mathbf{q}} \rightarrow \mathbf{q}$ requires the computation of the marginalization

$$\underset{f_{\mathsf{q}_g} \rightarrow q_g}{\mathfrak{M}}(q_g) = f_{\mathsf{q}_g}(q_g) \tag{APX.88}$$

and is sent through $\mathbf{q} \rightarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}}$

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g} \leftarrow q_g}{\mathfrak{M}}(q_g) \approx f_{\mathsf{q}_g}(q_g) \prod_{n' \in \mathfrak{G}_g} ((1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}) \tag{APX.89}$$

The message of the link $\mathbb{P}_{\mathbf{s}|\mathbf{q}} \rightarrow \mathbf{s}$ is obtained with

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g} \rightarrow s_n}{\mathfrak{M}}(s_n) \approx \int_{[0,1]} \mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g}(s_n|q_g) f_{\mathsf{q}_g}(q_g) \left[\prod_{n' \in \mathfrak{G}_g \setminus \{n\}} ((1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}) \, \mathrm{d}q_g \right] \tag{APX.90}$$

$$\approx \begin{cases} 1 - \hat{q}_{g,n} & \text{if } s_n = 0 \\ \hat{q}_{g,n} & \text{if } s_n = 1 \end{cases} \tag{APX.91}$$

where

$$\hat{q}_{g,n} = \mathbb{E}\left[\mathsf{q}_g; \underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_g \leftarrow q_g}}{\mathfrak{M}}(q_g)\right] \tag{APX.92}$$

$$= \frac{\int_{[0,1]} q_g f_{\mathsf{q}_g}(q_g) \left[\prod_{n' \in \mathfrak{G}_g \setminus \{n\}} \left((1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}\right) \mathrm{d}q_g\right]}{\int_{[0,1]} f_{\mathsf{q}_g}(q_g) \left[\prod_{n' \in \mathfrak{G}_g \setminus \{n\}} \left((1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}\right) \mathrm{d}q_g\right]} \tag{APX.93}$$

is the mean of the random activity probability given the state s_n. Next, the message is

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \leftarrow s_n}{\mathfrak{M}}(s_n) \approx \begin{cases} (1-\hat{q}_{g,n}) \prod\limits_{k' \in [K]\setminus\{k\}} \mathcal{CN}(0; r_{n'k'}, \tau_{r,n'k'}) & \text{if } s_n = 0 \\ \hat{q}_{g,n} \prod_{k' \in [K]\setminus\{k\}} \mathcal{CN}(0; r_{n'k'} - \mu_{\mathsf{h}}, \tau_{r,n'k'} + \tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases} \tag{APX.94}$$

$$\approx \begin{cases} (1-\hat{q}_{g,n}) \dfrac{\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} & \text{if } s_n = 0 \\ \hat{q}_{g,n} \dfrac{\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} & \text{if } s_n = 1 \end{cases} \tag{APX.95}$$

which consist in aggregating the messages for each antenna k for the link $\mathbf{s} \to f_{\mathbf{H}|\mathbf{s}}$. The message of the subsequent link $f_{\mathbf{H}|\mathbf{s}} \to \mathbf{H}$ averages over the state

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \to h_{nk}}{\mathfrak{M}}(h_{nk}) \approx (1-\hat{q}_{g,n})\, \delta(h_{nk}) \frac{\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} + \hat{q}_{g,n} \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}}) \frac{\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} \tag{APX.96}$$

IV.3 Beliefs

After the round trip of messages described by Eq. APX.80 and Eq. APX.87, estimates of the posterior beliefs are obtained based on Sec. C.II.3.

After Eq. D.39, the beliefs at variables q_g, s_n and h_{nk} are respectively

$$\underset{q_g}{\mathfrak{B}}(q_g) \approx \int_{[0,1]} f_{\mathsf{q}_g}(q_g) \left(\prod_{n' \in [N]} \left[(1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}\right] \mathrm{d}q_g \right) \tag{APX.97}$$

$$\underset{s_n}{\mathfrak{B}}(s_n) \approx \begin{cases} (1-\hat{q}_{g,n})\,\phi_{0,n} & \text{if } s_n = 0 \\ \hat{q}_{g,n}\phi_{1,n} & \text{if } s_n = 1 \end{cases}, \tag{APX.98}$$

and

$$\underset{h_{nk}}{\mathfrak{B}}(h_{nk}) \approx \mathcal{CN}(h_{nk}; r_{nk}, \tau_{r,nk}) \left[\frac{(1-\hat{q}_{g,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} \delta(h_{nk}) + \frac{\hat{q}_{g,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}}) \right]. \tag{APX.99}$$

IV.4 Estimates

Once the beliefs are derived, one can compute estimates of the system variables. We start with the estimates of the activity probabilities

$$\hat{q}_g = \mathbb{E}\left[\mathsf{q}_g; \underset{q_g}{\mathfrak{B}}(q_g)\right] = \frac{\int q_g \underset{q_g}{\mathfrak{B}}(q_g) \mathrm{d}q_g}{\int \underset{q_g}{\mathfrak{B}}(q_g) \mathrm{d}q_g} \tag{APX.100}$$

where the normalization ensures that the belief is a pdf. Substituting the belief inside the integrals leads to

$$\hat{q}_g \propto \int q_g f_{\mathsf{q}_g}(q_g) \left[\prod_{n' \in [N]} ((1-q_g)\phi_{0,n'} + q_g \phi_{1,n'}) \right] \mathrm{d}q_g. \tag{APX.101}$$

The estimate of the states is performed using individual log-likelihood ratio tests of the form

$$\hat{s}_n = \mathbb{1}\left(\mathsf{LLR}_n > 0\right) \tag{APX.102}$$

where

$$\mathsf{LLR}_n = \log \hat{q}_{g,n} + \log \phi_{0,n} - \log\left(1-\hat{q}_{g,n}\right) - \log \phi_{1,n} \tag{APX.103}$$

Note that the dependence on the group index g is removed in the notation LLR_n since the groups do not overlap.

Finally, the channel estimation is performed by computing

$$\hat{h}_{nk} = \frac{\int_{\mathbb{C}} h \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h}{\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h} \tag{APX.104}$$

from Eq. APX.99. We start by computing the denominator as

$$\begin{aligned}\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h =& \frac{(1-\hat{q}_{g,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} \int_{\mathbb{C}} \delta(h)\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \\ &+ \frac{\hat{q}_{g,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk}-\mu_{\mathsf{h}}, \tau_{r,nk}+\tau_{\mathsf{h}})} \int_{\mathbb{C}} \mathcal{CN}(h; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \end{aligned} \tag{APX.105}$$

$$= (1-\hat{q}_{g,n})\,\phi_{0,n} + \hat{q}_{g,n}\phi_{1,n} \tag{APX.106}$$

The numerator is

$$\begin{aligned}\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h =& \frac{(1-\hat{q}_{g,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} \int_{\mathbb{C}} h\delta(h)\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \\ &+ \frac{\hat{q}_{g,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk}-\mu_{\mathsf{h}}, \tau_{r,nk}+\tau_{\mathsf{h}})} \int_{\mathbb{C}} h\mathcal{CN}(h; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \end{aligned} \tag{APX.107}$$

$$= \hat{q}_{g,n}\phi_{1,n} \frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1} r_{nk}}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}. \tag{APX.108}$$

Finally

$$\hat{h}_{nk} = \frac{1}{1+e^{-\mathsf{LLR}_n}} \frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1} r_{nk}}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}} \tag{APX.109}$$

For the purpose of the updates of the GAMP part of the algorithm, we also compute the corresponding variance

$$\tau_{h,nk} = \mathbb{V}\left[\mathsf{h}_{nk}; \underset{h_{nk}}{\mathfrak{B}}\right] = \mathbb{E}\left[|\mathsf{h}_{nk}|^2; \underset{h_{nk}}{\mathfrak{B}}\right] - \left|\mathbb{E}\left[|\mathsf{h}_{nk}|; \underset{h_{nk}}{\mathfrak{B}}\right]\right|^2 = \mathbb{E}\left[|\mathsf{h}_{nk}|^2; \underset{h_{nk}}{\mathfrak{B}}\right] - \left|\hat{h}_{nk}\right|^2 \tag{APX.110}$$

The second moment of h_{nk} computation is similar to the mean computa-

tion, leading to

$$\mathbb{E}\left[\left|\mathsf{h}_{nk}\right|^2 ; \underset{h_{nk}}{\mathfrak{B}}\right] = \frac{1}{1+e^{-\mathsf{LLR}_n}}\left(\left|\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}}+\tau_{r,nk}^{-1}\mu_r}{\tau_{\mathsf{h}}^{-1}+\tau_{r,nk}^{-1}}\right|^2 + \frac{1}{\tau_{\mathsf{h}}^{-1}+\tau_{r,nk}^{-1}}\right). \quad \text{(APX.111)}$$

Finally, the variance reads

$$\tau_{h,nk} = \frac{1}{1+e^{-\mathsf{LLR}_{nk}}}\left(\left|\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}}+\tau_{r,nk}^{-1}\mu_r}{\tau_{\mathsf{h}}^{-1}+\tau_{r,nk}^{-1}}\right|^2 + \frac{1}{\tau_{\mathsf{h}}^{-1}+\tau_{r,nk}^{-1}}\right) - \left|\hat{h}_{nk}\right|^2. \quad \text{(APX.112)}$$

V Derivation of GHetA-HGAMP

This section provides a derivation of an instance of HGAMPs within the GHetAs framework. The final algorithm is given in Sec. E.III.2.

In the sequel, we assume that $n \in [N], k \in [K], m \in [M]$.

V.1 From the channel output to the activity probabilities

We first consider the propagation of the messages from $f_{\mathbf{Y}|\mathbf{H}}$ to $f_{\mathbf{q}}$ given the following chain

$$f_{\mathbf{q}} \leftarrow \mathbf{q} \leftarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}} \leftarrow \mathbf{s} \leftarrow f_{\mathbf{H}|\mathbf{s}} \leftarrow \mathbf{H} \leftarrow f_{\mathbf{Y}|\mathbf{H}} \quad \text{(APX.113)}$$

where we already dealt with the subchain $f_{\mathbf{H}|\mathbf{s}} \leftarrow \mathbf{H} \leftarrow f_{\mathbf{Y}|\mathbf{H}}$ using GAMP. The message of the next link $\mathbf{s} \leftarrow f_{\mathbf{H}|\mathbf{s}}$ becomes

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n}\to s_n}{\mathfrak{M}}(s_n) \approx \begin{cases} \int \delta(x)\mathcal{CN}(x; r_{nk}, \tau_{r,nk})\mathrm{d}x & \text{if } s_n = 0 \\ \int \mathcal{CN}(h; \mu_{\mathsf{h}}, \tau^x)\mathcal{CN}(x; r_{nk}, \tau_{r,nk})\mathrm{d}x & \text{if } s_n = 1 \end{cases} \quad \text{(APX.114)}$$

$$\approx \begin{cases} \mathcal{CN}(0; r_{nk}, \tau_{r,nk}) & \text{if } s_n = 0 \\ \mathcal{CN}(0; r_{nk}-\mu_{\mathsf{h}}, \tau_{r,nk}+\tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases}. \quad \text{(APX.115)}$$

The next link $\mathbb{P}_{\mathbf{s}|\mathbf{q}} \leftarrow \mathbf{s}$ collect all the messages for each antenna $k' \in [K]$ given the rules of Sec. C.II.3:

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \leftarrow s_n}{\mathfrak{M}} (s_n) \approx \begin{cases} \prod_{k' \in [K]} \mathcal{CN}(0; r_{nk}, \tau_{r,nk}) & \text{if } s_n = 0 \\ \prod_{k' \in [K]} \mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases} \tag{APX.116}$$

$$\approx \begin{cases} \phi_{0,n} \triangleq \mathcal{CN}(0; \boldsymbol{r}_n, \mathrm{diag}(\boldsymbol{\tau}_{r,n})) & \text{if } s_n = 0 \\ \phi_{1,n} \triangleq \mathcal{CN}(0; \boldsymbol{r}_n - \mu_{\mathsf{h}} \mathbf{1}_{K\times 1}, \mathrm{diag}(\boldsymbol{\tau}_{r,n} + \tau_{\mathsf{h}} \mathbf{1}_{K\times 1})) & \text{if } s_n = 1 \end{cases} \tag{APX.117}$$

where $\boldsymbol{r}_n = [r_{nk'}]^{\mathsf{T}}_{k' \in [K]}$ and $\boldsymbol{\tau}_{r,n} = [\tau_{r,nk'}]^{\mathsf{T}}_{k' \in [K]}$. Then it comes

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \rightarrow q_n}{\mathfrak{M}} (q_n) \approx (1 - q_n)\phi_{0,n} + q_n \phi_{1,n} \tag{APX.118}$$

for the link $\mathbf{q} \leftarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}}$ after averaging over the state variable before being propagated unchanged

$$\underset{f_{\mathbf{q}} \leftarrow q_n}{\mathfrak{M}} (q_n) \approx \underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \rightarrow q_n}{\mathfrak{M}} (q_n) \tag{APX.119}$$

with the last link $f_{\mathbf{q}} \leftarrow \mathbf{q}$.

V.2 From the activity probabilities to the channel output

The propagation is then conducted in the opposite direction from $f_{\mathbf{q}}$ to $f_{\mathbf{Y}|\mathbf{H}}$ reversing the chain of Eq. APX.113 as

$$f_{\mathbf{q}} \rightarrow \mathbf{q} \rightarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}} \rightarrow \mathbf{s} \rightarrow f_{\mathbf{H}|\mathbf{s}} \rightarrow \mathbf{H} \rightarrow f_{\mathbf{Y}|\mathbf{H}}. \tag{APX.120}$$

The first message corresponding to the link $f_{\mathbf{q}} \rightarrow \mathbf{q}$ requires the computation of the marginalization

$$\underset{f_{\mathbf{q}} \rightarrow q_n}{\mathfrak{M}} (q_n) \approx \int_{[0,1]^{N-1}} f_{\mathbf{q}}(\boldsymbol{q}) \left[\prod_{n' \in [N] \setminus \{n\}} ((1 - q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}) \, \mathrm{d}q_{n'} \right] \tag{APX.121}$$

and is sent through $\mathbf{q} \rightarrow \mathbb{P}_{\mathbf{s}|\mathbf{q}}$

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n} \leftarrow q_n}{\mathfrak{M}} (q_n) \approx \underset{f_{\mathbf{q}} \rightarrow q_n}{\mathfrak{M}} (q_n). \tag{APX.122}$$

The message of the link $\mathbb{P}_{\mathbf{s}|\mathbf{q}} \to \mathbf{s}$ is obtained with

$$\underset{\mathbb{P}_{\mathsf{s}_n|\mathsf{q}_n \to s_n}}{\mathfrak{M}}(s_n) \approx \begin{cases} 1-\hat{q}_{n,n} & \text{if } s_n = 0 \\ \hat{q}_{n,n} & \text{if } s_n = 1 \end{cases} \tag{APX.123}$$

where we use the shorthand notation

$$\hat{q}_{n,n} = \mathbb{E}\left[\mathsf{q}_n; \underset{f_{\mathbf{q}} \to \mathsf{q}_n}{\mathfrak{M}}\right] \tag{APX.124}$$

$$= \frac{\int_{[0,1]^{N-1}} f_{\mathbf{q}}(\boldsymbol{q}) \left[\prod_{n' \in [N]\setminus\{n\}} \left((1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}\right) \mathrm{d}q_{n'}\right]}{\int_{[0,1]^{N-1}} \left[\prod_{n' \in [N]\setminus\{n\}} \left((1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'}\right) \mathrm{d}q_{n'}\right]} \tag{APX.125}$$

compute the mean of the random activity probability. Next, the message is

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \leftarrow s_n}{\mathfrak{M}}(s_n) \approx \begin{cases} (1-\hat{q}_{n,n}) \prod\limits_{k' \in [K]\setminus\{k\}} \mathcal{CN}(0; r_{n'k'}, \tau_{r,n'k'}) & \text{if } s_n = 0 \\ \hat{q}_{n,n} \prod_{k' \in [K]\setminus\{k\}} \mathcal{CN}(0; r_{n'k'} - \mu_{\mathsf{h}}, \tau_{r,n'k'} + \tau_{\mathsf{h}}) & \text{if } s_n = 1 \end{cases} \tag{APX.126}$$

$$\approx \begin{cases} (1-\hat{q}_{n,n}) \dfrac{\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} & \text{if } s_n = 0 \\ \hat{q}_{n,n} \dfrac{\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} & \text{if } s_n = 1 \end{cases} \tag{APX.127}$$

which consist in aggregating the messages for each antenna the pair k for the link $\mathbf{s} \to f_{\mathbf{H}|\mathbf{s}}$. The message of the subsequent link $f_{\mathbf{H}|\mathbf{s}} \to \mathbf{H}$ averages over the state

$$\underset{f_{\mathsf{h}_{nk}|\mathsf{s}_n} \to h_{nk}}{\mathfrak{M}}(h_{nk}) \approx (1-\hat{q}_{n,n})\, \delta(h_{nk}) \frac{\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} + \hat{q}_{n,n} \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}}) \frac{\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} \tag{APX.128}$$

V.3 Beliefs

After the round trip of messages described by Eq. APX.113 and Eq. APX.120, estimates of the posterior beliefs are obtained based on Sec. C.II.3.

After Eqs. (E.59) to (E.61), the beliefs at variables q_n, s_n and h_{nk} are

respectively

$$\underset{q_n}{\mathfrak{B}}(q_n) \approx \int_{[0,1]^{N-1}} f_{\mathsf{q}}(\boldsymbol{q}) \left(\prod_{n' \in [N]} \left[(1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'} \right] \mathrm{d}q_{n'} \right) \tag{APX.129}$$

$$\underset{s_n}{\mathfrak{B}}(s_n) \approx \begin{cases} \left(1 - \mathbb{E}\left[\mathsf{q}_n; \underset{f_{\mathsf{q}} \to \mathsf{q}_n}{\mathfrak{M}}\right]\right) \phi_{0,n} & \text{if } s_n = 0 \\ \mathbb{E}\left[\mathsf{q}_n; \underset{f_{\mathsf{q}} \to \mathsf{q}_n}{\mathfrak{M}}\right] \phi_{1,n} & \text{if } s_n = 1 \end{cases}, \tag{APX.130}$$

and

$$\underset{h_{nk}}{\mathfrak{B}}(h_{nk}) \approx \mathcal{CN}(h_{nk}; r_{nk}, \tau_{r,nk}) \left[\frac{(1-\hat{q}_{n,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})} \delta(h_{nk}) + \frac{\hat{q}_{n,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk} - \mu_{\mathsf{h}}, \tau_{r,nk} + \tau_{\mathsf{h}})} \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}}) \right]. \tag{APX.131}$$

V.4 Estimates

Once the beliefs are derived, one can compute estimates of the system variables. We start with the estimates of the activity probabilities

$$\hat{q}_n = \mathbb{E}\left[\mathsf{q}_n; \underset{q_n}{\mathfrak{B}}(q_n)\right] \tag{APX.132}$$

$$= \frac{\int q_n \underset{q_n}{\mathfrak{B}}(q_n) \mathrm{d}q_n}{\int \underset{q_n}{\mathfrak{B}}(q_n) \mathrm{d}q_n} \tag{APX.133}$$

where the normalization ensures that the belief is a pdf. Substituting the belief inside the integrals leads to

$$\hat{q}_n \propto \int q_n f_{\mathsf{q}}(\boldsymbol{q}) \left[\prod_{n' \in [N]} \left((1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'} \right) \right] \mathrm{d}\boldsymbol{q}. \tag{APX.134}$$

The computation of such integral cannot be done analytically and is performed numerically, using Monte-Carlo integration. Hence, sampling S correlated vectors $\{\boldsymbol{q}_s\}_{s \in [S]}$ with Algo. E.1, allows to approximate the

estimate by

$$\hat{q}_n \approx \frac{\mathrm{Num}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}\}_{n'\in[N]}, \{\phi_{1,n'}\}_{n'\in[N]})}{\mathrm{Den}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}\}_{n'\in[N]}, \{\phi_{1,n'}\}_{n'\in[N]})} \tag{APX.135}$$

where the numerator is

$$\mathrm{Num}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}, \phi_{1,n'}\}_{n'\in[N]}) = \frac{1}{S}\sum_{s\in[S]} q_{n,s}\left[\prod_{n'\in[N]} ((1-q_{n',s})\phi_{0,n'} + q_{n',s}\phi_{1,n'})\right] \tag{APX.136}$$

and the denominator is

$$\mathrm{Den}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}, \phi_{1,n'}\}_{n'\in[N]}) = \frac{1}{S}\sum_{s\in[S]}\left[\prod_{n'\in[N]} ((1-q_{n',s})\phi_{0,n'} + q_{n',s}\phi_{1,n'})\right]. \tag{APX.137}$$

The estimate of the states is performed using individual log-likelihood ratio tests of the form

$$\hat{s}_n = \mathbb{1}\left(\mathsf{LLR}_n > 0\right) \tag{APX.138}$$

where

$$\mathsf{LLR}_n = \log \hat{q}_{n,n} + \log \phi_{0,n} - \log\left(1-\hat{q}_{n,n}\right) - \log \phi_{1,n} \tag{APX.139}$$

The log-likelihood ratio requires the computation of

$$\mathbb{E}\left[\mathsf{q}_n; \underset{f_{\mathbf{q}}\to\mathsf{q}_n}{\mathfrak{M}}\right] \propto \int_{[0,1]^N} q_n f_{\mathbf{q}}(\boldsymbol{q})\left[\prod_{n'\in[N]\setminus\{n\}} ((1-q_{n'})\phi_{0,n'} + q_{n'}\phi_{1,n'})\,\mathrm{d}q_{n'}.\right] \tag{APX.140}$$

which slightly differs from Eq. APX.134 since the product in the integral does not account for the term indexed by n. This expectation is also approximated using Monte-Carlo integration

$$\mathbb{E}\left[\mathsf{q}_n; \underset{f_{\mathbf{q}}\to\mathsf{q}_n}{\mathfrak{M}}\right] \approx \frac{\mathrm{Num}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}\}_{n'\in[N]\setminus\{n\}}, \{\phi_{1,n'}\}_{n'\in[N]\setminus\{n\}})}{\mathrm{Den}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}\}_{n'\in[N]\setminus\{n\}}, \{\phi_{1,n'}\}_{n'\in[N]\setminus\{n\}})} \tag{APX.141}$$

with numerator

$$\mathrm{Num}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}, \phi_{1,n'}\}_{n'\in[N]\setminus\{n\}}) = \frac{1}{S}\sum_{s\in[S]} q_{n,s} \prod_{n'\in[N]\setminus\{n\}} ((1-q_{n',s})\phi_{0,n'} + q_{n',s}\phi_{1,n'}) \tag{APX.142}$$

and denominator

$$\mathrm{Den}(\{\boldsymbol{q}_s\}_{s\in[S]}, \{\phi_{0,n'}, \phi_{1,n'}\}_{n'\in[N]\setminus\{n\}}) = \frac{1}{S}\sum_{s\in[S]} \prod_{n'\in[N]\setminus\{n\}} ((1-q_{n',s})\phi_{0,n'} + q_{n',s}\phi_{1,n'})\,. \tag{APX.143}$$

Finally, the channel estimation is performed by computing

$$\hat{h}_{nk} = \frac{\int_{\mathbb{C}} h\, \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h}{\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h} \tag{APX.144}$$

from Eq. APX.131. We start by computing the denominator as

$$\begin{aligned}\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h =& \frac{(1-\hat{q}_{n,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})}\int_{\mathbb{C}} \delta(h)\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \\ &+ \frac{\hat{q}_{n,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk}-\mu_{\mathsf{h}}, \tau_{r,nk}+\tau_{\mathsf{h}})}\int_{\mathbb{C}} \mathcal{CN}(h; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \end{aligned} \tag{APX.145}$$

$$= (1-\hat{q}_{n,n})\,\phi_{0,n} + \hat{q}_{n,n}\phi_{1,n} \tag{APX.146}$$

The numerator is

$$\begin{aligned}\int_{\mathbb{C}} \underset{h_{nk}}{\mathfrak{B}}(h)\mathrm{d}h =& \frac{(1-\hat{q}_{n,n})\,\phi_{0,n}}{\mathcal{CN}(0; r_{nk}, \tau_{r,nk})}\int_{\mathbb{C}} h\delta(h)\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \\ &+ \frac{\hat{q}_{n,n}\phi_{1,n}}{\mathcal{CN}(0; r_{nk}-\mu_{\mathsf{h}}, \tau_{r,nk}+\tau_{\mathsf{h}})}\int_{\mathbb{C}} h\mathcal{CN}(h; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})\mathcal{CN}(h; r_{nk}, \tau_{r,nk})\mathrm{d}h \end{aligned} \tag{APX.147}$$

$$= \hat{q}_{n,n}\phi_{1,n}\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1} r_{nk}}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}. \tag{APX.148}$$

Finally

$$\hat{h}_{nk} = \frac{1}{1+e^{-\mathrm{LLR}_n}}\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1} r_{nk}}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}} \tag{APX.149}$$

For the purpose of the updates of the GAMP part of the algorithm, we also compute the corresponding variance

$$\tau_{h,nk} = \mathbb{V}\left[\mathsf{h}_{nk}; \underset{h_{nk}}{\mathfrak{B}}\right] = \mathbb{E}\left[\left|\mathsf{h}_{nk}\right|^2; \underset{h_{nk}}{\mathfrak{B}}\right] - \left|\mathbb{E}\left[\left|\mathsf{h}_{nk}\right|; \underset{h_{nk}}{\mathfrak{B}}\right]\right|^2 = \mathbb{E}\left[\left|\mathsf{h}_{nk}\right|^2; \underset{h_{nk}}{\mathfrak{B}}\right] - \left|\hat{h}_{nk}\right|^2 \tag{APX.150}$$

The second moment of h_{nk} computation is similar to the mean computation, leading to

$$\mathbb{E}\left[\left|\mathsf{h}_{nk}\right|^2; \underset{h_{nk}}{\mathfrak{B}}\right] = \frac{1}{1+e^{-\mathsf{LLR}_n}}\left(\left|\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1}\mu_r}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}\right|^2 + \frac{1}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}\right). \tag{APX.151}$$

Finally, the variance reads

$$\tau_{h,nk} = \frac{1}{1+e^{-\mathsf{LLR}_{nk}}}\left(\left|\frac{\tau_{\mathsf{h}}^{-1}\mu_{\mathsf{h}} + \tau_{r,nk}^{-1}\mu_r}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}\right|^2 + \frac{1}{\tau_{\mathsf{h}}^{-1} + \tau_{r,nk}^{-1}}\right) - \left|\hat{h}_{nk}\right|^2. \tag{APX.152}$$

Résumé en français

I Introduction

I.1 Vue d'ensemble des réseaux sans-fil de nouvelles générations de la 5G à la 6G

Les télécommunications électromagnétiques sont soumises à un processus de standardisation depuis le début des années 1990, orchestré par des organismes internationaux tels que le 3GPP, l'ETSI, l'IEEE et l'ITU. Leur histoire a été marquée par les arrivées successives des générations de réseaux mobiles avec la 2G (1991), la 3G (2001), la 4G (2009) et, récemment, la 5G, chacune permettant la mise en œuvre de nouveaux services soutenus par des technologies de rupture.

La 5G s'appuie par exemple sur l'OFDM, une technologie introduite pour la 4G qui s'est avérée être d'une telle efficacité qu'elle est présente au sein de n'importe quel système de communication sans-fil. Dernière génération à avoir été standardisée [3], [4], elle a été conçue avec l'objectif de développer à la fois les communications pour les particuliers et l'industrie. Ainsi, les indicateurs de performances historiques tels que le débit de donnée utilisateur ou l'efficacité spectrale ne sont plus les seuls à être considérés comme étant de première importance.

Avec l'avènement et le déploiement mondial des réseaux 5G, il est également temps pour l'industrie et la recherche de commencer à penser le futur des télécommunications, au-delà de la 5G, voire même de la 6G.

I.2 Nouveaux cas d'usages

Afin de familiariser le lecteur ou la lectrice avec le contexte technologique de cette thèse, un panorama des cas d'usages de la 5G, B5G et 6G est proposé dans ce qui suit.

I.2.a Cas d'usage en 5G et B5G

Les services technologiques de la 5G et B5G s'articulent autour de trois grands cas d'usages: eMBB, uRLLC et mMTC.

Le premier, eMBB, est la suite direct de ce qui était déjà proposé en 4G-LTE, à savoir le dévelopemment et l'amélioration des systèmes de communications à très haut débit. On y trouve des applications dans les réseaux à très grande mobilité, les services de diffusion vidéo à très haute résolution vers des appareils mobiles ou encore la réalité augmentée et/ou virtuelle.

Le second cas d'usage est celui de la fiabilité. Ainsi, les services estampillés uRLLC sont ceux pour lesquelles il est capital de garantir des transmissions aux latences très faibles tout en diminuant leur taux d'erreur à des niveaux extrêmement stricts. Les applications concernées sont dites critiques et englobent l'IIoT, les véhicules autonomes (V2X et UAV) ou encore l'haptique pour la télésanté.

Le troisième cas d'usage porte sur les mMTC où les applications sont celles des communications massives entre machines. Contrairement à l'eMBB, le trafic des équipements connectés est sporadique du fait de communications opportunistes visant à économiser de l'énergie de fonctionnement. Les villes intelligentes, l'IoT et les capteurs santé forment la majorité du parc applicatif des mMTC.

I.2.b Cas d'usage en 6G

Bien que les cas d'usage de la 6G ne soient pas encore arrêtés, il semble être communément accepté qu'elle consistera soit à pallier les manquements technologiques de la 5G, soit à permettre le développement de services hybridants les usages de la 5G.

En particulier, deux hybridations sont envisagées. La première est

Indicateur	Valeur (vs. réseaux actuels)		
	5G	B5G	6G
Densité de connexion	1e6 périph./km^2 (10×)		
Latence	5 ms (0.5×)	1 ms (0.1×)	< 1 ms (< 0.1×)
Mobilité	500 km/h (+33%)		
Débit données max.	10 Gbps (10×)	100 Gbps (100×)	1 Tbps (1000×)
Fiabilité	99.999%	99.9999%	99.99999%
Efficacité spectrale et énergétique	10× bps/Hz/m^2/J (surfacique)	100× bps/Hz/m^2/J (surfacique)	1000× bps/Hz/m^3/J (volumétrique)
Débit données utilisateur expérimenté	100 Mbps (10×)	1 Gbps (100×)	10 Gbps (1000×)

Table H.1: Indicateurs de performances de la 5G à la 6G par rapport aux réseaux 4G actuels.

celle de l'uRLLC massif à l'intersection de l'uRLLC et des mMTC qui permettra le développement de réseaux IoT critiques à des échelles bien plus larges. La seconde hybridation possible combine l'eMBB et l'uRLLC afin de fortement fiabiliser les applications eMBB susmentionnées pour en améliorer l'expérience utilisateur.

Concernant les cas d'usage natif de la 6G, ils concernent principalement le développement de télécommunications non-conventionnelles. On peut citer pêle-mêle les communications satellitaires, les communications par lumière visible et les communications quantiques. À cela s'ajoute une intégration de l'intelligence artificielle comme composante essentielle au fonctionnement des réseaux.

I.2.c Indicateurs de performances

Le développement de la 5G jusqu'à la 6G est guidé par l'atteignabilité d'indicateurs de performances. Très grossièrement, on peut retenir que les 5G, B5G et 6G sont respectivement prévues pour être 10, 100 et 1000 fois la 4G, comme décrit dans Tab. H.1.

I.3 Accès multiple pour les réseaux sans-fil de nouvelle génération

Maintenant que le contexte technologique est mis en place, le regard est porté sur une facette importante des réseaux sans-fil, à savoir l'accès multiple (MA) d'équipements connectés à un point d'accès (AP) dans le but final de transmettre des données.

I.3.a Accès multiple orthogonal

De la 2G à la 4G, le MA a reposé sur un le principe fondamental des communications dites orthogonales. Du domaine temporel au domaine fréquentiel, en passant par des techniques d'accès par code, l'accès multiple orthogonal (OMA) exploite la séparation des ressources de commu-

nication pour permettre aux UEs de transmettre sans interférer les uns avec les autres.

Bien que cette approche soit très efficace pour séparer un nombre raisonnable d'UEs, i.e. tant que les ressources sont suffisantes pour ne pas être partagées entre plusieurs UEs, elle se heurte à l'augmentation conséquente de ces derniers avec le déploiement des réseaux 5Gs. L'OMA n'est donc pas taillé pour soutenir les cas d'usages de nouvelle génération, rendant nécessaire le dévelopemment d'un accès multiple non-orthogonal (NOMA).

I.3.b Accès multiple non-orthogonal

Le NOMA est un changement de paradigme important dans la conception des réseaux sans-fil. Contrairement à l'OMA, certaines ressources de communication peuvent être partagées entre les UEs. Il faut entendre par là qu'une même ressource (intervalle de temps, fréquence, puissance d'émission, codes, ...) peut être utilisée simultanément par plusieurs UEs. Les techniques NOMA sont aussi nombreuses qu'il existe de ressources de communication (ou de leurs combinaisons). On retiendra par exemple les techniques basées sur

- la puissance de transmission des UEs associées à un décodage d'annulation successive d'interférences;
- les codes ou séquences d'identification pseudo-aléatoires à faible densité;
- la théorie de l'acquisition comprimée (CS).

Ce sont ces dernières techniques d'accès qui ont été retenues pour cette thèse du fait de leur très bon couplage avec la théorie de l'estimation bayésienne, centrale pour les problèmes qui seront étudiés.

I.4 Défis

Il est maintenant temps d'ouvrir la discussion sur les défis qui sont considérés dans ce manuscrit. D'après la Sec. H.I.2, l'uRLLC et la mMTC sont deux cas d'usages qui diffèrent sensiblement de l'eMBB du fait que leurs applications reposent sur des indicateurs de performance qui ne sont pas centrés sur le débit2.

Pour ces cas d'usages, la procédure d'accès aléatoire (RA) est un problème important posé par le MA. Le RA est une étape importante dans

tous les réseaux sans-fil qui permet aux UEs d'accéder à un APs pour de futures transmissions de données.

Il est donc très difficile à concevoir de telle sorte à ce que les indicateurs de performance (passage à l'échelle, latence, fiabilité) de l'uRLLC et de la mMTC soient respectés. En particulier, il sera justifié dans la Sec. H.II que deux tâches importantes de traitement du signal qui peuvent permettre au RAs d'être plus efficace sont nécessaires:

1. la détection active des utilisateurs, qui permet une identification précise des UEs par un AP;
2. l'estimation du canal, qui permet à un APs d'acquérir des informations à grande échelle sur l'environnement afin d'améliorer la fiabilité de futures transmission en lien descendant.

Ces deux problèmes seront étudiés simultanément sous la dénomination d'AUDaCE dans les Sections H.IV and H.V.

I.5 Contributions

Un résumé des contributions de cette thèse est donné ci-après:

1. Un panorama des techniques bayésiennes de CS utiles pour l'AUDaCE est donné dans la Sec. H.III.
2. Deux nouveaux modèles d'activité pour le RA sont introduits.
   (a) L'activité groupée homogène de la Sec. H.IV.2.b qui repose sur des variables latentes induisant une structure de groupe;
   (b) L'activité groupée hétérogène de la Sec. H.V.2.a qui repose sur une application simple de la théorie des copules pour décrire des structures de dépendance statistique flexibles d'activité;
3. Une approche systématique de l'AUDaCE est exploitée pour un RAs spontané dans le cadre bayésien de la CS avec le développement de deux algorithmes à passage de message approximé généralisé hybride dans les Sections H.IV.3 and H.V.3.

II État de l'art

II.1 Accès aléatoire pour la mMTC et l'uRLLC

Chaque fois qu'un UE souhaite transmettre des données à un AP, il doit s'accorder avec ce dernier pour s'assurer qu'il en aura le droit en complétant au préalable une procédure de RA. L'accès est qualifié d'aléatoire puisque le moment de la première transmission est inconnu de l'AP.

Pour la version *New Radio* de la 5G (5G-NR), il existe pour le moment trois façons d'effectuer un RA. Les deux premières méthodes, que nous qualifions de *non-spontanées*, nécessitent une procédure en 3 ou 4 étapes. Elles sont initiées par l'envoi d'un signal préambule, unique (accès non-contentieux) ou choisi aléatoirement parmi un ensemble de préambules (accès contentieux). S'en suivent des échanges de messages avec l'AP qui aboutiront à l'autorisation (ou non) de l'UE à transmettre des données.

Un inconvénient de ces deux méthodes est justement l'échange de ces messages qui, dans le cadre de réseaux densément peuplés ou nécessitant des temps de latence très faibles, génèrent une surcharge liée au trafic de contrôle au détriment de la transmission de données utiles. Pour cette raison, une troisième méthode de RA, dite *spontanée* (GFRA), est possible et consiste à regrouper les 4 messages de l'approche contentieuse en 2 messages (voir Fig. B.1).

Comme évoquer dans Sec. H.I.3.b, permettre le GFRA pour la mMTC et l'uRLLC passera par l'utilisation de techniques NOMA. En particulier, l'activité parcimonieuse des UEs est un terrain fertile pour l'utilisation de CS-NOMA.

II.2 Modèle pour l'accès aléatoire spontané

II.2.a Transmission en bande de base

On considère que la transmission du préambule de l'UE $n \in [N]$ se fait pendant une fenêtre temporelle durant M symboles OFDMs. On suppose que les N UEs transmettent sur la même fréquence, comme cela peut être le cas dans les réseaux IoTs exploitant la technologie NB-IoTs. En notant $\mathbf{x}_n \in \mathbb{C}^M$ le signal en bande de base transmis par l'UE n, la matrice des N signaux est notée

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix} \in \mathbb{C}^{M \times N}. \tag{FR.1}$$

Cette matrice est transmise aux K antennes de l'AP au travers d'un canal à évanouissement lent. En notant h_{kn} le coefficient de canal entre l'UE n et l'antenne k, w_{mk} le coefficient de bruit sur l'antenne k pendant le symbole OFDM m, l'équation de transmission en bande de base s'écrit

$$\mathbf{Z} = \mathbf{XH} \in \mathbb{C}^{M\times K} \qquad \text{(sans bruit)}, \tag{FR.2a}$$

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \in \mathbb{C}^{M\times K} \qquad \text{(avec bruit)}. \tag{FR.2b}$$

II.2.b Modèles d'activité

États d'activité et motifs Pendant chaque intervalle de M symboles OFDM, les N UEs peuvent être séparés selon leur *état*: actif ou inactif. Un UE est dit actif s'il initie une procédure de RA, et inactif sinon. On associe à chaque état une variable aléatoire binaire s_n telle que

$$\begin{cases} \mathsf{s}_n = 0 & \Rightarrow \text{UE est inactif} \\ \mathsf{s}_n = 1 & \Rightarrow \text{UE est actif} \end{cases} \tag{FR.3}$$

L'état de tous les UEs est noté

$$\mathbf{s} = \begin{bmatrix} \mathsf{s}_1 & \dots & \mathsf{s}_N \end{bmatrix}^{\mathsf{T}}. \tag{FR.4}$$

que l'on nommera *motif* d'activité.

Le motif étant supposé aléatoire entre chaque fenêtre de transmission, la modélisation de sa pmf $\mathbb{P}_{\mathbf{s}}$ est importante. La litérature en comporte quelques exemples de caractérisation, brièvement décrits dans Tab. H.2.

Activité corrélée Nombre d'applications (V2X, IIoT, smart cities, etc) faisant intervenir le GFRA devraient exhiber des motifs d'activité corrélés, simplement dû au grand nombre d'UEs pouvant communiquer simultanément. Cependant, les modèles existants d'activité ne sont pas toujours adaptés à ces applications.

En effet, bien qu'ils permettent de couvrir des scénarios d'activité relativement différents, ils décrivent l'activité conjointe d'une façon paramétriquement simple mais limitée dans la dépendance statistique des états ou au contraire d'une façon paramétriquement très complexe pour des motifs à très forte corrélation, ce qui n'est pas adapté pour des réseaux de grande ampleur.

Il est donc intéressant de développer de nouveaux modèles qui pour-

Modèle	Probabilité $\mathbb{P}_{\mathbf{s}}(\boldsymbol{s})$
	Description
Activité indépendante	$\prod_{n\in[N]} \mathbb{P}_{\mathbf{s}_n}(s_n)$
	Une variable d'état pour chaque UE, indépendante des autres.
Activité de groupe indépendant	$\prod_{g=1}^{G} \mathbb{P}_{\mathbf{s}_g^{\mathfrak{G}}}(s_g^{\mathfrak{G}}) \prod_{n\in\mathfrak{G}_g} \delta(s_n - s_g^{\mathfrak{G}})$
	Les UEs sont répartis dans G groupes indépendants les uns des autres. L'état des UEs d'un groupe est régie par l'état de ce dernier $\mathbf{s}_g^{\mathfrak{G}}$.
Activité de groupe	$\sum_{\boldsymbol{s}^{\mathfrak{G}}\in\{0,1\}^G} \left(\prod_{n=1}^{N} \delta(s_n - \max_{g\in\mathfrak{N}_n}(s_g^{\mathfrak{G}}))\right) \left(\prod_{g=1}^{G} \mathbb{P}_{\mathbf{s}_g^{\mathfrak{G}}}(s_g^{\mathfrak{G}})\right)$
	Les UEs sont répartis dans G groupes, pouvant avoir des recouvrements. Un UE est actif si l'état d'un des groupes desquels il dépend est actif; il est inactif sinon.
Activité de Bernoulli multivariée	$\prod_{\boldsymbol{s}'\in\{0,1\}^N} p(\boldsymbol{s}') \prod_{n=1}^{N} s_n^{s'_n} (1-s_n)^{(1-s'_n)}$
	Ce modèle est le plus général de tous mais nécessite de spécifier les 2^N probabilités correspondant à chaque motif dans $\{0,1\}^N$.

Table H.2: Différentes caractérisations probabilistes du motif d'activité.

ront capturer efficacement ces motifs d'activité tout en assurant une flexibilité et une relative simplicité dans leur paramétrisation.

II.3 Construction de la matrice des préambules

En notant $\boldsymbol{p}_n$ le préambule de l'UE $n \in [N]$ et $\boldsymbol{P} = [\boldsymbol{p}_n]_{n\in[N]}$ la matrice des préambules, on peut modéliser la réception du signal reçu pendant une fenêtre de GFRA à l'aide de Eq. FR.2

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \in \mathbb{C}^{M\times K} \qquad \text{avec} \qquad \mathbf{Z} = \boldsymbol{P}\mathbf{H}. \tag{FR.5}$$

Dans le contexte du CS-NOMA, la construction de $\boldsymbol{P}$ doit vérifier la propriété d'isométrie restreinte (RIP)

$$\forall \boldsymbol{h} \in \mathbb{C}^N,\ \|\boldsymbol{h}\|_0 \leq \nu,\ (1-\delta_\nu)\|\boldsymbol{h}\|_2^2 \leq \|\boldsymbol{P}\boldsymbol{h}\|_2^2 \leq (1+\delta_\nu)\|\boldsymbol{h}\|_2^2 \tag{FR.6}$$

où $\nu \in [N]$ est le niveau de parcimonie et $\delta_\nu > 0$ est la constance d'isométrie restreinte (RIC). Cependant, vérifier qu'une matrice vérifie la RIP est connu pour être un problème NP-difficile, conduisant à plutôt

considérer sa *cohérence*

$$\mu(\boldsymbol{P}) = \max_{1 \leq n < n' \leq N} \left| \boldsymbol{p}_n^{\mathsf{H}} \boldsymbol{p}_{n'} \right|. \tag{FR.7}$$

En particulier, les matrices *Grassmanniennes* [40], [41], solutions du problème de conditionnement

$$\boldsymbol{P}^{\star} = \underset{\boldsymbol{P} \in \mathbb{C}^{M \times N}}{\arg\min}\, \mu(\boldsymbol{P}), \tag{FR.8}$$

ont prouvée être de bonnes candidates dans de nombreux scenarios (communications multi-antennes, multi-utilisateurs et spontanées), en apportant une multitude de technique de construction déterministe.

Par opposition à ses constructions algorithmiques, des constructions aléatoires peuvent être considérées tout en garantissant la RIP. On citera les matrices (sous-)gaussiennes, uniformément sphériques ou de Rademacher.

II.4 Détection d'utilisateur actifs et estimation de canal

Quand on considère le GFRA, deux problèmes se posent naturellement: la détection d'utilisateur actifs et l'estimation de canal (AUDaCE). À partir de l'observation du signal $\mathbf{Y} = \boldsymbol{Y}$ et la connaissance des préambules $\boldsymbol{P}$, l'AP doit déterminer quels sont les UEs qui effectuent un RA tout en estimant les coefficients de canal entre les UEs et ses antennes.

Deux formulations conjointes de ces deux problèmes existent. La première est celle d'une estimation MAP de la forme suivante

$$\boldsymbol{s}^{\star}, \boldsymbol{H}^{\star} = \underset{\boldsymbol{s} \in \{0,1\}^N, \boldsymbol{H} \in \mathbb{C}^{N \times K}}{\arg\max} f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}). \tag{FR.9}$$

La seconde approche consiste à définir la matrice par blocs $\mathbf{V}(\mathbf{s},\mathbf{H}) = [\mathbf{s},\mathbf{H}]$ et son estimateur MMSE

$$\boldsymbol{V}(\mathbf{s},\mathbf{H})^{\star} = \mathbb{E}[\mathbf{V}(\mathbf{s},\mathbf{H})] = \int_{\mathbb{C}^{N \times K}} \sum_{\boldsymbol{s} \in \{0,1\}^N} \boldsymbol{V}(\boldsymbol{s},\boldsymbol{H}) f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s},\boldsymbol{H} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{H}. \tag{FR.10}$$

Dans les deux cas, l'estimateur nécessite la connaissance de la pdf a

posteriori

$$f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s},\boldsymbol{H}\mid\boldsymbol{Y}) = f_{\mathbf{Y}}(\boldsymbol{Y})^{-1} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y}\mid\boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H}\mid\boldsymbol{s})\mathbb{P}_{\mathbf{s}}(\boldsymbol{s}) \tag{FR.11}$$

$$= \frac{f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y}\mid\boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H}\mid\boldsymbol{s})\mathbb{P}_{\mathbf{s}}(\boldsymbol{s})}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} f_{\mathbf{Y}|\mathbf{H},\mathbf{s}}(\boldsymbol{Y}\mid\boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H}\mid\boldsymbol{s})\mathbb{P}_{\mathbf{s}}(\boldsymbol{s})\mathrm{d}\boldsymbol{H}} \tag{FR.12}$$

dont la factorisation suppose que les variables du système $\mathbf{s},\mathbf{H},\mathbf{Y}$ forment la chaîne de Markov suivante

$$\mathbf{s} \to \mathbf{H} \to \mathbf{Y}. \tag{FR.13}$$

Selon la nature de cette densité, les optimaux de (FR.9) et (FR.10) peuvent être différents. En général, la résolution de ces problèmes de CS est NP-difficile. Elle nécessite donc des méthodes sous-optimales approximant les solutions idéales Eqs. (FR.9) and (FR.10).

II.5 Conclusion

Il apparaît que le GFRA est un catalyseur essentiel à l'uRLLC et la mMTC notamment pour accélérer les transmissions de données supplémentaires en réduisant la surcharge de contrôle. En particulier, l'AUDaCE joue un rôle important dans le GFRA pour identifier de manière fiable les UEs actifs et préparer les transmissions futures en estimant le canal entre l'AP et les UEs. Cependant, l'AUDaCE n'a été prise en compte qu'avec des modèles d'activité individuelle indépendants indépendante ou indépendante par groupe. Dans les Sections H.IV and H.V l'AUDaCE appliquée au GFRA avec des modèles de corrélation plus flexibles sera étudiée, tout en proposant des solutions algorithmiques à sa résolution. Cela sera rendu possible grâce aux techniques de CS-NOMA qui sont présentées dans la Sec. H.III.

III Prérequis algorithmiques d'acquisition comprimée bayésienne

III.1 Acquisition comprimée non-bayésienne

III.1.a Formulation du problème

Le problème d'acquisition comprimée consiste à reconstruire un signal $\boldsymbol{x} \in \mathbb{C}^N$ à partir de l'observation d'un signal $\boldsymbol{y} \in \mathbb{C}^M$. La transformation de $\boldsymbol{x}$ vers $\boldsymbol{y}$ est la plupart du temps exprimée sous la forme générale suivante:

$$\boldsymbol{y} = f(\boldsymbol{A}\boldsymbol{x}) \tag{FR.14}$$

où $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ est une matrice dite de mesure et $f : \mathbb{C}^M \to \mathbb{C}^N$ est une transformation non-linéaire des composantes de son argument. Le cas particulier le plus étudié est le suivant

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \tag{FR.15}$$

pour lequel la mesure du vecteur $\boldsymbol{x}$ par $\boldsymbol{A}$ est perturbée par un bruit $\boldsymbol{w} \in \mathbb{C}M$. Intuitivement, lorsque la dimension de l'espace de mesure M est plus petite que celle de l'espace du signal N, la reconstruction de $\boldsymbol{x}$ à partir d'une observation (bruitée) semble compromise. Cependant, lorsque $\boldsymbol{x}$ possède la particularité d'être *parcimonieux*, i.e. que certaines de ces composantes sont nulles, l'utilisation de cette information peut être exploitée pour reconstruire $\boldsymbol{x}$. Formellement, cela revient à résoudre le problème d'optimisation suivant:

$$\boldsymbol{x}^\star = \underset{x \in \mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_0 \qquad \text{tel que} \qquad \|y - \boldsymbol{A}\boldsymbol{x}\|_p \leq \epsilon \tag{FR.16}$$

où $p > 0$ et $\epsilon > 0$. Sa résolution est NP-difficile.

III.1.b Inventaire de méthodes non-bayésiennes

La litérature sur la CS regorge de méthodes de résolution pour Eq. FR.16. On peut citer les méthodes d'optimisation (BPDN, LASSO), les méthodes gloutonnes (MP, OMP, StOMP, MMP, CoSaMP, FBMP) et les méthodes itératives (IHTA, ISTA). Chacune de ces familles algorithmiques considère

une version modifiée du problème de base, à savoir

$$\boldsymbol{x}^\star = \underset{x\in\mathbb{C}^N}{\arg\min} \|\boldsymbol{x}\|_1 \qquad \text{tel que} \qquad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon \tag{FR.17}$$

qui a l'avantage de proposer une formulation convexe mais sous-optimale, pour $\epsilon > 0$.

III.2 Algorithme à propagation de croyance

La propagation de croyance désigne un ensemble de techniques d'inférence bayésienne se basant sur l'échange de messages entre les sommets d'un graphe de facteurs.

III.2.a Graphe de facteurs

Un *graphe de facteurs* $\mathfrak{G}(\mathfrak{F}, \mathfrak{V}, \mathfrak{E})$ est un graphe multipartite non-orienté qui représente les dépendences fonctionnelles des facteurs d'une fonction à ses variables.

Soit une fonction f dépendendant des variables $\boldsymbol{v} = [v_1, \ldots, v_I]$ et dont une factorisation possible est

$$f(\boldsymbol{v}) = \prod_{j\in[J]} f_j(\boldsymbol{v}_j) \tag{FR.18}$$

où les J sous-ensembles de variables $\{\boldsymbol{v}_j\}_{j\in[J]}$ vérifient

$$\begin{cases} \forall j \in [J],\ \boldsymbol{v}_j \subseteq \boldsymbol{v} \\ \bigcup_{j\in[J]} \boldsymbol{v}_j = \boldsymbol{v} \end{cases} \tag{FR.19}$$

avec de possibles recouvrements entre les sous-ensembles. En notant $\mathfrak{F} = \{f_j\}_{j\in[J]}$ l'ensemble des sommets *facteurs* et $\mathfrak{V} = \{v_i\}_{i\in[I]}$ l'ensemble des sommets *variables*, on peut construire l'ensemble des arêtes $\mathfrak{E} = \{(f_j, v_i) \mid \forall (i,j) \in [I]\times[J],\ v_i \in \boldsymbol{v}_j\}$. Ces trois composantes essentielles forment le graphe de facteurs de la fonction f, dans lequel deux sommets facteur f_j et variable v_i sont reliés par une arête si et seulement si le facteur f_j admet v_i comme l'une de ses variables. De façon général, une fonction peut être représentée par autant de graphe qu'elle possède de factorisation, le minimum étant nécessairement de un graphe, celui représentant la fonction non-factorisée.

III.2.b Inférence bayésienne

S'il existe un domaine pour lequel les graphes de facteurs possède une utilisation naturelle, c'est celui de l'inférence bayésienne. Elle englobe l'ensemble des méthodes déterminant la validité d'hypothèses probabilistes appliquées à un système à partir d'observation de ce dernier. Dans le cadre du GFRA, deux problèmes d'inférence bayésienne ont été identifiés, à savoir l'estimation du canal de communication et la détection d'utilisateurs actifs à partir de l'observation du signal reçu.

Parmi l'ensemble des méthodes d'inférence bayésienne, on en retiendra trois.

Maximum a posteriori C'est la méthode d'inférence pour décider de l'hypothèse la plus probable. Elle cherche donc à déterminer l'hypothèse qui *maximise* la probabilité ou vraisemblance d'une observation selon cette hypothèse. Formellement, elle s'écrit

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}} f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y}) \tag{FR.20}$$

où $\mathbf{x}$ désigne la variable aléatoire de l'hypothèse et $\mathbf{y}$ celle de l'observation.

Erreur au carré moyenne minimum Cette méthode d'inférence cherche l'hypothèse qui *minimisera* le carré de l'erreur en moyenne. En considérant que la variable d'hypothèse est à valeur (scalaire, vectorielle ou matricielle) dans $\mathbb{C}$, l'hypothèse choisit sera celle vérifiant

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} \mathbb{E}[\|\mathbf{x} - \boldsymbol{x}\|_2 \mid \mathbf{y} = \boldsymbol{y}] \tag{FR.21}$$

qui, sous les conditions que $\mathbb{E}[\|\mathbf{x}\|_2]$ et $\mathbb{E}\left[\|\mathbf{x}\|_2^2\right]$ existent et sont finies, peut s'écrire

$$\boldsymbol{x}^\star = \mathbb{E}[\mathbf{x} \mid \mathbf{y} = \boldsymbol{y}]\,. \tag{FR.22}$$

qui n'est rien d'autre qu'une moyenne a posteriori.

Inférence variationnelle Contrairement à l'estimation MAP ou MMSE, l'inférence variationnelle ne cherche pas à estimer directement la meilleure hypothèse selon un critère particulier mais plutôt la vraisemblance de toutes les hypothèses, i.e. la distribution a posteriori des hypothèses. L'inférence variationnelle cherche donc la distribution qui sera la plus

proche de la vraisemblance théorique en résolvant le problème suivant

$$g^\star = \underset{g\in\mathcal{D}}{\arg\min}\, \mathbb{KL}\left[g \middle\| f_{\mathbf{x}|\mathbf{y}}\right] \qquad \text{telle que} \qquad \mathbf{y} = \boldsymbol{y}. \tag{FR.23}$$

III.2.c Approximation par propagation de croyance

Pour chacun des problèmes d'inférence présentés au-dessus, la distribution a posteriori des hypothèses $f_{\mathbf{x}|\mathbf{y}}$ intervient. Cette densité étant la plupart du temps inconnue analytiquement, en utilisant le très célèbre théorème de Bayes, il est possible de considérer les densités a priori et de vraisemblance qui ont l'avantage d'être plus facilement connues en s'appuyant sur la nature physique du système considéré. Ainsi, on préfera effectuer le remplacement suivant

$$f_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x})}{\int f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y} \mid \boldsymbol{x}) f_{\mathbf{x}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}. \tag{FR.24}$$

Il est ensuite possible de construire le graphe de facteur basé sur la factorisation de Bayes et de considérer l'algorithme à propagation de croyance correspondant. Ce dernier consiste à échanger des messages entre les nœuds variables et facteurs. Les messages sont proportionnels à des densités de probabilité et s'écrivent comme suit:

Message d'un nœud facteur vers un nœud variable Il consiste en la marginalisation de l'agrégation de tous les messages des nœuds variables reçus par le nœud facteurs excepté le message du nœud variable destinataire. Selon la nature du problème d'inférence considéré (MAP, MMSE ou VI), le message aura la forme suivante:

$$\underset{f_{\mathbf{v}}\to \mathsf{v}_n}{\mathfrak{M}}\left(v_n\right) \propto \begin{cases} \max\limits_{\boldsymbol{v}_{\setminus n}} f_{\mathbf{v}}(\boldsymbol{v}) \prod\limits_{n'\in[N]\setminus\{n\}} \underset{f_{\mathbf{v}}\leftarrow \mathsf{v}_{n'}}{\mathfrak{M}}\left(v_{n'}\right) & \text{(MAP)} \\ \int f_{\mathbf{v}}(\boldsymbol{v}) \prod\limits_{n'\in[N]\setminus\{n\}} \underset{f_{\mathbf{v}}\leftarrow \mathsf{v}_{n'}}{\mathfrak{M}}\left(v_{n'}\right) \mathrm{d}v_{n'} & \text{(MMSE)} \\ \dfrac{\mathrm{Proj}\left(\int f_{\mathbf{v}}(\boldsymbol{v}) \prod_{n'\in[N]} \underset{f_{\mathbf{v}}\leftarrow \mathsf{v}_{n'}}{\mathfrak{M}}\left(v_{n'}\right) \mathrm{d}v_{n'}\right)}{\underset{f_{\mathbf{v}}\leftarrow \mathsf{v}_n}{\mathfrak{M}}\left(v_n\right)} & \text{(VIEF)} \end{cases} \tag{FR.25}$$

où v_n est la variable destinataire du message envoyé par le facteur $f_{\mathbf{v}}(\boldsymbol{v})$.

Message d'un nœud variable à un nœud facteur Ce message est simplement le produit de tous les messages envoyés par les nœuds facteurs voisins excepté celui provenant du facteur destinataire. Formellement, il s'écrit

$$\underset{f_m \leftarrow \mathsf{v}}{\mathfrak{M}}(v) \propto \prod_{m' \in [M] \setminus \{m\}} \underset{f_{m'} \rightarrow \mathsf{v}}{\mathfrak{M}}(v). \tag{FR.26}$$

Croyance d'une variable La croyance d'une variable est la pdf de cette dernière construite à partir de *tous* les messages provenant des nœuds facteurs voisins. Elle s'écrit donc

$$\underset{\mathsf{v}}{\mathfrak{B}}(v) \propto \prod_{m' \in [M]} \underset{f_{m'} \rightarrow \mathsf{v}}{\mathfrak{M}}(v) \tag{FR.27}$$

où la normalisation est implicite. C'est à partir de cette croyance qu'il est possible d'effectuer une estimation approximée MAP ou MMSE de la variable correspondante.

III.3 Algorithmes à passage de messages approximés

L'algorithme AMP a été introduit pour la première fois par les chercheurs Donoho, Maleki et Montanari pour étudié le problème de CS présenté dans Eq. FR.16.

Il a été montré qu'il est une approximation de BP dans le cas où le graphe de facteurs exhibe une structure dense, i.e. où une portion du graphe de facteur consiste en une couche de nœuds variables tous reliés aux nœuds d'une couche de nœuds facteurs. Ce caractère dense permet de faire l'approximation que les messages échangés entre les nœuds correspondent à des pdf de distributions gaussiennes. En combinant ce résultat avec la structure des messages échangés Eqs. (FR.25) and (FR.26), on peut réduire BP à l'échange des moyennes et variances de ces pdf ((voir Algo. C.2), au lieu des distributions initiales.

L'avantage de cette approximation est double. D'une part, il permet une réduction de la complexité de calcul puisqu'on peut se passer de l'étape de représentation de distributions continues et d'autre part, il permet une étude analytique des performances d'AMP dans le cas asymptotique du système Eq. FR.16, i.e. quand M croit linéairement avec N et $M(N)/N \rightarrow \eta < \infty$.

Originellement développé pour Eq. FR.16 avec $f_{\mathbf{y}|\mathbf{x}}$ gaussien, AMP a été généralisé plus tard par Rangan pour donner GAMP (voir Algo. C.3).

Cet généralisation d'algorithme est applicable avec $f_{\mathbf{y}|\mathbf{x}}$ d'une distribution quelconque.

Dans le cas où le graphe de facteurs possède en plus de la partie dense et de ses connexions entre sommets dite *faibles* des connexions entre sommets dite *fortes*, il est possible de combiné BP et GAMP pour donner un algorithme hybride du nom d'HGAMP. Cette hybridation est particulièrement intéressante pour réaliser des tâches d'inférence bayésienne incorporant des contraintes fortes sur les variables concernées par l'inférence. C'est donc cet algorithme qui sera utilisé tout du long des Sections H.IV and H.V.

IV Algorithme hybride généralisé à échange de messages approximés pour l'AUDaCE avec une activité homogène par groupe

IV.1 Introduction

Le développement d'une industrie connectée nécessite le déploiement à large échelle d'une multitude de capteurs sans-fil. Dans le cadre d'un site industriel certains capteurs sont disséminés sur les différentes infrastructure techniques afin d'en surveiller le bon fonctionnement. Dans ce genre de scénario, un des problèmes majeurs est la capacité d'identifier les capteurs signalant des dysfonctionnements.

En effet, lorsque plusieurs capteurs détectent un comportement anormale de la structure observée, ils vont devoir initier une procédure de RA. En particulier, c'est le GFRA qui sera privilégié afin de réduire la latence de cette procédure pour obtenir une réponse rapide de la part d'un système de décision centralisé relié à un AP.

Les équipements industriels étant surveillés par des groupes de capteurs, il est naturel que leur dysfonctionnement entraîne des procédures de GFRA simultanées, ou tout du moins corrélées en activité. C'est donc motivé par ce type de scénario que nous allons étudié le problème d'AUDaCE en introduisant un nouveau modèle d'activité qui tient compte d'une telle corrélation lorsque les capteurs d'un même équipement industriel sont supposés homogènes.

IV.2 Modèle et formulation du problème

IV.2.a Signal reçu

Le signal reçu par l'AP du site industriel s'écrit

$$\mathbf{Y} = \boldsymbol{P}\mathbf{H} + \mathbf{W} \tag{FR.28}$$

où on supposera les distributions suivantes des coefficients des matrices de canal et de bruit

$$\forall (n,k) \in [N] \times [K],\ \mathsf{h}_{nk} \mid \mathsf{s}_n = s \sim \begin{cases} \mathsf{Dirac}(0) & \text{si } s = 0 \\ \mathsf{CNorm}(\mu_{\mathsf{h}}, \tau_{\mathsf{h}}) & \text{si } s = 1 \end{cases} \tag{FR.29a}$$

$$\forall (m,k) \in [M] \times [K],\ \mathsf{w}_{mk} \sim \mathsf{CNorm}(\mu_{\mathsf{w}}, \tau_{\mathsf{w}})\,. \tag{FR.29b}$$

qui donnent les factorisations suivantes des densités

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \delta(h_{nk})^{1-s_n} \mathcal{CN}(h_{nk}; \mu_{\mathsf{h}}, \tau_{\mathsf{h}})^{s_n} \tag{FR.30a}$$

$$f_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{Y} \mid \boldsymbol{Z}) = \prod_{m=1}^{N}\prod_{k=1}^{K} \mathcal{CN}(y_{mk}; z_{mk} + \mu_{\mathsf{w}}, \tau_{\mathsf{w}}). \tag{FR.30b}$$

IV.2.b Motif d'activité homogène par groupe

Pour modéliser une activité homogène par groupe, nous considérons que la distribution des états d'activité est décrite par

$$\forall (n,g) \in [N] \times [G],\ \ \mathsf{s}_n \mid \mathsf{q}_g = q_g \sim \mathsf{Bern}(q_g) \tag{FR.31}$$

où $\{\mathsf{q}_g\}_{g\in[G]}$ forme l'ensemble des probabilités d'activité par groupe. Ainsi, tous les capteurs appartenant au groupe g auront la même probabilité d'être actif, mais pas nécessairement le même état. Sous ces conditions, on peut montrer que la corrélation entre les états de deux capteurs s'écrit

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } g_n \neq g_{n'} \\ \frac{1}{\alpha_g + \beta_g + 1} & \text{if } g_n = g_{n'} = g \end{cases} \tag{FR.32}$$

en supposant que

$$\forall g \in [G], \mathsf{q}_g = \mathsf{Beta}(\alpha_g, \beta_g) \qquad \text{où} \qquad \begin{cases} \alpha_g > 0 \\ \beta_g > 0 \end{cases} . \tag{FR.33}$$

IV.2.c Problème d'AUDaCE

Le problème d'AUDaCE est formulé ainsi

$$\begin{aligned} \boldsymbol{s}^\star, \boldsymbol{H}^\star &= \mathbb{E}[[\boldsymbol{s}, \boldsymbol{H}] \mid \boldsymbol{Y}] \\ &= \int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} [\boldsymbol{s}, \boldsymbol{H}] f_{\mathsf{S},\mathsf{H}|\mathsf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{H}. \end{aligned} \tag{FR.34}$$

où la pdf a posteriori est

$$f_{\mathsf{S},\mathsf{H}|\mathsf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) = \frac{\int_{[0,1]^G} f_{\mathsf{Y}|\mathsf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathsf{H}|\mathsf{S}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathsf{S}|\mathsf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathsf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \int_{[0,1]^G} f_{\mathsf{Y}|\mathsf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathsf{H}|\mathsf{S}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathsf{S}|\mathsf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathsf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{H}}. \tag{FR.35}$$

Cette formulation permet d'explicitement faire apparaître le modèle GHomA grâce à la pdf du vecteur des probabilités d'activité. La résolution de ce problème étant trop compliquée, on utilisera une approximation HGAMP pour approcher sa solution optimale.

IV.3 Algorithme pour l'AUDaCE avec activité hétérogène de groupe

En utilisant la factorisation de la pdf conjointe des variables systèmes, il est possible de déduire un graphe de facteurs associé au modèle de transmission avec GHomA dont une représentation est donnée par la Fig. D.4. S'en suit la construction de l'algorithme à propagation de croyance correspondant dont les messages sont résumés dans la Tab. D.1.

En s'appuyant sur le fait que le graphe de facteurs se décompose en deux parties, l'une dense pour la description du signal reçu et l'autre fortement connectée pour la description de l'activité homogène par groupe, il est possible d'approximer partiellement la propagation de croyance avec HGAMP pour construire l'Algo. D.1. Ce dernier possède une complexité de l'ordre de $O(NMK + NK + \sum_{g=1}^{G} U_g^2)$ où U_g indique la taille du groupe de capteur d'indice $g \in [G]$.

IV.4 Résultats numériques

L'algorithme a été évalué par le moyen de simulations de Monte-Carlo dont les paramètres sont décrits dans la Tab. D.3. Il en ressort que le nouvel algorithme GHomA-HGAMP performe mieux que d'autres algorithmes similaires basés sur GAMP. Cela provient du fait que GHomA-HGAMP bénéficie de l'information de corrélation a priori pour améliorer la détection des capteurs actifs dans le site industriel (2 à 3 fois moins d'erreur de détection selon le scénario) ainsi que l'estimation du canal de communication les séparant de l'AP (5dB de gain).

IV.5 Conclusion

Cette section à introduit un nouveau modèle d'activité corrélée ainsi qu'un algorithme efficace pour l'AUDaCE correspondant à ce modèle. Le modèle repose sur l'utilisation de variables latentes par groupe de capteurs décrivant la probabilité d'activité de chacun d'entre eux, supposant ainsi une certaine homogénéité dans leur capacité de détection d'anomalie. L'algorithme est quant à lui construit à partir des outils de CS bayésien, en particulier HGAMP.

Motivé par les bons résultats de GHomA-HGAMP, il semble naturel de pousser l'étude portant sur l'activité corrélée jusqu'à considérer des réseaux d'UEs pouvant former des groupes d'activité mais cette fois-ci hétérogène.

V Algorithme hybride généralisé à échange de messages approximés pour l'AUDaCE avec une activité hétérogène par groupe

V.1 Introduction

Dans la précédente Sec. H.IV, l'AUDaCE pour le GFRA a été étudié sous le modèle GHomA. Ce dernier est adapté aux réseaux sans-fil dans lesquels les UEs forment des groupes homogènes d'activité. Une telle hypothèse peut être valide selon le type de système considéré. Par exemple, pour un site industriel équipé de capteurs sans-fil dont la qualité de fabrication est la même, l'hypothèse d'homogénéité serait raisonnable. En revanche, il est naturel de penser que parmi l'ensemble de ces capteurs

certains seront amenés à être remplacés à un moment donné par des capteurs ayant les mêmes fonctionnalités mais dont la qualité de fabrication sera différente, entraînant une hétérogénéité dans le parc des UEs. En effet, cette différence de qualité pourra se traduire par des changements significatifs dans la capacité de détection de dysfonctionnement de l'équipement industriel surveillé.

Une telle hétérogénéité dans l'activité des capteurs est difficile à modéliser, motivant le développement d'un modèle à base de copules. Ces dernières sont célèbres pour leur flexibilité et leur capacité à modéliser efficacement des systèmes à dépendance statistique complexe entre ses constituants. L'AUDaCE sera donc formulée dans ce cadre hétérogène et, de nouveau, un algorithme de la famille de HGAMP sera proposé pour la résoudre.

V.2 Modèle et formulation du problème

Le modèle de transmission est le même que dans la Sec. H.IV.2.a. Nous donnons ci-après les éléments importants permettant la mise en place d'un modèle d'activité hétérogène par groupe (GHetA).

V.2.a Motif d'activité hétérogène par group

La première différence avec le modèle GHomA réside dans la modélisation des états qui seront supposés être distribués selon

$$\forall n \in [N],\ \mathsf{s}_n \mid \mathsf{q}_n = \mathsf{Bern}(q_n) \tag{FR.36}$$

où chaque variable d'état s_n est associée à sa propre variable latente de probabilité d'activité $\mathsf{q}_n \sim \mathsf{Beta}(\alpha_n, \beta_n)$.

La théorie des copules permets de modéliser le vecteur des probabilités d'activité selon la transformation suivante

$$\mathbf{c} \to \mathbf{u} \to \mathbf{q} \tag{FR.37}$$

où $\mathbf{c}$ désigne un vecteur dont la structure de corrélation est connue (par ex. gaussien) et $\mathbf{u}$ est un vecteur aléatoire corrélée dont les composantes sont uniformément distribuées sur $[0, 1]$. La conséquence d'une telle modélisation permet d'induire une dépendance statistique entre les composantes

de $\mathbf{q}$ à partir de celle de $\mathbf{c}$. La transformation de $\mathbf{c}$ à $\mathbf{q}$ s'écrit formellement

$$\boldsymbol{T}(\mathbf{c}) = \begin{bmatrix} T_1(\mathsf{c}_1) \\ \vdots \\ T_N(\mathsf{c}_N) \end{bmatrix} = \begin{bmatrix} (F_{\mathsf{q}_1}^{-1} \circ F_{\mathsf{c}_1})(\mathsf{c}_1) \\ \vdots \\ (F_{\mathsf{q}_N}^{-1} \circ F_{\mathsf{c}_N})(\mathsf{c}_N) \end{bmatrix} \tag{FR.38}$$

où $\{F_{\mathsf{c}_n}\}_{n\in[N]}$ et $\{F_{\mathsf{q}_n}\}_{n\in[N]}$ désigne respectivement les ensembles des cdfs marginales des composantes de $\mathbf{c}$ et $\mathbf{q}$.

En considérant un tel modèle, il est possible de donner l'expression de la corrélation entre deux états d'activité quelconque, à savoir

$$\mathrm{Cor}[\mathsf{s}_n, \mathsf{s}_{n'}] = \begin{cases} 1 & \text{if } n = n' \\ 0 & \text{if } s_n \perp s_{n'} \\ \frac{\mathrm{Cor}[\mathsf{q}_n, \mathsf{q}_{n'}]}{\sqrt{(\alpha_n+\beta_n+1)(\alpha_{n'}+\beta_{n'}+1)}} & \text{otherwise} \end{cases}. \tag{FR.39}$$

V.2.b Formulation de l'AUDaCE

Le problème d'AUDaCE est formulé ainsi

$$\begin{aligned} \boldsymbol{s}^\star, \boldsymbol{H}^\star &= \mathbb{E}[[\boldsymbol{s}, \boldsymbol{H}] \mid \boldsymbol{Y}] \\ &= \int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} [\boldsymbol{s}, \boldsymbol{H}] f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{H}. \end{aligned} \tag{FR.40}$$

où la pdf a posteriori est

$$f_{\mathbf{s},\mathbf{H}|\mathbf{Y}}(\boldsymbol{s}, \boldsymbol{H} \mid \boldsymbol{Y}) = \frac{\int_{[0,1]^G} f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}}{\int_{\mathbb{C}^{N\times K}} \sum_{\boldsymbol{s}\in\{0,1\}^N} \int_{[0,1]^G} f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y} \mid \boldsymbol{H}) f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H} \mid \boldsymbol{s}) \mathbb{P}_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s} \mid \boldsymbol{q}) f_{\mathbf{q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}\mathrm{d}\boldsymbol{H}}. \tag{FR.41}$$

Comme pour la formulation avec GHomA, cette formulation permet d'explicitement faire apparaître le modèle GHetA grâce à la pdf du vecteur des probabilités d'activité. De nouveau, on utilisera une approximation HGAMP pour approcher la solution optimale.

V.3 Algorithme pour l'AUDaCE avec activité hétérogène de groupe

Les variables du système forment la chaîne de Markov suivante

$$\mathbf{q} \to \mathbf{s} \to \mathbf{H} \to \mathbf{Y} \tag{FR.42}$$

permettant d'écrire la pdf conjointe a posteriori comme

$$f_{\mathbf{H},\mathbf{s},\mathbf{q}|\mathbf{Y}}(\boldsymbol{H},\boldsymbol{s},\boldsymbol{q}\mid\boldsymbol{Y})=\frac{1}{f_{\mathbf{Y}}(\boldsymbol{Y})}f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y}\mid\boldsymbol{H})f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H}\mid\boldsymbol{s})f_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s}\mid\boldsymbol{q})f_{\mathbf{q}}(\boldsymbol{q}). \quad \text{(FR.43)}$$

Chaque terme au numérateur peut être factorisé comme

$$f_{\mathbf{Y}|\mathbf{H}}(\boldsymbol{Y}\mid\boldsymbol{H})=\prod_{k\in[K]}\prod_{m\in[M]}f_{\mathsf{y}_{mk}|\mathbf{h}_k}(y_{mk}\mid\boldsymbol{h}_k) \quad \text{(FR.44)}$$

$$f_{\mathbf{H}|\mathbf{s}}(\boldsymbol{H}\mid\boldsymbol{s})=\prod_{k\in[K]}\prod_{n\in[N]}f_{\mathsf{h}_{nk}|\mathsf{s}_n}(h_{nk}\mid s_n) \quad \text{(FR.45)}$$

$$f_{\mathbf{s}|\mathbf{q}}(\boldsymbol{s}\mid\boldsymbol{q})=\prod_{n\in[N]}f_{\mathsf{s}_n|\mathsf{q}_n}(s_n\mid q_n) \quad \text{(FR.46)}$$

où $f_{\mathbf{q}}(\boldsymbol{q})$ reste non-factorisé.

En utilisant le graphe de facteurs de la Fig. E.4 pour cette nouvelle factorisation, il est possible d'écrire l'algorithme BP correspondant dans Tab. E.1. Ce dernier comportant une partie à connexion dense et une autre à connexions fortes, l'utilisation d'un algorithme de type HGAMP est naturelle pour approximer les messages de la partie dense. Cela résulte en l'algorithme GHetA-HGAMP, décrit dans par l'Algo. E.2. La complexité de cet algorithme est en $O(NMK+NK+NS)$ où S est le nombre d'échantillons de vecteur $\boldsymbol{q}_j$ corrélés nécessaire aux calculs des intégrales présentes dans Algo. E.2.

V.4 Résultats numériques

GHetA-HGAMP est évalué par simulations de Monte-Carlo, tout d'abord en étant comparé à GHomA-HGAMP et d'autres algorithmes basé sur GAMP, puis en testant sa robustesse avec corrélations biaisées par rapport à la corrélation réelle des scénarios considérés. Les paramètres de chaque simulation sont donnés dans les Tables E.2 and E.3 et les résultats dans les Figs. E.6 to E.9 et Figs. E.10 to E.13.

Il en ressort que GHetA-HGAMP

- est une solution pouvant s'adapter à tous les scénarios de corrélation en atteignant toujours les meilleures performances de détection et d'estimation vis-à-vis de ses concurrents;
- est robuste aux corrélations surestimant une faible vraie corrélation d'activité;

- nécessite de connaître la corrélation dans le cas où la vraie corrélation d'activité est forte pour des gains pouvant monter à 5dB en estimation de canal et diviser par 10 l'erreur de détection.

V.5 Conclusion

Cette section a introduit un nouveau modèle de dépendance statistique pour l'activité d'UEs dans le cas d'une procédure de GFRA. Ce modèle est flexible de part l'utilisation qu'il fait de la théorie des copules tout en proposant une approche systématique aux communications à activité hétérogène corrélée. Cela est confirmé par l'évaluation de l'algorithme GHetA-HGAMP qui a été proposé pour le problème d'AUDaCE.

Parmi les opportunités de travail à explorer suite à ce travail, on notera l'amélioration de la complexité de l'algorithme et l'exploration des différentes copules pour adapter les structures de dépendance statistique aux différents cas d'usage de communication hétérogène dans les réseaux sans-fil.

VI Conclusion et perspectives

VI.1 Aboutissements de la thèse

Cette thèse a étudié le problème du GFRA dans le contexte des mMTC et uRLLC au sein de 5G-NR. Il a été discuté de l'importance pour les systèmes de communication de tirer parti d'AUDaCE pour permettre la GFRA. En utilisant CS-NOMA, il est possible de traiter efficacement la structure des variables du système impliquées dans AUDaCE. En particulier, il a été démontré la capacité du cadre bayésien HGAMP à faire face à l'activité corrélée sous-jacente des UEs.

Tout d'abord, un modèle de modèle GHomA a été considéré au chapitre D. Il a été démontré qu'il permet de généraliser l'activité indépendante par groupe avec l'introduction de variables de groupe latentes associées aux probabilités d'activité de chaque UE au sein des groupes. Sur la base de ce nouveau modèle, un algorithme LBP a été construit avant d'être approximé par une instance de HGAMP, à savoir GHomA-HGAMP. Ce dernier a démontré numériquement son intérêt pour AUDaCE dans certains régimes d'activité face à d'autres algorithmes basés sur GAMP qui ne tiennent pas compte de la corrélation induite par GHomA.

Deuxièmement, un modèle de motif GHetA généralisant le modèle de motif GHomA a été introduit. Cette généralisation s'appuie sur la théorie des copules, ce qui permet de construire des structures de dépendance générales entre les états d'activité des UEs. Avec une approche similaire à celle de GHomA, des instances de LBP et de HGAMP ont été développées, conduisant à GHetA-HGAMP. L'algorithme a montré sa capacité à traiter efficacement AUDaCE dans différents scénarios de corrélation d'activité, alors que les algorithmes GHomA-HGAMP et GAMP modifié se sont avérés efficaces uniquement dans les cas de forte et faible corrélation d'activité.

Parmi les autres directions qui pourraient être envisagées, nous présentons et donnons dans les sections suivantes un aperçu des travaux futurs qui pourraient être réalisés sur la base des travaux de cette thèse.

VI.2 Extension aux communications multi-porteuses OFDM

Pour porter le travail d'AUDaCE à des communications multi-porteuses OFDM, on peut considérer la factorisation suivante du tenseur de canal $\mathcal{H} \in \mathbb{C}^{N \times K \times F}$

$$f_{\mathsf{H}|\mathsf{s}}(\mathcal{H} \mid \boldsymbol{s}; \boldsymbol{B}) = \prod_{n \in [N]} \prod_{k \in [K]} \prod_{f \in [F]} f_{\mathsf{h}_{nkf}|\mathsf{s}_n}(h_{nkf} \mid s_n; b_{nf}) \tag{FR.47}$$

où $\boldsymbol{B} \in \{0,1\}^{N,F}$ est une matrice d'allocation des F fréquences aux N UEs. Le graphe de facteurs de la Fig. E.4 serait alors pourvu de nouveau sommets facteurs et variable dans la partie dense pour chaque nouvelle fréquence ajoutée au modèle. La partie liée à l'activité resterait toutefois la même. Cette tensorisation du modèle n'est donc pas un problème à l'utilisation de GHomA-HGAMP ou GHetA-HGAMP.

VI.3 Transmission de données

Au lieu de considérer le GFRA, on pourrait considérer la transmission de donnée spontanée, i.e. la transmission de données sans attendre une autorisation de la part de l'AP. Un tel changement signifierait que la matrice déterministe des préambules $\boldsymbol{P}$ serait remplacée par une matrice aléatoire contenant des données $\mathbf{X}$. Le problème d'AUDaCE est donc transformé en un problème de DrAUDaCE dont on cherche à estimer conjointement le canal et les données tout en détectant les UEs actifs à

partir d'une transmission de la forme

$$\mathbf{Y} = \mathbf{Z} + \mathbf{W} \qquad \text{with} \qquad \mathbf{Z} = \mathbf{XH}. \tag{FR.48}$$

C'est donc un problème d'inférence bayésienne *bilinéaire* dans la mesure où on cherche à retrouver le produit bruité de deux variables.

Il est donc possible de combiner GHetA-HGAMP avec l'algorithme BiGVAMP pour ajouter la fonctionnalité d'estimation des données.

BIBLIOGRAPHY

[1] 3GPP. "Release 13." (), URL: https://www.3gpp.org/release-13 (cit. on p. 3).

[2] ——, "TR 21.914: Release 14," 3GPP, Jun. 8, 2018, p. 103, URL: https://www.3gpp.org/ftp/Specs/archive/21_series/21.914/ (cit. on p. 3).

[3] ——, "TR 21.915: Release 15," 3GPP, Oct. 1, 2019, p. 118, URL: https://www.3gpp.org/ftp/Specs/archive/21_series/21.915/ (cit. on pp. 3, 15, 17, 18, 19, 20, 153).

[4] ——, "TR 21.916: Release 16," 3GPP, Sep. 14, 2021, p. 157, URL: https://www.3gpp.org/ftp/Specs/archive/21_series/21.916/ (cit. on pp. 3, 15, 17, 153).

[5] ITU, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," p. 21, Sep. 2015, URL: https://www.itu.int/rec/R-REC-M.2083/en (cit. on pp. 5, 8).

[6] A. Dogra, R. K. Jha, and S. Jain, "A Survey on Beyond 5G Network With the Advent of 6G: Architecture and Emerging Technologies," *IEEE Access*, vol. 9, pp. 67 512–67 547, 2021, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2020.3031234 (cit. on pp. 5, 6).

[7] S. K. Ong, M. L. Yuan, and A. Y. C. Nee, "Augmented reality applications in manufacturing: A survey," *International Journal of Production Research*, vol. 46, no. 10, pp. 2707–2742, May 15, 2008, ISSN: 0020-7543, 1366-588X, URL: https://doi.org/10.1080/00207540601064773 (cit. on p. 6).

[8] D. Kamińska, T. Sapiński, S. Wiak, *et al.*, "Virtual Reality and Its Applications in Education: Survey," *Information*, vol. 10, no. 10, p. 318, Oct. 16, 2019, ISSN: 2078-2489, URL: https://doi.org/10.3390/info10100318 (cit. on p. 6).

[9] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile Augmented Reality Survey: From Where We Are to Where We Go," *IEEE Access*, vol. 5, pp. 6917–6950, 2017, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2017.2698164 (cit. on p. 6).

[10] L. P. Berg and J. M. Vance, "Industry use of virtual reality in product design and manufacturing: A survey," *Virtual Reality*, vol. 21, no. 1, pp. 1–17, Mar. 2017, ISSN: 1359-4338, 1434-9957, URL: https://doi.org/10.1007/s10055-016-0293-9 (cit. on p. 6).

[11] P. Popovski, J. J. Nielsen, C. Stefanovic, *et al.*, "Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar. 2018, ISSN: 0890-8044, URL: https://doi.org/10.1109/MNET.2018.1700258 (cit. on p. 6).

[12] P. Popovski, Č. Stefanović, J. J. Nielsen, *et al.*, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019, ISSN: 1558-0857, URL: https://doi.org/10.1109/TCOMM.2019.2914652 (cit. on p. 6).

[13] I. Rodriguez, R. S. Mogensen, A. Fink, *et al.*, "An Experimental Framework for 5G Wireless System Integration into Industry 4.0 Applications," *Energies*, vol. 14, no. 15, p. 4444, 15 Jan. 2021, ISSN: 1996-1073, URL: https://doi.org/10.3390/en14154444 (cit. on p. 6).

[14] A. Alalewi, I. Dayoub, and S. Cherkaoui, "On 5G-V2X Use Cases and Enabling Technologies: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 107 710–107 737, 2021, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2021.3100472 (cit. on p. 6).

[15] A. Masaracchia, Y. Li, K. K. Nguyen, *et al.*, "UAV-Enabled Ultra-Reliable Low-Latency Communications for 6G: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 137 338–137 352, 2021, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2021.3117902 (cit. on p. 6).

[16] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile-Internet-Based Telesurgery System for Healthcare 4.0: An Architecture, Research Challenges, and Future Directions," *IEEE Netw.*, vol. 33, no. 6, pp. 22–29, Nov. 2019, ISSN: 1558-156X, URL: https://doi.org/10.1109/MNET.001.1900063 (cit. on p. 6).

[17] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards Enabling Critical mMTC: A Review of URLLC Within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, 2020, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2020.3010271 (cit. on p. 7).

[18] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020, ISSN: 1558-156X, URL: https://doi.org/10.1109/MNET.001.1900287 (cit. on pp. 7, 8).

[19] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A Speculative Study on 6G," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020, ISSN: 1558-0687, URL: https://doi.org/10.1109/MWC.001.1900488 (cit. on p. 7).

[20] E. Calvanese Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, *et al.*, "6G: The Next Frontier: From Holographic Messaging to Artificial Intelligence Using Subterahertz and Visible Light Communication," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 42–50, Sep. 2019, ISSN: 1556-6080, URL: https://doi.org/10.1109/MVT.2019.2921162 (cit. on p. 7).

[21] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, and M. Asaduzzaman, "Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future," *IEEE Access*, vol. 7, pp. 46 317–46 350, 2019, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2019.2909490 (cit. on p. 7).

[22] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018, ISSN: 1558-2256, URL: https://doi.org/10.1109/JPROC.2018.2867029 (cit. on p. 8).

[23] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive Access for 5G and Beyond," *IEEE J. Select. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021, ISSN: 0733-8716, 1558-0008, URL: https://doi.org/10.1109/JSAC.2020.3019724 (cit. on pp. 10, 18).

[24] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 2294–2323, 2018, ISSN: 1553-877X, URL: https://doi.org/10.1109/COMST.2018.2835558 (cit. on p. 10).

[25] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1805–1838, thirdquarter 2020, ISSN: 1553-877X, URL: https://doi.org/10.1109/COMST.2020.2996032 (cit. on pp. 10, 60).

[26] Y. Yuan, S. Wang, Y. Wu, *et al.*, "NOMA for Next-Generation Massive IoT: Performance Potential and Technology Directions," *IEEE Commun. Mag.*, vol. 59, no. 7, pp. 115–121, Jul. 2021, ISSN: 1558-1896, URL: https://doi.org/10.1109/MCOM.001.2000997 (cit. on p. 10).

[27] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996, ISSN: 1557-9654, URL: https://doi.org/10.1109/18.485709 (cit. on p. 10).

[28] L. Chetot, J.-M. Gorce, and J.-M. Kelif, "Fundamental Limits in Cellular Networks with Point Process Partial Area Statistics," in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, Avignon, France: IEEE, Jun. 2019, pp. 1–8, ISBN: 978-3-903176-20-1, URL: https://doi.org/10.23919/WiOPT47501.2019.9144101 (cit. on p. 12).

[29] D. Duchemin, L. Chetot, J.-M. Gorce, and C. Goursaud, "Détecteur pour l'accès aléatoire massif entre machines avec connaissance statistique du canal en lien ascendant," p. 5, (cit. on p. 12).

[30] D. Duchemin, L. Chetot, J.-M. Gorce, and C. Goursaud, "Coded random access for massive MTC under statistical channel knowledge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, France: IEEE, Jul. 2019, pp. 1–5, ISBN: 978-1-5386-6528-2, URL: https://doi.org/10.1109/SPAWC.2019.8815491 (cit. on p. 12).

[31] L. Chetot, M. Egan, and J.-M. Gorce, "Joint Identification and Channel Estimation for Fault Detection in Industrial IoT with Correlated Sensors," *IEEE Access*, 2021, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2021.3106736 (cit. on p. 12).

[32] 3GPP, "TS 38.211: Physical channels and modulation," 3GPP, Sep. 2021, p. 131, URL: https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/ (cit. on p. 20).

[33] ——, "TR 37.824: Coexistence between NB-IoT and NR," 3GPP, Jun. 2020, p. 46, URL: https://www.3gpp.org/ftp/Specs/archive/37_series/37.824/ (cit. on p. 21).

[34] M. Chafii, F. Bader, and J. Palicot, "SC-FDMA with index modulation for M2M and IoT uplink applications," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona: IEEE, Apr. 2018, pp. 1–5, ISBN: 978-1-5386-1734-2, URL: https://doi.org/10.1109/WCNC.2018.8377028 (cit. on p. 21).

[35] B. Dai, S. Ding, and G. Wahba, "Multivariate Bernoulli distribution," *Bernoulli*, vol. 19, no. 4, pp. 1465–1483, Sep. 2013, ISSN: 1350-7265, URL: https://doi.org/10.3150/12-BEJSP10 (cit. on p. 26).

[36] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, May 2008, ISSN: 1631073X, URL: https://doi.org/10.1016/j.crma.2008.03.014 (cit. on p. 28).

[37] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser New York, NY, ISBN: 978-0-8176-4948-7, URL: https://doi.org/10.1007/978-0-8176-4948-7 (cit. on pp. 28, 35).

[38] A. S. Bandeira, M. Fickus, D. G. Mixon, and P. Wong, "The Road to Deterministic Matrices with the Restricted Isometry Property," *J Fourier Anal Appl*, vol. 19, no. 6, pp. 1123–1149, Dec. 2013, ISSN: 1069-5869, 1531-5851, URL: https://doi.org/10.1007/s00041-013-9293-2 (cit. on p. 28).

[39] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, Eds., *Compressed Sensing and Its Applications*, ser. Applied and Numerical Harmonic Analysis. Cham: Springer International Publishing, 2015, ISBN: 978-3-319-16041-2 978-3-319-16042-9, URL: https://doi.org/10.1007/978-3-319-16042-9 (cit. on p. 28).

[40] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing Lines, Planes, etc.: Packings in Grassmannian Spaces," *Exp. Math.*, vol. 5, no. 2, pp. 139–159, Jan. 1, 1996, ISSN: 1058-6458, URL: https://doi.org/10.1080/10586458.1996.10504585 (cit. on pp. 28, 161).

[41] T. Strohmer and R. W. Heath, "Grassmannian frames with applications to coding and communication," *Applied and Computational Harmonic Analysis*, vol. 14, no. 3, pp. 257–275, May 2003, ISSN: 10635203, URL: https://doi.org/10.1016/S1063-5203(03)00023-X (cit. on pp. 28, 161).

[42] B. Alexeev, J. Cahill, and D. G. Mixon, "Full Spark Frames," *J Fourier Anal Appl*, vol. 18, no. 6, pp. 1167–1194, Dec. 2012, ISSN: 1069-5869, 1531-5851, URL: https://doi.org/10.1007/s00041-012-9235-4 (cit. on p. 28).

[43] M. Fickus and D. G. Mixon, "Tables of the existence of equiangular tight frames," Jun. 16, 2016, URL: http://arxiv.org/abs/1504.00253 (cit. on p. 28).

[44] G. Kutyniok, A. Pezeshki, R. Calderbank, and T. Liu, "Robust dimension reduction, fusion frames, and Grassmannian packings," *Applied and Computational Harmonic Analysis*, vol. 26, no. 1, pp. 64–76, Jan. 2009, ISSN: 10635203, URL: https://doi.org/10.1016/j.acha.2008.03.001 (cit. on p. 28).

[45] J. Tropp, I. Dhillon, R. Heath, and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2004.839492 (cit. on p. 28).

[46] B. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 1962–1973, Sept./2000, ISSN: 00189448, URL: https://doi.org/10.1109/18.868472 (cit. on p. 28).

[47] I. Kammoun and J.-C. Belfiore, "A new family of Grassmann space-time codes for non-coherent MIMO systems," *IEEE Commun. Lett.*, vol. 7, no. 11, pp. 528–530, Nov. 2003, ISSN: 1089-7798, URL: https://doi.org/10.1109/LCOMM.2003.820081 (cit. on p. 28).

[48] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002, ISSN: 0018-9448, URL: https://doi.org/10.1109/18.978730 (cit. on p. 28).

[49] A. Decurninge and M. Guillaud, "Cube-Split: Structured Quantizers on the Grassmannian of Lines," *2017 IEEE Wirel. Commun. Netw. Conf. WCNC*, pp. 1–6, Mar. 2017, URL: https://doi.org/10.1109/WCNC.2017.7925902 (cit. on p. 28).

[50] K. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Design and analysis of a practical codebook for non-coherent communications," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, IEEE, Oct. 2017, pp. 1237–1241, URL: https://doi.org/10.1109/ACSSC.2017.8335549 (cit. on p. 28).

[51] N. Y. Yu, "Design of Non-Orthogonal Sequences Using a Two-Stage Genetic Algorithm for Grant-Free Massive Connectivity," Aug. 1, 2021, URL: http://arxiv.org/abs/2108.00361 (cit. on p. 28).

[52] M. Akrout, A. Housseini, F. Bellili, and A. Mezghani, "Bilinear Generalized Vector Approximate Message Passing," Sep. 14, 2020, URL: http://arxiv.org/abs/2009.06854 (cit. on pp. 32, 57, 128).

[53] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, Jan. 2010, pp. 1–5, URL: https://doi.org/10.1109/ITWKSPS.2010.5503193 (cit. on pp. 32, 39, 49, 51).

[54] Q. Zou, H. Zhang, D. Cai, and H. Yang, "A Low-Complexity Joint User Activity, Channel and Data Estimation for Grant-Free Massive MIMO Systems," *IEEE Signal Process. Lett.*, vol. 27, pp. 1290–1294, 2020, ISSN: 1558-2361, URL: https://doi.org/10.1109/LSP.2020.3008550 (cit. on pp. 33, 60, 128).

[55] A. E. Kalor, O. A. Hanna, and P. Popovski, "Random Access Schemes in Wireless Systems with Correlated User Activity," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jun. 2018, pp. 1–5, URL: https://doi.org/10.1109/SPAWC.2018.8445866 (cit. on p. 33).

[56] C. Zheng, M. Egan, L. Clavier, A. E. Kalør, and P. Popovski, "Stochastic Resource Optimization of Random Access for Transmitters With Correlated Activation," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 3055–3059, Sep. 2021, ISSN: 1558-2558, URL: https://doi.org/10.1109/LCOMM.2021.3090110 (cit. on pp. 33, 60).

[57] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2006.871582 (cit. on p. 35).

[58] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, ISSN: 00359246, URL: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (cit. on p. 37).

[59] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, Jan. 2001, ISSN: 0036-1445, 1095-7200, URL: https://doi.org/10.1137/S003614450037906X (cit. on p. 37).

[60] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006, ISSN: 0010-3640, 1097-0312, URL: https://doi.org/10.1002/cpa.20124 (cit. on p. 37).

[61] E. J. Candes and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2006.885507 (cit. on p. 37).

[62] S. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec./1993, ISSN: 1053587X, URL: https://doi.org/10.1109/78.258082 (cit. on p. 37).

[63] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, Nov. 1993, 40–44 vol.1, URL: https://doi.org/10.1109/ACSSC.1993.342465 (cit. on p. 38).

[64] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1094–1121, Feb. 2012, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2011.2173241 (cit. on p. 38).

[65] Suhyuk Kwon, Jian Wang, and Byonghyo Shim, "Multipath Matching Pursuit," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2986–3001, May 2014, ISSN: 0018-9448, 1557-9654, URL: https://doi.org/10.1109/TIT.2014.2310482 (cit. on p. 38).

[66] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, May 2009, ISSN: 10635203, URL: https://doi.org/10.1016/j.acha.2008.07.002 (cit. on p. 38).

[67] P. Schniter, L. C. Potter, and J. Ziniel, "Fast bayesian matching pursuit," in *2008 Information Theory and Applications Workshop*, San Diego, CA, USA: IEEE, Jan. 2008, pp. 326–333, ISBN: 978-1-4244-2670-6, URL: https://doi.org/10.1109/ITA.2008.4601068 (cit. on pp. 38, 60).

[68] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, Nov. 2009, ISSN: 10635203, URL: https://doi.org/10.1016/j.acha.2009.04.002 (cit. on p. 38).

[69] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004, ISSN: 0010-3640, 1097-0312, URL: https://doi.org/10.1002/cpa.20042 (cit. on p. 38).

[70] K. Bredies and D. A. Lorenz, "Linear Convergence of Iterative Soft-Thresholding," *J Fourier Anal Appl*, vol. 14, no. 5-6, pp. 813–837, Dec. 2008, ISSN: 1069-5869, 1531-5851, URL: https://doi.org/10.1007/s00041-008-9041-1 (cit. on p. 38).

[71] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009, ISSN: 1936-4954, URL: https://doi.org/10.1137/080716542 (cit. on p. 38).

[72] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J Royal Statistical Soc B*, vol. 68, no. 1, pp. 49–67, Feb. 2006, ISSN: 1369-7412, 1467-9868, URL: https://doi.org/10.1111/j.1467-9868.2005.00532.x (cit. on p. 39).

[73] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001, ISSN: 1557-9654, URL: https://doi.org/10.1109/18.910572 (cit. on p. 40).

[74] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Boston, MA: Springer US, 2008, ISBN: 978-0-387-76542-6 978-0-387-76544-0, URL: https://doi.org/10.1007/978-0-387-76544-0 (cit. on pp. 42, 43).

[75] A. Hjorungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*. 2011, ISBN: 978-0-521-19264-4 (cit. on p. 44).

[76] D. J. C. MacKay, D. J. C. M. Kay, and v. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Sep. 25, 2003, 694 pp., ISBN: 978-0-521-64298-9 (cit. on p. 45).

[77] T. P. Minka, "Expectation Propagation for approximate Bayesian inference," version 1, 2001, URL: https://doi.org/10.48550/ARXIV.1301.2294 (cit. on pp. 45, 49).

[78] ——, "A family of algorithms for approximate Bayesian inference," Massachusetts Institute of Technology, 2001, 75 pp., URL: https://tminka.github.io/papers/ep/minka-thesis.pdf (cit. on p. 45).

[79] T. Minka, "Divergence measures and message passing," p. 17, 2005 (cit. on p. 45).

[80] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Gao, "Unifying Message Passing Algorithms Under the Framework of Constrained Bethe Free Energy Minimization," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 7, pp. 4144–4158, Jul. 2021, ISSN: 1558-2248, URL: https://doi.org/10.1109/TWC.2021.3056193 (cit. on p. 45).

[81] J. Pearl, "Reverend bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, ser. AAAI'82, Pittsburgh, Pennsylvania: AAAI Press, Aug. 18, 1982, pp. 133–136, URL: https://aaai.org/Papers/AAAI/1982/AAAI82-032.pdf (cit. on p. 45).

[82] ——, "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence*, vol. 29, no. 3, pp. 241–288, Sep. 1986, ISSN: 00043702, URL: https://doi.org/10.1016/0004-3702(86)90072-X (cit. on p. 45).

[83] Y. Weiss, "Belief Propagation and Revision in Networks with Loops," Massachusetts Institute of Technology, 1997, p. 15, URL: https://dspace.mit.edu/bitstream/handle/1721.1/7249/AIM-1616.pdf?sequence=2 (cit. on p. 49).

[84] A. T. Ihler, J. W. Fischer III, and A. S. Willsky, "Loopy Belief Propagation: Convergence and Effects of Message Errors," *J. Mach. Learn. Res.*, vol. 6, pp. 905–936, Dec. 1, 2005, ISSN: 1532-4435 (cit. on p. 49).

[85] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of Loopy Belief Propagation," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, ser. UAI'05, Arlington, Virginia, USA: AUAI Press, Jul. 26, 2005, pp. 396–403, ISBN: 978-0-9749039-1-0 (cit. on p. 49).

[86] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *PNAS*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 10, 2009, ISSN: 0027-8424, 1091-6490, URL: https://doi.org/10.1073/pnas.0909892106 (cit. on p. 49).

[87] ——, "Message passing algorithms for compressed sensing: II. analysis and validation," in *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, Jan. 2010, pp. 1–5, URL: https://doi.org/10.1109/ITWKSPS.2010.5503228 (cit. on p. 49).

[88] A. Maleki, "Approximate message passing algorithms for compressed sensing," Stanford, 2010, 308 pp. (cit. on p. 51).

[89] Q. Zou and H. Yang, "A Concise Tutorial on Approximate Message Passing," Jan. 19, 2022, URL: http://arxiv.org/abs/2201.07487 (cit. on p. 51).

[90] Xiangming Meng, Sheng Wu, Linling Kuang, and Jianhua Lu, "An Expectation Propagation Perspective on Approximate Message Passing," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1194–1197, Aug. 2015, ISSN: 1070-9908, 1558-2361, URL: https://doi.org/10.1109/LSP.2015.2391287 (cit. on p. 52).

[91] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," ser. 2011 IEEE International Symposium on Information Theory Proceedings, 2011-07, 2011, pp. 2168–2172, URL: https://doi.org/10.1109/ISIT.2011.6033942 (cit. on pp. 52, 62, 75).

[92] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, "Concise Derivation for Generalized Approximate Message Passing Using Expectation Propagation," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1835–1839, Dec. 2018, ISSN: 1558-2361, URL: https://doi.org/10.1109/LSP.2018.2876806 (cit. on pp. 53, 132).

[93] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2021–2025, URL: https://doi.org/10.1109/ICASSP.2015.7178325 (cit. on p. 54).

[94] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar, "On the Convergence of Approximate Message Passing With Arbitrary Matrices," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5339–5351, Sep. 2019, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2019.2913109 (cit. on p. 54).

[95] S. Rangan, A. K. Fletcher, V. K. Goyal, E. Byrne, and P. Schniter, "Hybrid Approximate Message Passing," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4577–4592, Sep. 2017, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2017.2713759 (cit. on pp. 55, 56, 61, 62, 73, 75, 78).

[96] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "A GAMP-Based Low Complexity Sparse Bayesian Learning Algorithm," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 294–308, Jan. 2018, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2017.2764855 (cit. on p. 57).

[97] J. Vila and P. Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2011, pp. 799–803, URL: https://doi.org/10.1109/ACSSC.2011.6190117 (cit. on p. 57).

[98] J. P. Vila and P. Schniter, "Expectation-Maximization Gaussian-Mixture Approximate Message Passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2013.2272287 (cit. on p. 57).

[99] Q. Zou, H. Zhang, and H. Yang, "Expectation-Maximization-Aided Hybrid Generalized Expectation Consistent for Sparse Signal Reconstruction," *IEEE Signal Process. Lett.*, pp. 1–1, 2021, ISSN: 1558-2361, URL: https://doi.org/10.1109/LSP.2021.3065600 (cit. on pp. 57, 60).

[100] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019, ISSN: 1557-9654, URL: https://doi.org/10.1109/TIT.2019.2916359 (cit. on pp. 57, 129).

[101] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 1525–1529, URL: https://doi.org/10.1109/ACSSC.2016.7869633 (cit. on p. 57).

[102] A. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Expectation consistent approximate inference: Generalizations and convergence," in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jul. 2016, pp. 190–194, URL: https://doi.org/10.1109/ISIT.2016.7541287 (cit. on p. 57).

[103] R. Ayachi, M. Akrout, V. Shyianov, F. Bellili, and A. Mezghani, "Massive Unsourced Random Access Based on Bilinear Vector Approximate Message Passing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 5283–5287, URL: https://doi.org/10.1109/ICASSP43922.2022.9747338 (cit. on p. 57).

[104] X. Meng and J. Zhu, "Bilinear Adaptive Generalized Vector Approximate Message Passing," *IEEE Access*, vol. 7, pp. 4807–4815, 2019, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2018.2887261 (cit. on p. 57).

[105] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear Generalized Approximate Message Passing," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014, ISSN: 1053-587X, 1941-0476, URL: https://doi.org/10.1109/TSP.2014.2357776 (cit. on pp. 57, 128).

[106] S. Sarkar, A. K. Fletcher, S. Rangan, and P. Schniter, "Bilinear Recovery Using Adaptive Vector-AMP," *IEEE Trans. Signal Process.*, vol. 67, no. 13, pp. 3383–3396, Jul. 2019, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2019.2916100 (cit. on p. 57).

[107] J. T. Parker, "Approximate Message Passing Algorithms for Generalized Bilinear Inference," Ohio State University, 2014 (cit. on pp. 57, 128).

[108] Z. Yuan, Q. Guo, and M. Luo, "Approximate Message Passing With Unitary Transformation for Robust Bilinear Recovery," *IEEE Trans. Signal Process.*, vol. 69, pp. 617–630, 2021, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2020.3044847 (cit. on p. 57).

[109] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher, "Matrix inference and estimation in multi-layer models," *J. Stat. Mech.*, vol. 2021, no. 12, p. 124 004, Dec. 2021, ISSN: 1742-5468, URL: https://doi.org/10.1088/1742-5468/ac3a75 (cit. on p. 57).

[110] Q. Zou, H. Zhang, and H. Yang, "Multi-Layer Bilinear Generalized Approximate Message Passing," *IEEE Trans. Signal Process.*, vol. 69, pp. 4529–4543, 2021, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2021.3100305 (cit. on p. 57).

[111] M. Borgerding, P. Schniter, and S. Rangan, "AMP-Inspired Deep Networks for Sparse Linear Inverse Problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 15, 2017, ISSN: 1053-587X, 1941-0476, URL: https://doi.org/10.1109/TSP.2017.2708040 (cit. on p. 57).

[112] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-Learned Approximate Message Passing for Asynchronous Massive Connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, Aug. 2021, ISSN: 1536-1276, 1558-2248, URL: https://doi.org/10.1109/TWC.2021.3067903 (cit. on p. 57).

[113] M. Erdelj, N. Mitton, and E. Natalizio, "Applications of industrial wireless sensor networks," in *Industrial Wireless Sensor Networks: Applications, Protocols, and Standards*, Boca Raton, FL: CRC Press, 2013, pp. 1–22 (cit. on p. 59).

[114] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018, ISSN: 1558-0792, URL: https://doi.org/10.1109/MSP.2018.2844952 (cit. on p. 60).

[115] K. He, Y. Li, C. Yin, and Y. Zhang, "A novel compressed sensing-based non-orthogonal multiple access scheme for massive MTC in 5G systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 81, Apr. 19, 2018, ISSN: 1687-1499, URL: https://doi.org/10.1186/s13638-018-1079-4 (cit. on p. 60).

[116] Y. Huang, Y. He, L. Shi, T. Cheng, Y. Sui, and W. He, "A Sparsity-Based Adaptive Channel Estimation Algorithm for Massive MIMO Wireless Powered Communication Networks," *IEEE Access*, vol. 7, pp. 124 106–124 115, 2019, ISSN: 2169-3536, URL: https://doi.org/10.1109/ACCESS.2019.2937183 (cit. on p. 60).

[117] A. Pramanik *et al.*, "Compressed sensing channel estimation in massive MIMO," *IET Commun.*, vol. 13, no. 19, pp. 3145–3152, 2019 (cit. on p. 60).

[118] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive Sensing-Based Adaptive Active User Detection and Channel Estimation: Massive Access Meets Massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020, ISSN: 1941-0476, URL: https://doi.org/10.1109/TSP.2020.2967175 (cit. on pp. 60, 78).

[119] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan, and L. Hanzo, "Iterative Joint Channel Estimation, User Activity Tracking, and Data Detection for FTN-NOMA Systems Supporting Random Access," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2963–2977, May 2020, ISSN: 1558-0857, URL: https://doi.org/10.1109/TCOMM.2020.2975169 (cit. on p. 60).

[120] Q. Zou, H. Zhang, D. Cai, and H. Yang, "Message Passing Based Joint Channel and User Activity Estimation for Uplink Grant-Free Massive MIMO Systems With Low-Precision ADCs," *IEEE Signal Process. Lett.*, vol. 27, pp. 506–510, 2020, ISSN: 1558-2361, URL: https://doi.org/10.1109/LSP.2020.2979534 (cit. on pp. 60, 62, 73).

[121] K. B. Lee, S. Cheon, and C. O. Kim, "A Convolutional Neural Network for Fault Classification and Diagnosis in Semiconductor Manufacturing Processes," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017, ISSN: 1558-2345, URL: https://doi.org/10.1109/TSM.2017.2676245 (cit. on pp. 61, 79).

[122] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula Methods in Finance*. John Wiley & Sons, Oct. 22, 2004, 312 pp., ISBN: 978-0-470-86345-9 (cit. on p. 90).

[123] R. B. Sinitsyn and F. J. Yanovsky, "MIMO radar copula ambiguity function," in *2012 9th European Radar Conference*, Oct. 2012, pp. 146–149 (cit. on p. 90).

[124] C. Zheng, M. Egan, L. Clavier, G. W. Peters, and J.-M. Gorce, "Copula-Based Interference Models for IoT Wireless Networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6, URL: https://doi.org/10.1109/ICC.2019.8761783 (cit. on p. 90).

[125] N. Deligiannis, J. F. C. Mota, E. Zimos, and M. R. D. Rodrigues, "Heterogeneous Networked Data Recovery From Compressive Measurements Using a Copula Prior," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5333–5347, Dec. 2017, ISSN: 1558-0857, URL: https://doi.org/10.1109/TCOMM.2017.2746099 (cit. on p. 90).

[126] K. Ghavami and M. Naraghi-Pour, "Blind Channel Estimation and Symbol Detection for Multi-Cell Massive MIMO Systems by Expectation Propagation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 943–954, Feb. 2018, ISSN: 1536-1276, URL: https://doi.org/10.1109/TWC.2017.2772837 (cit. on p. 127).

[127] X. Meng, L. Zhang, C. Wang, *et al.*, "Advanced NOMA Receivers from a Unified Variational Inference Perspective," *IEEE J. Sel. Areas Commun.*, pp. 1–1, 2020, ISSN: 1558-0008, URL: https://doi.org/10.1109/JSAC.2020.3018834 (cit. on p. 127).

[128] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang, and P. Schniter, "Multi-User Detection Based on Expectation Propagation for the Non-Coherent SIMO Multiple Access Channel," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 9, pp. 6145–6161, Sep. 2020, ISSN: 1558-2248, URL: https://doi.org/10.1109/TWC.2020.3000419 (cit. on p. 127).

[129] Y. Mei, Z. Gao, Y. Wu, *et al.*, "Compressive Sensing Based Joint Activity and Data Detection for Grant-Free Massive IoT Access," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2021, ISSN: 1558-2248, URL: https://doi.org/10.1109/TWC.2021.3107576 (cit. on p. 127).

[130] Till Tantau, "The TikZ and PGF Packages," URL: https://mirrors.ircam.fr/pub/CTAN/graphics/pgf/base/doc/pgfmanual.pdf (cit. on p. 131).

[131] *Inkscape*, URL: https://inkscape.org/ (cit. on p. 131).

[132] *Python*, Python Software Foundation, URL: https://www.python.org/ (cit. on p. 131).

[133] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 17, 2020, ISSN: 0028-0836, 1476-4687, URL: https://doi.org/10.1038/s41586-020-2649-2 (cit. on p. 131).

[134] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nat Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2, 2020, ISSN: 1548-7091, 1548-7105, URL: https://doi.org/10.1038/s41592-019-0686-2 (cit. on p. 131).

[135] T. A. Caswell, M. Droettboom, A. Lee, *et al.*, *Matplotlib*, version v3.5.0, Zenodo, Nov. 16, 2021, URL: https://doi.org/10.5281/ZENODO.5706396 (cit. on p. 131).

[136] ENS Lyon, *Pole Scientifique de Modélisation Numérique*, URL: http://www.ens-lyon.fr/PSMN/doku.php?id=documentation:accueil (cit. on p. 131).

FOLIO ADMINISTRATIF

THÈSE DE L'UNIVERSITÉ DE LYON OPERÉE AU SEIN DE L'INSA LYON

NOM : CHETOT *Date de soutenance* : 07/07/2022

Prénoms : Lélio, Georges

Titre : Activity Models and Bayesian Estimation Algorithms for Wireless Grant-Free Random Access

Nature : Doctorat *Numéro d'ordre* : 2022LYSEI062

École doctorale : Électronique, Électrotechnique, Automatique (EEA), n° 160

Spécialité : Traitement du signal et de l'image

Résumé :
Les nouveaux réseaux sans fil de cinquième génération (5G) ont commencé récemment a être déployés dans le monde. Leur arrivée amène l'émergence d'un large spectre de nouveaux services dont les stricts prérequis technologiques font que les performances de la 5G peuvent être vues comme celles de la 4G en dix fois plus important. Ces services sont articulés autour des usages des communications mobiles à large bande (eMBB), ultra-fiables à faible latence (uRLLC) et massives entre machines (mMTC), où chacun d'entre eux nécessite le développement de technologies clés. Ces dernières joueront également un rôle important dans l'émergence de la 6G.
Cette thèse se focalise sur l'accès aléatoire spontané (GFRA) comme un facilitateur de l'uRLLC and du mMTC. Ce nouveau protocole introduit pour la version New Radio de la 5G vise à réduire le surplus de données de l'accès aléatoire pour diminuer la latence d'accès d'un équipement utilisateur (UE) à un point d'accès (AP).
Atteindre un GFRA efficace est capital pour une pléthore d'applications 5G tels que les réseaux sans-fils d'internet des objets (IoT). L'étude de nouvelles techniques d'accès non-orthogonales (NOMA) est donc considérée. En empruntant à la théorie de l'acquisition comprimée (CS), en particulier bayésienne, de nouveaux algorithmes à échange généralisé de messages approximés (GAMP) ont été développés pour résoudre conjointement les problèmes de détection d'utilisateur actifs et d'estimation de canal (AUDaCE). Le premier est crucial pour identifier proprement les UEs transmettant dans le contexte de réseaux denses à large échelle; le second est tout aussi important puisqu'il permet à un AP de transmettre des données aux UEs détectés de façon fiable.
Contrairement aux travaux existants sur le sujet, cette thèse étudie le problème d'AUDaCE pour des réseaux sans-fil dans lesquels l'activité des UEs est supposée corrélée, ce qui serait typiquement le cas pour des réseaux densément peuplés. Pour cela, deux nouveaux modèles d'activité sont présentés. Le premier suppose que l'activité des UEs est caractérisée par des groupes homogènes d'activité (GHomA) tandis que le second considère plus généralement des groupes hétérogènes d'activité (GHetA).
Pour chaque modèle, un algorithme hybride GAMP (HGAMP) est développé. Avec l'aide de variables latentes de groupe modélisant les probabilités d'activités des UEs, l'algorithme GHomA-HGAMP peut résoudre l'AUDaCE pour GFRA en exploitant l'homogénéité d'activité des groupes. Si l'activité est hétérogène, chaque UE possède sa propre probabilité d'activité corrélée à celles des autres UEs, permettant d'utiliser la théories des copules pour développer GHetA-HGAMP.
Une amélioration significative des performances de ces deux nouveaux algorithmes par rapport à celles d'algorithmes existants de même nature (GAMP modifié et HGAMP pour parcimonie de groupe) est démontrée par le biais de nombreuses simulations numériques. En particulier, l'estimation de canal est améliorée jusqu'à 4dB pour une détection d'utilisateur avec deux fois moins d'erreurs.
Ainsi cette thèse propose une approche systématique de l'AUDaCE pour des réseaux sans-fils avec des activités corrélées en utilisant des outils bayésiens. Une esquisse de l'utilisation de cette approche pour des communications multi-porteuses et pour des transmissions spontanées de données avec AUDaCE est proposée afin de conclure cette thèse.

Mots-clés : 5G, mMTC, uRLLC, réseaux sans-fil, accès aléatoire, détection d'utilisateur, estimation de canal, AUDaCE, traitement du signal statistique, acquisition comprimée, inférence bayésienne, HGAMP

Laboratoire(s) de recherche : Centre of Innovation in Telecommunications and Integration of Service (CITI)

Directeur de thèse : Jean-Marie GORCE

Président du jury :

Composition du jury : Philippe CIBLAT, Catherine DOUILLARD, Dejan VUKOBRATOVIĆ, Aline ROUMY, Čedomir STEFANOVIĆ, Jean-Marie GORCE, Malcolm EGAN