Statistical physics and approximate message passing algorithms for sparse linear estimation problems in signal processing and coding theory

Jean Barbier

STATISTICAL PHYSICS AND APPROXIMATE MESSAGE PASSING ALGORITHMS FOR SPARSE LINEAR ESTIMATION PROBLEMS IN SIGNAL PROCESSING AND CODING THEORY

Thèse n.
présentée le 18 septembre 2015
à l'École Normale Supérieure de Paris.
Travail effectué au laboratoire de physique statistique
de l'École Normale Supérieure de Paris
sous la direction de Prof. Florent KRZAKALA
et au sein de l'école doctorale physique en île de France

Université Paris Diderot (Paris 7) Sorbonne Paris cité
pour l'obtention du grade de Docteur ès Sciences
spécialité physique théorique, par

Jean BARBIER

acceptée sur la proposition du jury:

Prof. Laurent DAUDET, examinateur
Prof. Silvio FRANZ, examinateur
Prof. Florent KRZAKALA, directeur
Prof. Marc LELARGE, examinateur
Prof. Nicolas MACRIS, rapporteur
Prof. Marc MÉZARD, examinateur
Prof. Federico RICCI-TERSENGHI, examinateur
Prof. David SAAD, rapporteur

Paris, École Normale Supérieure, 2015.

La science c'est plutôt cool quand même...

A mes parents qui m'ont tout donné et mes soeurs que j'aime par dessus tout.
Merci de m'avoir supporté jusqu'ici...

Remerciements

Je tiens avant tout à remercier mes parents qui m'ont toujours laissé totalement libre de mes choix et m'ont tout donné, qui m'ont soutenu pendant cette longue période souvent difficile, parfois très difficile et toujours merveilleuse qu'ont été mes études. Merci à mes soeurs, Louise et Virginie. Merci à toute ma famille, présente dans les joies et difficultés.

Je remercie mon directeur Florent, le maitre Jedi, un ami. Merci de m'avoir fait confiance, de m'avoir enseigné par la pratique le vrai sens critique, de m'avoir présenté tant de personnes incroyables, de m'avoir offert ces expériences enrichissantes en école et ailleurs, d'avoir été insupportable quand il le fallait vraiment et d'avoir toujours su trouver l'équilibre entre pression et liberté, travail et humour, entre aide et indépendance. Je n'aurais réellement pas pu vivre une meilleure expérience pour mon doctorat.

Merci à Eric Tramel et Francesco Caltagirone pour leur sympathie et leurs explications.

Je tiens à remercier ceux qui ont rendu ma thèse encore plus agréable par leur amitié, qui ont transformé tous mes repas (et soirées) en moments toujours plus marrants, qui m'ont presentés leurs amis... Merci Thim, Thomas, le petit Christophe, Alaa, Alice, Antoine, Sophie, Ralph et merci à tous les autres aussi.

Je voudrais également remercier ceux qui m'ont permis d'en arriver là. En particulier Riccardo Zecchina qui m'a fait découvrir les domaines de l'inférence et de la belle physique statistique du désordre, Alessandro Pelizzola qui a tout fait pour rendre mon année à Turin si enrichissante et simple, Silvio Franz et Emmanuel Trizac pour m'avoir fait confiance en m'envoyant en Italie, merci à Marc Mézard pour son influence directe ou indirecte dans tous les travaux auxquels j'ai pu m'intéresser pendant ces trois années et demie, sur tous les papiers que j'ai pu lire et pour m'avoir montré ce que c'est que de vraiment savoir skier... Merci à Lenka Zdeborova pour ces collaborations fructueuses, Laurent Daudet pour m'avoir aidé à prendre des décisions importantes.

Merci aussi à Rüdiger Urbanke et Nicolas Macris pour leur accueil à Lausanne. J'attends avec impatience les années à venir... Merci aux autres membres de mon jury de prendre le temps pour ma soutenance et avoir la patience de lire ma thèse: David Saad, Federico Ricci-Tersenghi et Marc Lelarge.

Je n'oublie pas les membres de mon premier lieu de travail, le laboratoire de physico-chimie théorique de l'École Supérieure de physique et chimie de Paris, en particulier Élie Raphaël et Thomas Salez pour leur sympathie perpétuelle, ainsi que Justine et Antoine.

Je dois beaucoup à mes colocataires qui m'auront supporté malgré mes crises de nerfs suite à trop de message-passing et avec qui j'aurais tellement rigolé: Charlotte, PH et Manon, vous

êtes les meilleurs.
Merci Auré d'avoir rendu mes études encore plus intéressantes, du début à la fin.
Je n'oublie pas Brian et tous les serveurs de Chez Léa sans qui cette période de rédaction aurait été très différente...
Je tiens également à remercier la DGA de m'avoir financé.
La liste pourrait encore continuer longtemps ayant rencontré tellement de gens intéressants et sympathiques durant ces années, merci à vous tous..

Paris, 20 Mai 2015 J. B.

Abstract

This thesis is interested in the application of statistical physics methods and inference to signal processing and coding theory, more precisely, to sparse linear estimation problems.

The main tools are essentially the graphical models and the approximate message-passing algorithm together with the cavity method (referred as the state evolution analysis in the signal processing context) for its theoretical analysis. We will also use the replica method of statistical physics of disordered systems which allows to associate to the studied problems a cost function referred as the potential of free entropy in physics. It allows to predict the different phases of typical complexity of the problem as a function of external parameters such as the noise level or the number of measurements one has about the signal: the inference can be typically easy, hard or impossible. We will see that the hard phase corresponds to a regime of coexistence of the actual solution together with another unwanted solution of the message passing equations. In this phase, it represents a metastable state which is not the true equilibrium solution. This phenomenon can be linked to supercooled water blocked in the liquid state below its freezing critical temperature.

Thanks to this understanding of blocking phenomenon of the algorithm, we will use a method that allows to overcome the metastability mimicing the strategy adopted by nature itself for supercooled water: the nucleation and spatial coupling. In supercooled water, a weak localized perturbation is enough to create a crystal nucleus that will propagate in all the medium thanks to the physical couplings between closeby atoms. The same process will help the algorithm to find the signal, thanks to the introduction of a nucleus containing local information about the signal. It will then spread as a "reconstruction wave" similar to the crystal in the water.

After an introduction to statistical inference and sparse linear estimation, we will introduce the necessary tools. Then we will move to applications of these notions. They will be divided into two parts.

The signal processing part will focus essentially on the compressed sensing problem where we seek to infer a sparse signal from a small number of linear projections of it that can be noisy. We will study in details the influence of structured operators instead of purely random ones used originally in compressed sensing. These allow a substantial gain in computational complexity and necessary memory allocation, which are necessary conditions in order to work with very large signals. We will see that the combined use of such operators with spatial coupling allows the implementation of an highly optimized algorithm able to reach near to optimal performances. We will also study the algorithm behavior for reconstruction of

approximately sparse signals, a fundamental question for the application of compressed sensing to real life problems. A direct application will be studied via the reconstruction of images measured by fluorescence microscopy. The reconstruction of "natural" images will be considered as well.

In coding theory, we will look at the message-passing decoding performances for two distincts real noisy channel models. A first scheme where the signal to infer will be the noise itself will be presented. The second one, the sparse superposition codes for the additive white Gaussian noise channel is the first example of error correction scheme directly interpreted as a structured compressed sensing problem. Here we will apply all the tools developed in this thesis for finally obtaining a very promising decoder that allows to decode at very high transmission rates, very close of the fundamental channel limit.

Keywords: Statistical physics, disordered systems, mean field theory, signal processing, Bayesian inference, statistical learning, coding theory, linear estimation, sparsity, approximate sparsity, compressed sensing, spatial coupling, Gaussian channel, error correcting codes, sparse superposition codes, approximate message passing algorithm, cavity method, state evolution analysis, replica method.

Résumé

Cette thèse s'intéresse à l'application de méthodes de physique statistique des systèmes désordonnés ainsi que de l'inférence à des problèmes issus du traitement du signal et de la théorie du codage, plus précisément, aux problèmes parcimonieux d'estimation linéaire.

Les outils utilisés sont essentiellement les modèles graphiques et l'algorithme approximé de passage de messages ainsi que la méthode de la cavité (appelée analyse de l'évolution d'état dans le contexte du traitement de signal) pour son analyse théorique. Nous aurons également recours à la méthode des répliques de la physique des systèmes désordonnées qui permet d'associer aux problèmes rencontrés une fonction de coût appelé potentiel ou entropie libre en physique. Celle-ci permettra de prédire les différentes phases de complexité typique du problème, en fonction de paramètres externes tels que le niveau de bruit ou le nombre de mesures liées au signal auquel l'on a accès : l'inférence pourra être ainsi typiquement simple, possible mais difficile et enfin impossible. Nous verrons que la phase difficile correspond à un régime où coexistent la solution recherchée ainsi qu'une autre solution des équations de passage de messages. Dans cette phase, celle-ci est un état métastable et ne représente donc pas l'équilibre thermodynamique. Ce phénomène peut-être rapproché de la surfusion de l'eau, bloquée dans l'état liquide à une température où elle devrait être solide pour être à l'équilibre.

Via cette compréhension du phénomène de blocage de l'algorithme, nous utiliserons une méthode permettant de franchir l'état métastable en imitant la stratégie adoptée par la nature pour la surfusion : la nucléation et le couplage spatial. Dans de l'eau en état métastable liquide, il suffit d'une légère perturbation localisée pour que se créer un noyau de cristal qui va rapidement se propager dans tout le système de proche en proche grâce aux couplages physiques entre atomes. Le même procédé sera utilisé pour aider l'algorithme à retrouver le signal, et ce grâce à l'introduction d'un noyau contenant de l'information locale sur le signal. Celui-ci se propagera ensuite via une "onde de reconstruction" similaire à la propagation de proche en proche du cristal dans l'eau.

Après une introduction à l'inférence statistique et aux problèmes d'estimation linéaires, on introduira les outils nécessaires. Seront ensuite présentées des applications de ces notions. Celles-ci seront divisées en deux parties.

La partie traitement du signal se concentre essentiellement sur le problème de l'acquisition comprimée où l'on cherche à inférer un signal parcimonieux dont on connaît un nombre restreint de projections linéaires qui peuvent être bruitées. Est étudiée en profondeur l'in-

fluence de l'utilisation d'opérateurs structurés à la place des matrices aléatoires utilisées originellement en acquisition comprimée. Ceux-ci permettent un gain substantiel en temps de traitement et en allocation de mémoire, conditions nécessaires pour le traitement algorithmique de très grands signaux. Nous verrons que l'utilisation combinée de tels opérateurs avec la méthode du couplage spatial permet d'obtenir un algorithme de reconstruction extrêmement optimisé et s'approchant des performances optimales. Nous étudierons également le comportement de l'algorithme confronté à des signaux seulement approximativement parcimonieux, question fondamentale pour l'application concrète de l'acquisition comprimée sur des signaux physiques réels. Une application directe sera étudiée au travers de la reconstruction d'images mesurées par microscopie à fluorescence. La reconstruction d'images dites "naturelles" sera également étudiée.

En théorie du codage, seront étudiées les performances du décodeur basé sur le passage de message pour deux modèles distincts de canaux continus. Nous étudierons un schéma où le signal inféré sera en fait le bruit que l'on pourra ainsi soustraire au signal reçu. Le second, les codes de superposition parcimonieuse pour le canal additif Gaussien est le premier exemple de schéma de codes correcteurs d'erreurs pouvant être directement interprété comme un problème d'acquisition comprimée structuré. Dans ce schéma, nous appliquerons une grande partie des techniques étudiée dans cette thèse pour finalement obtenir un décodeur ayant des résultats très prometteurs à des taux d'information transmise extrêmement proches de la limite théorique de transmission du canal.

Mots clefs : Physique statistique, systèmes désordonnés, théorie du champ moyen, traitement du signal, inférence Bayesienne, apprentissage statistique, théorie du codage, estimation linéaire, parcimonie, parcimonie approximative, acquisition comprimée, couplage spatial, canal Gaussien, codes correcteurs d'erreurs, codes de superposition parcimonieuse, algorithme approximé de passage de messages, méthode de la cavité, analyse d'évolution des états, méthode des répliques.

Contents

List of Figures

List of symbols and acronyms

a	:	a generic quantity, usually a scalar if not precised further.
$\mathbf{a}$	:	a vector.
$\mathbf{A}$	:	a matrix.
$a_i = (\mathbf{a})_i$	:	the i^{th} component of $\mathbf{a}$. The two notations are equivalent.
A_{ij}	:	the element at the i^{th} line and j^{th} column of $\mathbf{A}$.
$\mathbf{A}_{i,\bullet}$	:	the i^{th} line of $\mathbf{A}$.
$\mathbf{A}_{\bullet,i}$	:	the i^{th} column of $\mathbf{A}$.
$\mathbf{s}$	:	generally the signal or message to infer.
$\mathbf{x}$	:	an intermediate variable for estimating $\mathbf{s}$.
$\hat{\mathbf{x}}$	:	the final estimate of the signal $\mathbf{s}$.
$\mathbf{y}$	:	the measurement vector or codeword.
$\mathbf{F}$	:	the measurement or coding operator.
$\hat{a}$	:	the estimate of the quantity a.
K	:	number of non zero components in the sparse signal $\mathbf{s}$.
N	:	number of scalar components of the signal or message to reconstruct $\mathbf{s}$.
L	:	number of vector components of the signal or message to reconstruct $\mathbf{s}$.
M	:	number of scalar components of the measure or codeword $\mathbf{y}$.
$:=$	:	equal by definition.
$\rho := K/N$	:	density of non zero components of $\mathbf{s}$.
$\alpha := M/N$	:	the measurement rate.

$a\|b$	:	a "given" (or such that) b.
$\mathcal{N}(u\|\mu,\sigma^2)$	:	a Gaussian distribution for the random variable u with mean μ and variance σ^2.
$\delta(x)$	:	the delta Dirac function (which is formally a distribution) that is a probability density giving a infinite weight to the single value $x=0$.
δ_{ij}	:	the kronecker symbol, which is one if $i=j$, 0 else.
$\partial i := \{j : (ij) \in E\}$	:	the set of neighbors nodes to node i in the graphical representation of the problem (E is the set of edges).
$\partial i \setminus j$	:	the set of neighbors nodes to node i except the node j in the graphical representation of the problem.
$u \sim P(u\|\boldsymbol{\theta})$	:	u is a random variable with distribution $P(u\|\boldsymbol{\theta})$ that depends on a vector of parameters $\boldsymbol{\theta}$.
$\mathbb{E}_x(y(x)) = \mathbb{E}_P(y(x))$	:	the average of the function $y(x)$ with respect to the random variable $x \sim P(x)$. The two notations are used equivalently when there are no possible ambiguities.
$<\mathbf{x}> := 1/N \sum_i^N x_i$	:	the empirical average of the vector $\mathbf{x}$ (here there are N components).
$\{x_i\|f(x_i)\}_{g(i)}^b$	:	the ensemble made of $\{x_i\}$ that verify the conditions $f(x_i)\ \forall\ i \in \{1,...,b\|g(i)$ is true$\}$. As for any operations such as sums, products, etc, if the lower bound of the index is not explicited, it means it starts from 1. For example, $\{a_i\|a_i>0\}_i^N = \{a_i\|a_i>0\}_{i=1}^N$, $\sum_i^N x_i = \sum_{i=1}^N x_i$, etc.
$[x_i\|f(x_i)]_{g(i)}^b$	:	the concatenation of $\{x_i\|f(x_i)\}_{g(i)}^b$ to form a vector of cardinality that depends on g and b.
$[a,b]$	:	the simple concatenation of a and b to form a vector.
$\mathbb{I}(E)$	:	the indicator function, which is 1 if the condition E is true, 0 else.
Δ	:	the variance of the i.i.d Gaussian measurement noise.
$\boldsymbol{\xi} := [\xi_\mu \sim \mathcal{N}(\xi_\mu\|0,\Delta)]_\mu^M$	:	the i.i.d Gaussian measurement noise vector.
$\|\|\mathbf{y}\|\|_p := 1/M\left(\sum_i^M \|y_i\|^p\right)^{1/p}$	:	the rescaled ℓ_p norm of $\mathbf{y}$, which is here of size M.
$\text{snr} := \|\|\mathbf{y}\|\|_2^2/\Delta$	:	the signal to noise ratio, where y is the codeword.
$z \in O(u)$	:	z is a quantity of the same order as u i.e. $z = Cu$ where C is a constant that does not scale with the problem size N.

$z \in o(u)$	:	z is a quantity at least an order smaller than u i.e. $\lim \frac{z}{u} = 0$ in a proper limit that depends on the context.
$a \approx b$	:	a and b are equal up to a negligible difference $\in o(a)$.
$\mathbf{x}_{\setminus i} := [x_j]_{j \neq i}^N$	:	the vector made of the $N-1$ components of $\mathbf{x}$ that are not the i^{th} one.
$\mathbf{x}_{a \setminus i} := [x_j \in \partial a]_{j \neq i}$	:	the vector of components of $\mathbf{x}$ that are neighbors of the factor a, except the i^{th} one.
$d\mathbf{x} := \prod_i^N dx_i$	:	integration over all the components of the size N vector $\mathbf{x}$.
$\mathscr{D}\mathbf{x} := \prod_i^N \mathscr{D}x_i = \prod_i^N dx_i \mathscr{N}(x_i \vert 0,1)$	:	integration over all the components of the size N vector $\mathbf{x}$ with unit centered Gaussian measure.
$\mathbf{X}^\intercal$	:	the transpose of a vector or matrix.
$\mathbf{x}^a := [x_i^a]_i^N$	:	the component wise power operation for a vector (or matrix).
$\mathbf{XY} := \mathbf{Z}$ where $Z_{ij} = X_{ij} Y_{ij}$	:	the component wise product between matrices or vectors of same size.
$\mathbf{x}^\intercal \mathbf{y} = \mathbf{x}\mathbf{y}^\intercal := \sum_i^N x_i y_i$	:	the scalar product between two vectors of size N. The two notations are equivalent as we always assume the vectors to have proper dimensions to apply the product.
$\mathbf{Fx} := [\sum_i^N F_{\mu i} x_i]_\mu^M$	:	the matrix product between $\mathbf{F}$ of size $M \times N$ and $\mathbf{x}$ of size N.
$(\mathbf{Fx})_\mu := \sum_i^N F_{\mu i} x_i$	:	the μ^{th} component of the vector $\mathbf{Fx}$.
$\mathrm{Var}_P(u) := \mathbb{E}_P(u^2) - \mathbb{E}_P(u)^2$	:	the variance of the random variable u with distribution P.
$\mathrm{inv}(\mathbf{A})$	:	the inverse of the matrix A.
∂_x	:	the partial derivative with respect to x.
$\vert\mathbf{x}\vert$	:	the number of components of $\mathbf{x}$ or the cardinality of an ensemble.
$a \propto b$	:	a is proportional to b, i.e. they are equal up to a constant.

CS	:	compressed sensing.
MSE	:	mean squarre error.
MMSE	:	minimum mean squarre error.
MAP	:	maximum à posteriori.
i.e.	:	id est.
N-d	:	N-dimensional.
BP	:	belief propagation.
AMP	:	approximate message passing.
SE	:	state evolution analysis.
i.i.d	:	independent and identically distributed.
AWGN	:	additive white Gaussian noise.
EM	:	expectation maximization.

Part I

Main contributions and structure of the thesis

1 Main contributions

My work has been concentrated around two main axis: i) signal processing through compressed sensing and its application in image reconstructions and ii) coding theory over real channels and its links to compressed sensing. I will present here my main contributions in these fields, dividing my work into practical achievements through algorithms design and the theoretical and asymptotic studies. In addition, I've worked on a combinatorial optimization problem, namely the independent set problem, in order to get familiar with the cavity method and the diverse phase transitions that occur in such problems. I will start by briefly present this piece of work that I won't detail in this thesis. This choice has been made for sake of coherence of the thesis: all the problems I've worked on, except this one, belong to the class of sparse linear estimation problems and a common methodology is used, based on the approximate message-passing algorithm and the state evolution and replica analyzes for the asymptotic studies.

1.1 Combinatorial optimization

1.1.1 Study of the independent set problem, or hard core model on random regular graphs by the one step replica symmetry breaking cavity method

In this work [1], we have studied the NP-hard independent set problem on random regular graphs, the dual of the vertex cover problem better known as the hard-core model in the physics literature. This model is of great interest as it can be seen as a lattice version of the hard spheres, a fundamental model in physics. The aim of this theoretical work was to reconciliate the two extreme regimes corresponding to the high and low connectivities of the graph. Both were known for a long time but each with a totally different behavior. While in the low connectivity regime, the problem displays a continuous full replica symmetry breaking transition as the density of particles increases in the graph, it was proven in the mathematical literature that in the high connectivity limit, the opposite phenomenon happens: the space of solution breaks discontinuously into exponentially many well separated components, a behavior typically found in glassy systems, at a density which is the half of the maximum one.

The main result obtained through the cavity method is the obtention of the full phase diagram of the problem for all connectivities. The computation of the different phase transitions in the problem by population dynamics in the 1RSB framework shows that the change in behavior between a continuous full RSB regime and the appearance of the discontinuous 1RSB transition happens at connectivity $K = 16$. It appears that between $16 \leq K < 20$, despite the existence of a stable 1RSB glassy phase, the continous transition remains if the density of particles is too high until for $K \geq 20$, the 1RSB phase becomes stable for all densities until the maximum one. This shows that this model is the simplest mean field model of the glass and jamming transitions, and can be used to get insights on more complex models such as the hard spheres in high dimensions. In addition, the asymptotic analysis in the cavity framework is in perfect agreement with the rigorous results at high connectivity, which supports further the validity of the cavity method in such problems despite it is not yet rigorously established.

1.2 Signal processing and compressed sensing

1.2.1 Generic expectation maximization approximate message-passing solver for compressed sensing

Practical achievements : I've implemented a modular AMP solver for compressed sensing in MATLAB, that includes a lot of different possible priors for the signal model. In addition, most of the free parameters in these priors and the noise variance can be learned efficiently through expectation maximisation. All the algorithms can be found at `https://github.com/jeanbarbier/BPCS_common`.

1.2.2 Approximate message-passing for approximate sparsity in compressed sensing

Practical achievements : My first work [2] during this thesis was focused on the study of the AMP performances and behavior when dealing with signals that are only approximately sparse, sometimes referred as compressible. We implemented a specifically designed prior for approximate sparsity, and the expectation maximisation learning of all the parameters of this prior.

Theoretical results : We performed the static and dynamical asymptotic analyzes thanks to the replica and state evolution techniques respectively. We extracted how the AMP performances change as a function of the variance of the small components part of the signal, a kind of effective noise, and what are the best possible results from the Bayesian point of view. A first order phase transition blocking the AMP solver under some measurement ratio appears, but we have shown how the spatial coupling strategy can restore the optimality of the AMP solver and until which level of effective noise it makes sense to use this strategy.

1.2.3 Influence of structured operators with approximate message-passing for the compressed sensing of real and complex signals

Practical achievements : A large amount of work has been put into the combination of this AMP solver with structured operators, based of fast Hadamard and Fourier transforms [3]. Furthermore, I've developed a set of routines for applying the spatial coupling strategy in combination with these structured operators. The result is a very fast AMP solver yet optimal from the information theoretic point of view. It is able to deal with very large signals as the use of such operators allows to side step the memory issues that quickly arise working with the large matrices that one must store if not structured.

Theoretical results : Side to side with the developement of the AMP solver combined with full or spatially-coupled structured operators, we have studied how the use of such operators influence the performances of AMP in noiseless compressed sensing of real or complex signals. The point is that AMP together with the state evolution analysis has originally been derived for i.i.d matrices but we have numerically shown that despite the state evolution does not describe properly the AMP dynamic with structured operators, it remains an accurate predictive tool for its final performances as the reconstruction quality is the same than with i.i.d matrices. Furthermore, it appeared that structured operators improves the rate of convergence of AMP. In addition, the study of the spatial coupling strategy have shown that it performs very well with such operators as well and allows to make optimal inference as long as the signal density is not too large.

1.2.4 "Total variation" like reconstruction of natural images by approximate message-passing

Practical achievements : In order to reconstruct "natural images" (i.e. that are sparse in the discrete gradient space) in the compressive regime, I've worked on an AMP implementation mimicing the total-variation optimization algorithms. The result is an algorithm able to compete with the best optimization solvers in terms of reconstruction results, but with fewer parameters to tune as most of them can be learned efficiently.

1.2.5 Compressive fluorescence microscopy images reconstruction with approximate message-passing

Practical achievements : Still in the field of compressive imaging, I've developed an AMP implementation for reconstruction of images measured by compressive fluorescence microscopy, based on an approximate sparsity prior. The images here are highly sparse in the direct pixel space. The AMP overcomes the ℓ_1 optimization solvers in terms of reconstruction quality, speed and minimum undersampling ratio to get good results. Furthermore, all the free parameters of the model can be learned efficiently.

1.3 Coding theory for real channels

1.3.1 Error correction of real signals corrupted by an approximately sparse Gaussian noise

Practical achievements : My first introduction to the field of coding theory is through a work [4] that naturally followed the study of approximate sparsity in compressed sensing. In this work, the aim is error correction over a channel that adds to a real transmitted message an approximately sparse Gaussian noise with some large components, the others being a smaller amplitude background noise. The aim is the reconstruction of the noise in order to cancel it at the end. We naturally used our previously developed AMP solver for approximately sparse signals to design an efficient decoder. In addition, the use of spatial coupling in this context allowed the decoder to perform at high rate, well above the results obtained with previously developed convex optimization based solvers.

Theoretical results : Based on the state evolution analysis for compressed sensing of approximately sparse signals, we predicted the asymptotic performances of our AMP based decoder, which shows that it is robust to the background noise, in the sense that the reconstruction error grows continuously with the variance of the small components of the noise.

1.3.2 Study of the approximate message-passing decoder for sparse superposition codes over the additive white Gaussian noise channel

Practical achievements : We presented the first decoder based on AMP for the sparse superposition codes, a capacity achieving error correction scheme over the AWGN channel. We exposed the first close connection between the compressed sensing theory and error correcting codes, as the sparse superposition codes decoding can directly be interpreted as a compressed sensing problem for signals with structured sparsity, or equivalently of signals with vectorial components instead of scalar ones, for which compressed sensing has originally been developed.

Our first paper [5] on this topic studied the decoder when the coding matrix is i.i.d Gaussian and the power allocation of the transmitted message is constant. Despite the presence of a first order phase transition blocking the AMP decoder well before the capacity, the numerical results have shown that the perfomances are overcoming a previous decoder based on soft thresholding methods, the adaptative successive decoder. This decoder exhibits very poor results with respect to AMP with full coding matrices for any reasonnable codeword sizes, despite being asymptotically capacity achieving.

In a second more in-depth study of our decoder [6], we included both non constant power allocation of the signal and the spatial coupling strategy to our scheme. Numerical studies suggested that despite improvements thanks to the power allocation, a well designed spatially-coupled coding matrix allows for better results both in terms of the rate of transmission and in

robustness to noise. Numerical tests also show that the combined use of power allocation and spatial coupling lower the efficiency of the scheme with respect to power allocation or spatial coupling used alone. In addition, we tested our structured Hadamard-based spatially-coupled operators that allow to perform Bayesian optimal decoding. The finite size effects in this setting have been quantified and show that this strategy is very efficient and allows to decode perfectly at very high rates, even for small sizes of the codeword.

Theoretical results : Relying on this connection with compressed sensing, we derived the state evolution analysis of the decoder in the most general setting. It allows the prediction of the asymptotic dynamical behavior of the decoder for power allocated signals, encoded with or without spatially-coupled i.i.d operators. Based on our work on structured operators, we conjectured that the final perfomances of the decoder can be accurately predicted by this analysis, despite small descrepancies during the dynamic. Again, it appeared that structured operators allows for faster convergence.

In addition, we performed the heuristic replica analysis of the coding scheme in order to compute the performances of the minimum mean square error estimator. This analysis is coherent with the previous rigorous results on the scheme as it shows that this Bayesian optimal estimator reaches asymptotically the capacity of the channel. The results suggest that the Bayes optimal estimator converge to the capacity with a rate following a power law as a function of the section size B, a fundamental parameter of the scheme. Both the derived state evolution recursions and replica potential are actually quite general, and can be applied for the prediction of the AMP behavior on any problem dealing with group sparsity, where the groups of variables are not overlaping each other.

2 Structure of the thesis and important remarks

This thesis is decomposed into three main parts. I wrote the first one, "Fundamental concepts and tools" keeping constantly the following question in mind:

What would have been really useful for me to know in order to gain a lot of time starting my PhD three years ago ?

I have thus tried to make a (very subjective) introduction to what I consider as fundamental methods and ideas useful to a starting PhD student with a statistical physics background as me and who wants to work in the fascinating field of statistical inference and graphical models. I assume very few (if not at all) knowledge about the general theory of statistical inference but also that the reader have some notions in statistical physics of disordered systems and spin glasses, as I will make efforts to establish links with this fundamental field of research. My goal is not to explain the physics of disordered systems, as many great books already exist and can be found in the references, but more to see how it can be of great interest in the apparently unrelated field of sparse linear estimation, the main subject of this thesis.

I want to emphasize an important point, not only true for this first part but for all this manuscript:

This work does not aim at mathematical rigor.

It is worth to mention it as the kind of problems treated in this thesis are classicaly studied by people of the computer science and signal processing, information and coding theory or applied mathematics communities who are used to more, or even perfectly rigorous treatments. But I am (hopefully at this time...) a statistical physicist, and the tools developed in this field, at least part of them, are not yet proven rigorously. This remark leads to another important one:

Despite the use of not (yet) rigorous methods, the theoretical results presented in this manuscript are conjectured to be exact.

It must be understood that the rigor in the treatment is not a necessary condition for the results

to be exact. The statistical physics methods used in this thesis, mainly the replica method used for asymptotic studies, are not rigorous but there exist an incredibly large amount of work and models where it has been proven to be exact, even sometimes rigorous, especially in the field of combinatorial optimization problems. Furthemore, even when not proven rigorously, numerical studies are always supporting the results of these methods. In opposite the cavity method, referred as the state evolution in the signal processing literature (and in this thesis as well) *is* rigorous for the prediction of the AMP behavior for sparse linear estimation problems (except for the vectorial components cases, but yet conjectured exact in this case).

The next two parts expose most of the original results and applications of my thesis, in addition to the asymptotic results of this first part. All the tools presented in the first part will be applied here.

The second part "Signal processing" is mostly related to compressed sensing. Chapter six presents a study of the influence of approximate sparsity in compressed sensing solved thanks to AMP. The seventh chapter studies how the use of structured operators such as Hadamard and Fourier ones change the AMP behavior in compressed sensing of real and complex signals. Furthermore, we will see that these can be combined with the spatial coupling strategy to perform optimal inference both from the theoretical and algorithmic point of views. Chapter eight is devoted to my work on compressive imaging, where the AMP algorithm is applied to the reconstruction of two different kind of images: i) the so called "natural" images, that have a compressible discrete gradient and ii) sparse images in the pixel basis, which have been obtained by fluorescence microscopy technic.

The last part "Coding theory" contains all my results on how the AMP algorithm can be used as a very efficient decoder for real noisy channels. In the chapter nine, the superposition codes for the additive white Gaussian noise channel are studied in depth. These represent the first direct link between error correction and compressed sensing. In this chapter, all the presented analytical tools are used to perform the asymptotic study of the decoder and the algorithmic tools as well: the spatial coupling and Hadamard-based structured operators are combined with AMP to get a capacity achieving decoder with very good finite size performances as shown by numerical studies. The last chapter introduce a different real channel model which adds gross Gaussian distributed errors to the signal in addition to a small Gaussian background noise. The algorithm developed in the chapter about approximate sparsity will be combined to spatial coupling to perform error correction at high rates.

Part II Fundamental concepts and tools

3 Statistical inference and linear estimation problems for the physicist layman

This first chapter which is voluntarily not technical at all is a general introduction to the main questions and techniques related to statistical inference and which are relevant to the present thesis. It can be skept by the reader familiar with statistical inference and compressed sensing as it does not include any original results. It is oriented towards a statistical physics student interested in working on inference related problems. Effort will be made in order to draw connections with physics, especially the statistical physics of disordered systems. It is thus assumed that the reader posseses a basic knowledge of this field.

First, I will define the general problem of statistical inference (or statistical estimation) and give the main distinctions among inference problems. I will also explain the difference between a direct and an inverse problem and show why in the context of statistical inference, only inverse problems really matter as opposed to statistical physics which has been created to deal with direct problems and later on extended to inverse ones. Canonical examples, yet very important both from the applicative and theoretical point of views will be presented. I will also discuss the notion of bias-variance tradeoff, which will help us to understand the fundamental limitations of statistical inference.

I will then focus on the model which is at the core of the present thesis, namely linear sparse estimation problems and compressed sensing, with a particular emphasis on the applications of compressed sensing in modern technologies.

I will introduce some basic notions of complexity theory, discussing the important tradeoff between statistical and computationnal efficiency of an algorithm. This question is essential in the modern context of "Big data" generating technologies where the sets of data produced become so large that solving the desired problem is not enough anymore: it must be done in a fast way as well.

I will then present two distinct methodologies to deal with sparse linear estimation. First I will very briefly present the convex optimization approach to compressed sensing which has been

used and studied since the appearance of the field in 2006, and which remains an important tool for nowadays applications. I will try to give some insights behind the principle of ℓ_1 norm optimization for inducing sparsity in the solution.

After a short introduction to some useful concepts of information theory, especially the notion of entropy and mutual information between random variables, I will move on to the main methodology underlying the techniques used in this thesis, namely the theory of Bayesian inference. The main principles will be exposed and then the modelisation of the sparse linear estimation problem thanks to these tools is discussed. I will underline the flexibility and the advantages of the method compared to an optimization approach. I will discuss the notion of estimator and give insights about why the minimum mean square error estimator is the appropriate choice in the continuous framework.

Finally I discuss the coding theory for the additive white Gaussian noise channel in the probabilistic Bayesian setting. The problem of communication through a noisy channel and the notion of capacity will be presented. Then we end up discussing the linear coding strategy and give a geometrical interpretation of the decoding problem.

3.1 What is statistical inference ?

Before to enter the details of the problems studied in the present thesis, we present very briefly what are *statistical inference* problems, also referred as *inverse problems*, *estimation problems* or *learning* depending on the community. Great introductions can be found such as [7–10].

3.1.1 General formulation of a statistical inference problem

Inference refers to the process of drawings conclusions about some system or phenomenon in a rational way from observations related to it, and a possible a priori knowledge about it. We are here interested in statistical inference, that will be mainly applied to signal processing problems. Assume you have access to some data y, that have been generated through some process f related to some system properties of interest represented by the so-called *signal* s. The general relation linking these objects is given by:

$$y(\boldsymbol{\theta}_s, \boldsymbol{\theta}_f, \boldsymbol{\theta}_{out}) = P_{out}\left(f\left(s(\boldsymbol{\theta}_s)|\boldsymbol{\theta}_f\right)|\boldsymbol{\theta}_{out}\right) \tag{3.1}$$

y is also referred as the observations, responses of the system or measurements in the present thesis, s as the input, the predictors or signal in the present signal processing context. These two quantities can be a scalar, vectors, matrices, sets of labels, etc. f models the *deterministic* part of the data generating process and can be any function, whereas P_{out} is a *stochastic* function linking the processed signal to the actual observations, used to model some kind of *noise*. We will refer to it as the *channel*, this terminology coming from the communication theory, where the noise models how the non ideal transmitting channel alters what is travelling

through it. In full generality, noise just means an incontrolable and undesired source of randomness that alters the observations of the system. In some cases, the noise can be correlated in some way to the signal s or process f such as in blind calibration [11], but in all the remaining, we will always consider the noise to be additive and uncorrelated with them. $\boldsymbol{\theta}_s$ are parameters of the signal such as some of its statistical properties like mean, variance, etc that can be known a priori or not. $\boldsymbol{\theta}_f$ are those of the process f, for example the number of Fourier coefficients taken in a partial discrete Fourier transform and finally $\boldsymbol{\theta}_{out}$ are parameters of the channel, usually the statistical properties of the noise.

The problem is to estimate the signal s from the knowledge of the observations y, the process f and the channel model P_{out}. The signal can be thought as a fixed realization of a random process, for example a message some emitter has sent you or some image. The channel $\boldsymbol{\theta}_{out}$, signal $\boldsymbol{\theta}_s$ and processing function $\boldsymbol{\theta}_f$ parameters can be unknown too and it is a part of the task to learn them as well. This can be done efficiently by statistical procedures such as *expectation maximization* based methods that will be detailed in sec. 4.3.8.

For sake of readibility, we drop these parameters dependencies and re-write the general statistical inference model as:

$$y = P_{out}\big(f(s)\big) \tag{3.2}$$

and the dependency on all the parameters of the problem $\boldsymbol{\theta} := [\boldsymbol{\theta}_s, \boldsymbol{\theta}_f, \boldsymbol{\theta}_{out}]$ is now always implicit.

An important remark is that in the present thesis, we mostly consider the signal s to be the unknown and f to be known as this interpretation is more relevant in the problems studied here. But actually, it would be perfectly equivalent to change their respective roles and let the process f becoming the unknown and s to be some known inputs. Depending on the problem it can be more natural to think of f as the unknow process that gave rise to the observations from controled inputs. It is really a matter of tastes. For example, in the classical reference [8], f is most of the time the object of interest that we want to infer, but in [9], the authors adopted the same convention as in the present thesis of always considering the signal s as the unknown. In examples where the unknown will be way more easily interpreted as f, it will be explicited, but in the rest s is always the infered quantity. Let us now define more precisely the different kind of statistical inference problems and give some vocabulary.

3.1.2 Inverse versus direct problems

The statistical inference problem (3.2) is by nature an *inverse problem*, in the sense that it consists in estimating properties of the system *from* some noisy observations about it, as opposed to the *direct problem* which is to obtain these observations. Getting observations about a complex system is usually quite easy compared to the associated inverse problem.

Let us give some examples to emphasize this disymmetry in complexity between direct and

inverse problems. It is easy to gather data about the past stock prices which are observations correlated to many parameters of the market and to the behavior of plenty of buyers and sellers with their own strategies, but it is highly difficult to infer from these the future prices, that must be in some way correlated to previous ones. It is nowadays quite easy to measure time series of the activity of many neurons in parallel, but the inverse problem consisting in infering the network of connections between the neurons from which result these activities is very hard [12]. In an epidemic spreading of some disease, we can partially know at some time t who are contaminated or not and have some idea of the network of connections between people, from which we would like to infer back the source of the disease: the patient(s) zero [13]. The same question can be asked for the identifications of the source of an internet virus, where it is even more easy to get the network of connections between computers. These are highly non trivial inverse dynamical problems.

Statistical physics arised at the beginning of the 19^{th} to deal with direct problems. The aim was to link the microscopic properties of the system to its macroscopic ones, impossible to derive directly from the quantum mechanics, so the knowledge about the fundamental interactions between the atoms to the physical observables and order parameters like temperature, pressure, average magnetization, etc. But as we will see, the methodology of statistical physics and especially its tools to compute thermodynamical averages over some disorder is really useful in the signal processing and inverse problems context, where the atoms are replaced by the signal components, the interactions by the constraints extracted from the observations that must verify these variables and the order parameter or observable that we would like to predict is the typical error we will make in the inference of the signal. Fig. 3.1 is a table summarizing the connections between quantitities and notions of statistical physics and those of inference and signal processing (defined in this chapter).

3.1.3 Estimation versus prediction

Inference can be important for two main reasons. In one hand, one could aim at accurately estimating the signal that gave rise to the observations. If the signal models some system, inference really is about *understanding* it. For example in seismology, the signal s of interest could be the 3-d density field of the floor in some area. Perturbations by located explosions could be performed and the vertical displacement y of the floor in some places could be measured. The relation between the signal and the measures f, even non trivial is a priori obtainable (at least approximately) from the physics of waves propagations in complex media and the locations of the explosions. The noise here comes from the approximations in the modelisation of f and the partial measurements.

In another hand, one aim could be to perform *predictions*. In this setting, it is easier to think as the signal s to be known and it is the process f which becomes the unknown object of interest. The goal is to get an estimate of it $\hat{f}$ which is able to accurately output responses to new, yet unobserved signals. This is for example the case for economical purposes. A trader is not

Statistical physics	**Inference**
Hamiltonian	Cost function
Particules, atoms, spins	Signal components
Microstates	All the possible measured signals
Macrostate	The final signal estimate $\hat{\mathbf{x}}$, or estimator
Physical phases: liquid, solid, gas, glass, etc	Computational phases: Easy, hard, impossible inference
Boltzmann distribution	Posterior distribution
Partition function $Z(\mathbf{y}, \mathbf{F}, \boldsymbol{\theta})$	Absolute probability of the measure $P(\mathbf{y}\|\boldsymbol{\theta}, \mathbf{F})$
External field	Prior distribution
External parameters: temperature, volume, chemical potential, etc	Noise variance Δ, signal to noise ratio snr, measurement rate α, signal densisty ρ, etc
Order parameter: average magnetization, correlation functions, Edwards-Anderson order parameter for spin glasses, etc	Mean square error MSE, bit error rate, etc
Quenched disorder: spin interactions, impurities in the medium, etc	Observations, sensing or coding matrix and noise realizations
Free energy/entropy	Potential function

Figure 3.1 – Relations between the statistical physics quantities and vocabulary with the inference and signal processing one, focused on the quantites useful in the present thesis, mainly related to compressed sensing and error correcting codes.

really interested in understanding the complex relations defining the market, so to estimate accurately f but more to be able to predict future prices y from the knowledge of previous ones s thanks to an estimator of the market process $\hat{f}$ that have a good predictive potential, despite it can have few common features with the true market behavior f, that can be way too complex to infer anyway from few partial observations.

All the problems that will be studied in depth in this thesis are estimation ones: we will always infer a signal that will be processed through some known transform $f(\mathbf{s}) = \mathbf{Fs}$, a matrix product.

3.1.4 Supervised versus unsupervised inference

All the previously discused examples and the model (3.2) belong to the class of *supervised* learning: problems where both observations y and the process f are known. It is thus a *fitting* problem and we can interpret the observations y associated with f as a training data set, that allows to "teach" the inference algorithm to perform its task in a supervised way. Again, the s and f roles can be switched without loss of generality. An example of a supervised problem is *classification* where one seeks for an algorithm able to class data in groups. For example if one want to design an algorithm able to distinguish between pictures of boys and girls, the

algorithm would be fed with a large amount of pictures, each with its corresponding label: boy or girl. Then, the algorithm $\hat{f}$ is expected to perfom the task well if fed with new unlabeled data, so to perform predictions.

In constrast, in an *unsupervised* inference task, one has only access to pure data y without any associated training signal s (as in the classification task) or process f (as in the seismology example). So fitting loses its sense. The aim is more to find patterns in the data that can be interpreted a posteriori. The paradigmatic problem in this class is *clustering* where you have access to data you wish to class into a relatively small number of groups compared to the number of data points [8, 14, 15]. For example, this is fundamental in recommander systems and collaborative filtering techniques that aim at clustering a set of buyers into groups, each representing a quite different consumer profile with different buying habits. Then, with some data about a new buyer (what he bought, when, the fequency, etc) he can be labeled with one of these typical profiles extracted from the clustering analysis, and thus the large amount of information gathered from the other consumers associated to this group can be used to predict products that will match this particular buyer need with high probability. Another classical problem falling in this class it the *community detection* where the data is a set of relations (a graph of connections) between unlabeled variables and one aims at labeling these points [14, 16–18]. This is performed everyday by Facebook which want to infer your potential friends (so labeled as Friend of Mr. You, in contrast with not friend of Mr. You) from the knowledge of the friendship connections among all their users.

The assumption behind this kind of techniques is that despite each data points are different, in reality a small number of parameters (the labels) is enough to describe the entire data set accurately: this is called a *dimensionality reduction* assumption and stands at the roots of many modern techniques dealing with very large data sets.

3.1.5 Parametric versus non parametric inference

Another important distinction is between *parametric* and *non parametric* problems. In the non parametric setting [8, 19], no or very few prior knowledge is assumed about the object to infer, so no prior model is assumed to perform the inference. For example in prediction, the estimate $\hat{f}$ of f is chosen among *all* possible functions that are "smooth" enough, with only parameter being the level of smoothness λ. This is a very large space, denoted by $\mathscr{S}(\lambda)$. Non parametric fitting is generally performed by applying a regularizing kernel to the observations. The obtained processing function $\hat{f}$ is thus "just" a smoother version of the observations and it thus does not require any particular a priori structure or shape for f.

On the opposite, parametric inference [8, 10] makes strong assumptions about the structure of the infered object. For example, in the linear regression problem, ones assumes that $\mathbf{y} = f(\mathbf{X}) = \mathbf{f}^\intercal \mathbf{X}$, where $\mathbf{f}$ is a vector of coefficients linking the vector of observations to the matrix $\mathbf{X}$, where each column is an input. The inference task is thus here to estimate these coefficients. Thus f is now chosen among a way more restricted space which is here $\mathbb{R}^N$,

instead of $\mathscr{S}(\lambda)$. One could even more constrain the functionnal space of the model, requiring for example that only a small fraction of the coefficients of $\mathbf{f}$ are different from zero, i.e. that $\mathbf{f}$ is *sparse*. In this example, it is clear that calling f the process or the signal is irrelevant.

A natural question that arise from this discussion is: *Why would anyone constrain the functionnal space in which the processing function is chosen ?* that can be rephrased as: *Why would anyone prefere parametric to non parametric inference ?* Of course more degrees of freedom in the choice of f or s allows a priori for a better fit of the data but a more flexible model lower its *interpretability* i.e. it makes difficult the task of interpreting the relations between the data and observations. In contrast, a quite constrained sparse linear model is directly interpretable: the observations depend (approximately at least) only on a small subset of the inputs components, identified by the non-zero coefficients of the estimate of $\mathbf{f}$. This can be very useful in a medical application. For example, we could gather observations $[y_i]_1^N \in \{1,0\}^N$ telling if yes or not the patient i has some given cancer. The lines of the input matrix $\mathbf{X}$ could correspond to conditions possibly correlated to this particular cancer such as health features or habits: is the person smoking, obese, doing sport regularly, a man, an O blood type, etc : $x_{ki} \in \{1,0\}$ tells if yes or not the i^{th} patient verifies the condition k. Then, if the inference outputs a sparse vector of coefficients $\hat{\mathbf{f}}$ such that $\mathbf{y} = \hat{\mathbf{f}}^\intercal \mathbf{X}$, one gets deep insights about which features participate or not to this cancer (the features corresponding to the non zero coefficients of $\hat{\mathbf{f}}$), and with which weight (the amplitudes of the non zero coefficients of $\hat{\mathbf{f}}$): here model interpretability is absolutely essential to identify the true causes of the cancer. But going back to the trader example that want to make accurate predictions about future stock prices, he does not not really care about the interpretability, only good predictions matter. In this case, non parametric inference can be more appropriate.

Another disadvantage of non-parametric inference is its high potential of *overfitting* the noise. It means that it is a difficult task to estimate the proper smoothing parameter λ of the observations: a too smooth model will have a very poor predictive potential whereas a too rough model will fit the noise in the data instead of the data itself which again will generate a bad model. This is probably one of the most fundamental problem in any statistical learning problem, reffered as the *bias-variance tradeoff* [8].

3.1.6 The bias-variance tradeoff: what a good estimator is ?

A fundamental question in any statistical estimation problem is: *Can we quantify the error we will make using a given estimator for the quantity we wish to infer ?* This is in general very difficult to answer in a practical setting, actually most of the theoretical part of this thesis will be dedicated to this specific question. In spite of that, it appears that in such problems, we can always differentiate three distincts sources of error, namely the *bias*, the *variance* and the *irreducible error*. The problem is of course to quantify them. Let us demonstrate what they are and how they can be interpreted, so that we can assert what are the best results we can hope to obtain. We first restrict the fully general model (3.2) to additive noise channels which

is of interest in the present thesis and quite a general model in the continuous framework: $P_{out}(\tilde{\mathbf{y}}) = \tilde{\mathbf{y}} + \boldsymbol{\xi}$. We assume first that we want to infer the processing function. The appropriate object to quantify this estimation error is called the *prediction risk*:

$$R_p := \mathbb{E}_{\mathbf{y}}\left(||\mathbf{y} - \hat{\mathbf{y}}||_2^2\right) \tag{3.3}$$
$$=< \mathbb{E}_{y_\mu}\left((y_\mu - \hat{y}_\mu)^2\right) > \tag{3.4}$$

where we used that all the measurements are independent. It is the average mean square error between the observation $\mathbf{y}(\mathbf{s}, \boldsymbol{\xi}|f)$ given by (3.2) and the prediction $\hat{\mathbf{y}} := \hat{f}(\mathbf{s}|\mathbf{y})$. The average is performed over the problem realization $\mathbf{y}$, i.e. over the noise $\boldsymbol{\xi}$ and the input data $\mathbf{s}$. We could consider also the case where $\mathbf{s}$ is fixed, it would not change the analysis, just at the end the averages with respect to $\mathbf{s}$ in the various sources of error that we will identify would disappear. f is of course independent of $\mathbf{s}$ and $\boldsymbol{\xi}$. Let us derive the equations for only one component, the final result being the average over all the components $\mu \in \{1, \dots, M\}$. We denote $f := \left(f(\mathbf{s})\right)_\mu$ and $\hat{f} := \left(\hat{f}(\mathbf{s}|\mathbf{y})\right)_\mu$ and skeep the μ index for sake of readibility:

$$\mathbb{E}_y\left((y - \hat{y})^2\right) = \mathbb{E}_{\xi,\mathbf{s}}\left((f + \xi - \hat{f})^2\right) \tag{3.5}$$
$$= \mathbb{E}_{\xi,\mathbf{s}}\left(\xi^2 + f^2 + \hat{f}^2 - 2f\hat{f} + 2\xi(f - \hat{f})\right) \tag{3.6}$$
$$= \Delta + \mathbb{E}_{\mathbf{s}}\left(f^2 - 2f\mathbb{E}_\xi(\hat{f}) + \mathbb{E}_\xi(\hat{f}^2)\right) \tag{3.7}$$
$$= \Delta + \mathbb{E}_{\mathbf{s}}\left(\mathbb{E}_\xi(\hat{f}^2) - \mathbb{E}_\xi(\hat{f})^2 + f^2 - 2f\mathbb{E}_\xi(\hat{f}) + \mathbb{E}_\xi(\hat{f})^2\right) \tag{3.8}$$
$$= \Delta + \mathrm{Var}(\hat{f}) + \mathbb{E}_{\mathbf{s}}\left((f - \mathbb{E}_\xi(\hat{f}))^2\right) \tag{3.9}$$
$$=: R_{p,\mu} \tag{3.10}$$
$$\Rightarrow R_p = \frac{1}{M}\sum_{\mu}^{M} R_{p,\mu} \tag{3.11}$$

where we have used the fact that the noise has zero mean and variance Δ and that the input $\mathbf{s}$ is independent of the noise $\boldsymbol{\xi}$ so $\mathbb{E}_\xi(f) = f$. Three quantities with transparent interpretation appeared:

- The *variance* $< \mathrm{Var}(\hat{f}) > =< \mathbb{E}_{\mathbf{s}}\left(\mathbb{E}_\xi(\hat{f}^2) - \mathbb{E}_\xi(\hat{f})^2\right) >$ which quantifies the average (over the signal realization) sensitivity of the estimator to fluctuations in the observations due to the noise. An high variance estimator would change a lot as a function of small perturbations in the observations and is thus not robust.

- The squared *bias* $< \mathbb{E}_{\mathbf{s}}\left((f - \mathbb{E}_\xi(\hat{f}))^2\right) >$ which represents the systematic error induced by the estimator if its average with respect to the noise differs from the true processing function f, thus a constant shift between the estimator and the observations.

- The *irreducible error* Δ which is the error induced by the precense of the noise (the noise is i.i.d thus $\Delta_\mu = \Delta$). It is called irreducible as it is purely random and inherently present in the observations, and thus cannot be canceled in any manner and should not be fitted.

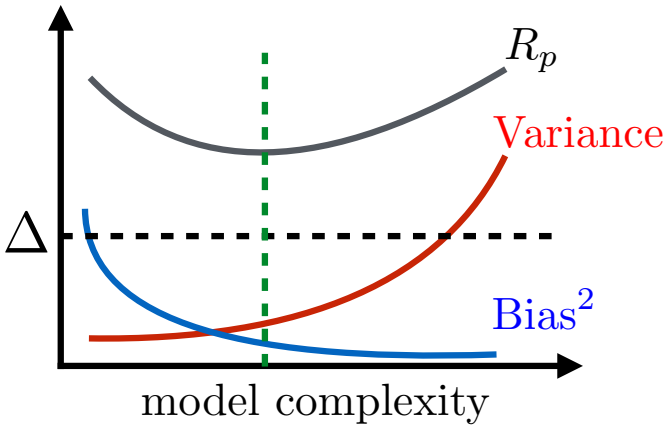

Figure 3.2 – Graphical representation of the bias-variance tradeoff. The horizontal axis is the model complexity, the black dashed curve is the irreducible error Δ and the grey curve, the prediction risk which is the sum of Δ, the red variance curve and the blue squared bias curve. The sum of the squared bias and variance terms is the reducible error, that can be asymptotically canceled if we have access to a very large number of data points or if one has directly access to the data generating model. The optimal estimator is given by the model with complexity corresponding to the minimum of R_p, represented by the green dashed line.

The sum of the variance and squared bias terms is the *reducible error* as it is the term that can be lowered by adjusting the estimator. This quantity is a function of the *model complexity*, i.e. of the number of degrees of freedom of $\hat{f}$ or its "roughness" in the non parametric case. As the model complexity increases, the bias decreases monotically as the observations are fitted more accurately but the pay-off is that the variance monotically increases until a point where it actually overcomes the gain in error due to the bias decrease. The optimal complexity of the estimator is the one that enables the estimator to fit enough the observations such that the bias is low but not too much such that the variance is not too high. This is summarized by the Fig. 3.2. The green line which represents the optimal complexity (i.e. which minimizes the prediction risk) separates an underfitting regime on its left from the overfitting regime. The associated optimal estimator is denoted as $\hat{f}_{opt}$. The prediction risk cannot fall under the irreducible error due to the noise, the black dashed line on the plot. The reducible error is the gap between Δ and the prediction risk at the optimal complexity on the plot Fig. 3.2.

What about the case of interest in the present thesis, the inference of the signal? The error estimate of interest in this case is the *risk* associated to the mean square error. Here one must be very careful: depending on the authors and especially on the adopted point of view (frequentist or Bayesian statistics), the risk can be defined in different ways. Here we place ourselves in the Bayesian framework and assume that we have access to data $\mathbf{y}$ but not to the true signal $\mathbf{s}$. We represent the signal by an auxilliary variable $\mathbf{x}$ to which we associate a posterior distribution $P(\mathbf{x}|\mathbf{y})$, see sec. 3.6.1 for details. In this framework, the definition of the risk $R(\hat{\mathbf{x}}|\mathbf{y})$ of an estimator $\hat{\mathbf{x}}$ is the average of its mean square error MSE with respect to $\mathbf{x}$

weighted by its posterior:

$$MSE(\hat{\mathbf{x}},\mathbf{s}) := ||\hat{\mathbf{x}}-\mathbf{s}||_2^2 = \frac{1}{N}\sum_i^N (\hat{x}_i - s_i)^2 =< (\hat{\mathbf{x}}-\mathbf{s})^2 > \tag{3.12}$$

$$R(\hat{\mathbf{x}}|\mathbf{y}) := \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{MSE(\hat{\mathbf{x}},\mathbf{x})\} = \frac{1}{N}\sum_i^N \int dx_i P(x_i|\mathbf{y})(\hat{x}_i - x_i)^2 \tag{3.13}$$

The true MSE (3.12) cannot be computed as $\mathbf{s}$ in unknown but $R(\hat{\mathbf{x}}|\mathbf{y})$ can be if we are able to estimate $P(\mathbf{x}|\mathbf{y})$. $P(x_i|\mathbf{y})$ is the marginal posterior of x_i. This risk, which is the one we refer to in this thesis, is linked to the so-called *Bayesian risk* $R_B(\hat{\mathbf{x}})$ as:

$$R_B(\hat{\mathbf{x}}) = \int d\mathbf{y} P(\mathbf{y}) R(\hat{\mathbf{x}}|\mathbf{y}) \tag{3.14}$$

where we average over the data for which it is supposed that we have a prior distribution $P(\mathbf{y})$. But in this thesis we always consider the data to be fixed. See [9, 10] for frequentist definitions of the risk.

In the special case of a linear orthogonal f and defining the inverse of f as $g := \text{inv}(f)$ one can write:

$$s_i = g(y_i - \xi_i) = g(y_i) + \bar{\xi}_i \tag{3.15}$$

which as the same form as (3.2) and where $\bar{\xi}_i$ is a new effective noise. In properly rescaled problems, $\bar{\xi}_i$ has also a variance $\in O(\Delta)$. Thus all the previous discussion and demontration remains valid (considering only the average over the noise) and we obtain that the prediction risk and the risk are equal up to a multiplicative factor $R_p = cR$ [9], and thus the three sources of error remain the same with identical interpretations. This is a priori not justified when f is not invertible as in the present thesis, where highly underdetermined linear systems will be studied but nevertheless, the three sources of errors actually remain the same. See [8] for a very nice introduction about statistical learning and the different sources of error in inference. [20] recently studied the influence of the reducible error in the prior mismatch case in the Bayesian framework, see sec. 3.6.

3.1.7 Another source of error: the finite size effects

There are two ways to cancel the reducible error: to have access to an infinite number of observations generated from the same system or having directly access to the data generating model. For example if one wants to infer the signal $\mathbf{s}$, P_{out}, f, and all the parameters of the problem $\boldsymbol{\theta}$ including those of the signal $\boldsymbol{\theta}_s$ in (3.2) must be known. But as the data is finite $N < \infty$, even when the model is perfectly known and thus the reducible error is inexistant, there can remain another source of error related to the algorithm that performs inference: the finite size effects that induce a *finite size error* $\epsilon(N)$, where $\lim_{N\to\infty}\epsilon(N) = 0$. In the present thesis, the inference algorithm that we will use is the approximate message-passing algorithm

derived and discussed in sec. 4.3. It is based on "law of large numbers-like" arguments and is only rigorous in the limit $N \to \infty$. Thus when using it on finite size systems, the algorithm becomes an approximation and this can induce finite size errors.

The optimal estimator (i.e. which has no reducible error) thus verifies:

$$\hat{x}_{i,opt} = s_i + \epsilon(N) + \bar{\xi}_i \tag{3.16}$$

where $\bar{\xi}_i$ is an effective error with variance $\in O(\Delta)$ as well when the system is properly scaled. This discussion shows that to get a good estimator, one must reduce as much as possible the reducible error and also study carefully the finite size effects associated to the inference algorithm used to get the estimate.

3.1.8 Some important parametric supervised inference problems

All the problems treated in the present thesis belong to the class of parametric supervised problems of the form (3.2). A non exhaustive list of such problems could include:

Denoising : $f(\mathbf{s}) = \mathbf{s}$, $P_{out}(\,|\boldsymbol{\theta}_{out})$
This is the simplest (in terms of definition) parametric statistical inference problem where one aims at reconstructing a corrupted signal by a noisy channel $P_{out}(\,|\boldsymbol{\theta}_{out})$, such as an AWGN channel of particular interest in this thesis: $P_{out}(\tilde{\mathbf{y}}|\Delta) = \tilde{\mathbf{y}} + \boldsymbol{\xi}$ where Δ is the variance of the AWGN $\boldsymbol{\xi}$ with i.i.d components of zero mean.

AWGN corrupted linear estimation problems : $f(\mathbf{s}|\mathbf{F}) = \mathbf{Fs}$, $P_{out}(\tilde{\mathbf{y}}|\Delta) = \tilde{\mathbf{y}} + \boldsymbol{\xi}$
This model, amongst which belongs AWGN compressed sensing and linear error correcting codes over the AWGN channel, is at the core of this thesis. This is also a parametric problem as there are few (possibly unknown) free parameters, here the noise variance and some others, parametrizing the prior knowledge about $\mathbf{s}$, see the section on Bayesian inference sec. 3.6 for more details about the notion of prior. Denoising over an AWGN channel can be seen as a particular instance of this problem where $\mathbf{F}$ is the identity matrix.

Binary symmetric or erasure channel models : $f(\mathbf{s}|\mathbf{H}) = \mathrm{mod}(\mathbf{Hs}, 2)$, $P_{out}(\tilde{\mathbf{y}}|\epsilon) = \mathbf{z}$
where $\mathrm{mod}(\,,2)$ is the modulo 2 component-wise operation and $z_i = \tilde{y}_i$ with probability $(1-\epsilon)$, $z_i = \star$ or $-\tilde{y}_i$ with probability ϵ for the binary erasure channel or the binary symmetric channel respectively. Here $\star$ means lack of information. These are classical models in communication theory.

Matrix completion : $f(\mathbf{S}) = \mathbf{S}$, $P_{out}(\tilde{\mathbf{Y}}|\epsilon) = \mathbf{Z}$
where $Z_{ij} = \tilde{Y}_{ij}$ with probability $(1-\epsilon)$, $Z_{ij} = \star$ with probability ϵ, usually close to one. This problem is fundamental in the field of recommander systems and collaborative filtering, where prior knowledge about the matrix $\mathbf{S}$ can be that it is low rank and one seeks for the missing entries. Despite its usefulness in common fields, this problem is different from the unsupervised tasks of clustering or community detection.

Classification : $f(\mathbf{z}|\mathbf{s},\mathscr{C}) = \left[c_{z_i}|c_{z_i} \in \mathscr{C}\right]_i^N$, $P_{out}(\tilde{\mathbf{y}}) = \tilde{\mathbf{y}}$
The classification problem is a canonical problem of parametric supervised learning. The classifier $f(z_i|\mathbf{s},\mathscr{C})$ ouputs the class $c_{z_i} \in \mathscr{C}$ to which belongs the item z_i. Here the vector to infer $\mathbf{s}$ is actually a set of parameters allowing the classifier to perform its task properly. To do so, one must have a set on inputs objects $\mathbf{z}^{train}$ and their associated classes $\{c_{z_i^{train}}\}_i^N$ to teach the classifier how to distinguish between the classes, i.e. learn $\mathbf{s}$ which basically draws plans between the different classes of $\mathscr{C}$ in the items space. This problem stands at the roots of modern image recognition and deep neural networks theory.

The inverse Ising problem : $f(\mathbf{s}|\mathbf{h},\mathbf{J}) = \{m_i, \{c_{ij}\}_{j\neq i}\}_i^N$, $P_{out}(\tilde{\mathbf{y}}|\epsilon,\Delta) = \tilde{\mathbf{z}}$
where $\tilde{z}_i = \tilde{y}_i + \xi$ with probability $1-\epsilon$, $\tilde{z}_i = \star$ with probability ϵ and $\xi \sim \mathscr{N}(\xi|0,\Delta)$. Here, one has access to partial noisy observations of the means $\{m_i\}_i^N$ and two point correlations $\{\{c_{ij}\}_{j\neq i}\}_i^N$ of the variables $\{s_i\}_i^N$ (for example measured experimentally). The aim is reconstructing the pairwise network of interactions $\{\{J_{ij}\}_{j\neq i}\}_i^N$ between these and the external fields $\{h_i\}_i^N$, such that the averages and correlations of the variables computed with respect to the Ising measure $P(\mathbf{s}|\mathbf{h},\mathbf{J}) \propto \exp\left(-\sum_{i,j\neq i}^{N,N} J_{ij}s_is_j - \sum_i^N h_is_i\right)$ matche the observed ones. This is a problem of great interest especially in phylogenetics and neurosciences. It is useful in any network reconstruction problems where one does not have access to higher than second order statistics about the variables that form the networks which is generally the case due to the restricted size of the samples in biology. This Ising form of the measure is derived by maximum entropy criterion as we shown in sec. 4.1.3.

Let us now focus on the specific problem of interest in this thesis, namely sparse linear estimation with i.i.d AWGN corruption, but it must be understood that the general methodology discussed hereafter (particularly Bayesian inference and message-passing algorithms) could be applied in most of these problems as they are special instances of the general model (3.2). For example, some references sharing the same methodology as the one developed in this work could include [21] for classification, [22] for clustering (interesting links between message-passing and spectral methods can be found [14,15]) or [23,24] for all the details about inference in communications over binary channels. See [12,25] and the references therein for applications of the inverse Ising problem, including in biology and [26] for a review of efficient algorithms to deal with this problem.

3.2 Linear estimation problems and compressed sensing

In this thesis, the studied problems belong to the class of noisy linear estimation problems under AWGN corruption which general form can be written as:

$$\mathbf{y} = \mathbf{F}\mathbf{s} + \boldsymbol{\xi} \tag{3.17}$$

$$\Leftrightarrow y_\mu = \sum_i^N F_{\mu i}s_i + \xi_\mu = (\mathbf{F}\mathbf{s})_\mu + \xi_\mu \ \forall \mu \in \{1,\dots,M\} \tag{3.18}$$

where $\xi_\mu \sim \mathcal{N}(\xi_\mu|0,\Delta)\ \forall \mu \in \{1,\dots,M\}$ and we place ourselves in the general continuous framework $\mathbf{F} \in \mathbb{R}^{MN}$, $\mathbf{s} \in \mathbb{R}^N$. Let us start by underlying some fundamental differences between interpreting (3.18) as a linear *fitting* model, i.e. we fit some data as a linear combination of some basis $\mathbf{F}$ vectors, the approximate coefficients being the signal we want to infer or as the true *generating* model of the data, i.e. we know that the data has been generated through (3.18) and we look for the unknown signal $\mathbf{s}$. Despite some common features in terms of algorithmic techniques that can be used for estimation in both cases, there is a deep difference in the behavior of the problem as in the latter case, there exist a particular state, the true solution $\mathbf{s}$, which can have a larger statistical weight than any other approximate solutions. In statistical physics, the construction of a problem from a given solution such as in inference defines the so called *planted ensemble*, see [24, 27] for the statistical physics of this ensemble.

3.2.1 Approximate fitting versus inference

For this discussion, we place ourselves in the noiseless setting $\Delta = 0$. Algebra tells us that the system (3.18) can be solved exactly if the number of measurements is *at least equal* to the number of variables $M \geq N$ and are linearly independent one of the other. This can be done by simple matrix inversion, selecting a subset of N lines of the original matrix and the associated measures and to inverse the resulting system. But this is true only if the observations *were really generated* through a linear model such as (3.18), which implies that there indeed exists a solution to the system. But one could just have access to data without really knowing its generating process and want to approximately fit these data as a linear combination of some basis functions or atoms (the columns of $\mathbf{F}$), that form the so-called dictionnary in this context.

Let us consider (3.18) as an approximation model. If the number of observations is $M < N$, the system is said *underdetermined*: there are too many possible solutions, usually an infinite number. If in contrast, $M > N$ the system is *overdetermined* and there could be no solution verifying all the observations exactly. To deal with these situations, one can use the *least square estimate* $\hat{\mathbf{x}}_{LS}$. It consists in finding the linear combination of the basis functions that minimize the empirical prediction risk (3.3):

$$\hat{\mathbf{x}}_{LS} = \underset{\mathbf{x}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{F}\mathbf{x}||_2^2 \tag{3.19}$$

$$= \operatorname{inv}\left(\mathbf{F}^\intercal \mathbf{F}\right) \mathbf{F}^\intercal \mathbf{y} \tag{3.20}$$

$$:= \mathbf{F}^* \mathbf{y} \tag{3.21}$$

$\mathbf{F}^* := \operatorname{inv}(\mathbf{F}^\intercal \mathbf{F})\, \mathbf{F}^\intercal$ is the so-called pseuso-inverse of $\mathbf{F}$. This solution has some caveats such as the fact that it won't produce any sparsity in the solution which can make the interpretation of the solution quite complicated as previously discussed in sec. 3.1.5. More advanced methods such as the sparsity inducing LASSO which will be presented in sec. 3.4, solved by linear programming techniques, can be used to improve the interpretability by sparsifying the solution. If an approximately sparse solution is found, it means that the data can be thought essentially as linearly depending on a small subset of features, the basis vectors associated

to the estimated non zero components of the fit $\hat{\mathbf{x}}$, selected among an original larger set of possibilities: this is called model selection. An ensemble of advanced techniques exist to deal with fitting linear models, details can be found in [8, 10].

But what happens if we know that the data has actually been generated by the linear model (3.18)? The notion of overdetermined system loses its senses. For any $M \geq N$, finding the solution is trivial because it *exists*. Inference for sparse linear models thus deals with situations where $M < N$, or even $M \ll N$ but we know that the linear system have a solution for sure as the data was produced in this way, which makes all the difference. One could think of the least square estimate as a strategy, but in an inference problem, the aim is not to minimize the prediction risk but the mean square error (3.12) and a minimum prediction risk solution can (and *will* in most of the cases) have a very high MSE.

So how to do? Still, albebra requires as many constraints as variables to infer over. Hopefully, additional input information about the solution can counterbalance the missing constraints: sparsity is our savior.

3.2.2 Sparsity and compressed sensing

A new paradigm in signal processing is the notion of *sparsity* and *compressibility* : a signal is said to be K-sparse if there exist a basis $\mathbf{\Psi}$ in which the representation of the signal in it has only K components that are non zero, that we call its support. A K-compressible signal is a signal that is "well" approximated by a K-sparse one. More precisely, if the signal is approximated by keeping only its K components with largest amplitude in an appropriate basis, then the ℓ_p norm of the difference between this approximation and the signal in this basis decays as a power law:

$$||\mathbf{s}_K - \mathbf{s}||_p < CK^{-u} \tag{3.22}$$

for some constants C and $u > 0$ where $\mathbf{s}_K$ denotes the best K-sparse approximation of the compressible signal $\mathbf{s}$. It basically means that the amplitude of the sorted components of $\mathbf{s}$ decays at least as a power law, see [28] for more details on this notion. In this thesis, I will sometimes use the terminology of sparse signals even for compressible ones.

Compressed sensing, introduced 10 years ago in a series of papers by Donoho, Candès, Tao and Romberg [29–32] is the theory and ensemble of techniques behind the measurement protocol and reconstruction process of sparse and compressible signals from few (noisy or not) linear observations. Mathematically speaking, it is the field of research focused on solving a priori undetermined systems of linear equations, using sparsity assumptions about the solution. A very nice and simple introduction to compressed sensing (seen from the convex optimization point of view) can be found in [33], see [34, 35] for the probabilistic point of view as adopted in this thesis.

Another fundamental aspect of compressed sensing, complementary to sparsity, is the notion

of *coherence*. In order to be able to infer back the sparse signal **s** from a model like (3.18), the measurement (or "sensing") matrix **F** must be as incoherent as possible with the sparsifying basis $\mathbf{\Psi}$ of **s**. It means that each basis vector of **F** must be as orthogonal as possible to *all* the basis vectors of $\mathbf{\Psi}$ at the same time or equivalently any basis vector of **F** *cannot* be expressed nor well approximated by a sparse linear combination of the basis vectors of $\mathbf{\Psi}$. In this way, all measurements y_μ will contain an approximately equal amount of information about all the components of **s** expressed in $\mathbf{\Psi}$. Intuitively, the $O(K)$ measurements select a set of possible solutions to the linear system (3.18) and the sparsity assumption select among these the sparsest one which can be unique in the noiseless setting. As this solution has only K non zero values, the $O(K)$ measurements are enough to fix their amplitudes. The coherence between two matrices **A** and **B** is formally defined as:

$$\mu(\mathbf{A},\mathbf{B}) = \sqrt{N} \max_{1\leq k,l\leq N} |(\mathbf{A}_{\bullet,k})^\intercal \mathbf{B}_{\bullet,l}| \tag{3.23}$$

which is thus a direct measure of the correlation between the matrices.

Constructing a maximally incoherent sensing matrix for a given sparsifying basis $\mathbf{\Psi}$ is a computationally very hard problem and cannot be solved in general. But here the intuition suggesting that a purely random sensing matrix (that we will always take i.i.d Gaussian) must be highly incoherent with any $\mathbf{\Psi}$ with high probability (i.e. tends to one as the matrices size diverge) is valid. Indeed, drawing a random i.i.d Gaussian matrix will give rise to basis vectors uncorrelated with the $\mathbf{\Psi}$ ones, exactly as a white noise has a flat spectrum in any basies. This is one among many others advantages of working in high dimensions. It is quite instinctive to see that in a very high dimensional space parametrized by some basis, if you draw a random vector, it has very low probability to be close to aligned to one of the vector basis as there are so many available directions. Working with random Gaussian i.i.d matrices has another great advantage: it allows to perform analytical predictions in the large size limit $N \to \infty$ using techniques mainly based on "law of large numbers like" arguments, standing at the roots of the state evolution analysis, see sec. 5.3.

Reconstructing **s** from the knowledge of the measurement matrix **F** and few AWGN corrupted observations **y** given by (3.18) is thus a compressed sensing problem as long as **s** is sparse and **F** is "random enough" (this notion will be studied in great details in the chapter about structured operators in compressed sensing, chap. 7). The knowledge of the sparse nature of the signal allows one to solve the reconstruction problem in an efficient manner.

What does an efficient manner actually means? First, compressed sensing theory shows that for a K-sparse signal of size N, its reconstruction can be performed from a number of linear observations that grows with K instead of N. So for very sparse signals with a density $\rho := K/N \ll 1$, the theoretically required number of measurements can be very low. This is litterally a revolution in signal processing as it overcomes the Shannon-Nyquist theorem that states that if some physical signal's highest represented frequency is f (the so-called Nyquist rate), i.e. its Fourier coefficients are all zeros for any frequency higher than f, then at least

$2f$ discrete samples of this signal are required for perfect reconstruction. Basically, it means that *without any prior knowledge about a signal* of size N in its discrete representation, $O(N)$ measurements are required to estimate it, and this independently of its true informational content, usually carried by a small support of size $K \ll N$.

Assume that you want to measure a pure sinusoidal signal with very high frequency oscillations. The Shannon-Nyquist theorem a priori constrain you to perform many measurements of this signal. But if in addition you have now the prior knowledge that this signal is a pure sinusoide, you can use this supplementary information on the sparsity of this signal in the Fourier space to infer it from far fewer measurements. This is what compressed sensing is all about: performing the very least number of operations for estimating a signal.

The terminology compressed sensing comes from this paradigmatic change. In usual sensing/compression strategies, one performs many measurements, where again "many" is fixed by the Shannon-Nyquist theorem. Once the signal has been estimated, one can try *a posteriori* to find a sparsifying basis for it and thus locate its zero or small components: this is performed thanks to compression algorithms such as JPEG2000 for images or Fourier analysis for sounds. At the end one stores only the informative support of the signal. This two-steps procedure is quite inefficient: most of the $O(N)$ performed measures contain highly redundant information as the zeros of the signal do not contribute to them, and thus they correlate only the K informative components. This is why *afterwards* compression can be done, thanks to this inherent redundancy contained in the observations obtained by usual sampling techniques applied to sparse signals.

Compressed sensing is an "all in one" procedure that optimally uses all the knowledge one have about the measured signal, at least in the Bayesian framework (convex optimization procedures discussed in sec. 3.4 usually only assume sparsity as opposed to Bayesian inference sec. 3.6 that allows to integrate more information in the model). By carefully designing the measurement protocol, one can maximally reduce the redundancy inside each sample which drastically lower the required number of them. This allows to directly reconstruct the most compressed form of the signal in a fixed sparsifying basis, chosen thanks to the a priori knowledge about the signal: *The signal I want to measure is a natural image, so I know that it is a priori sparse in the wavelet basis. I will thus try to directly estimate the few important coefficients of the signal in this basis. My signal is a sound, so it should be sparse in the Fourier basis, etc.*.

3.2.3 Why is compressed sensing so useful?

In many applications, measurements can be costly in energy, time and money and reducing the required number of such samples can have a great impact. But one could ask that despite compressed sensing is a beautiful mathematical theory that should allow for great improvements in signal processing, is it actually relevant for real life applications? The anwser is yes. The notion of sparsity or compressibility of a signal is not only a great theoretical property but

also an actual feature of virtually all signals in the nature. The following hypothesis appears to be almost always validated: if a signal carries some information, it must have some kind of structure in an appropriate representation, i.e. it is not pure noise, the only signal for which it exists no sparsifying basis. We will focus in this manuscript on some important applications such as image reconstructions (see chap. 8) and error correcting codes for communications (see part. IV) but we can name many others. A non hexaustive list of relevant examples could include:

- Medical imagery such as in magnetic resonance imagery [36] where the measurements are very long and costly. Sparsity of the image in the wavelet or discrete cosine basis can be exploited to lower the number of required measured Fourier coefficients to reconstruct good quality images.

- Deep space imagery and radio interferometry [37]. Probing the deep space is highly costly as massive telescopes are required, preferably all over the world to perform independent measurements. But many features of interest are highly sparse in appropriate basies such as intensity fields of compact astrophysical objects or the imprint of cosmic strings in the temperature field of the cosmic microwave background radiation.

- New optical devices such as one pixel cameras [38]. Imagine that some highly accurate new sensor is really costly. Thanks to compressed sensing, a unique sensor can be used to measure images.

- Still in the field of optics, people are nowadays trying to use the randomness of physical media such as layers of white paint as the compressive imaging device [39]. Scattering media are thus promising candidates for designing efficient and compact compressive imager. In parallel, algorithms for the estimation of the sensing matrix generated by natural scattering media are developed [40].

- Compressed sensing is also applied in acoustics, for example in problems of vibrating source localization [41] or sampling of the 3-d acoustic field and the room impulse responses [42], that characterize the reverberation properties of the room. In both applications, the number of microphones required can be greatly reduced using compressed sensing techniques because, for example, the room impulse responses can be considered sparse in the time domain. This could be applied in virtual reality, video games and electronic music, where the use of a space-varying reverberation extracted from real environments improve the impression of immersion.

- Compressed sensing has a great potential in group testing with applications in genetic screening, compressed genotyping or optimal blood testing [43]. Imagine you have many samples of blood and you know that few of them are infected. How to optimally mix samples to reduce the number of required tests to find the infected samples? Compressed sensing theory answers this question.

- The applicability of compressed sensing theory is nowadays also extended to problems with high computational cost. It can help to greatly lower the complexity of some matrix reconstruction problems such as the computation of sparse Hessian matrices, useful in density functional theory [44].

The list could go on for a while: radar detection, efficient measurements and reconstructions of complex wave functions, data compression, efficient analog-to-information conversion, etc. A massive list of applications can be found online: `http://dsp.rice.edu/cs`, `http://nuit-blanche.blogspot.fr`.

Summing up, compressed sensing theory answers the questions: how one should design an optimal measurement process for a signal for which we have some information, such as sparsity. And how to reconstruct it efficiently from this few optimal measures. Compressed sensing will ouput the sparsest solution to a problem, and is in this sense a kind of modern Occam's razor: it finds the "minimal solution" to the problem and this from the minimal amount (or close to) of information theoretically required to solve the problem.

Let us now focus on the second fundamental efficiency aspect of compressed sensing. Now it is well defined, one could argue: *All that is really nice, at least on the paper, but can the estimation problem be actually solved in an amount of time which is not of the order of the age of the Universe?* In other words, are there efficient algorithms that can solve a given compressed sensing problem in a fast way? Fortunatly yes. An extensive subfield of research in compressed sensing focuses on developing such computationally efficient and yet accurate solvers, like in the present thesis. We will expose the two actual main ways of solving a compressed sensing problem, but before that we present some basic notions of complexity theory.

3.3 The tradeoff between statistical and computationnal efficiency

With the explosion of the size of the data sets generated in modern scientific, medical, social and economical applications, a major point to consider in todays inference techniques is the tradeoff between *statistical* and *computationnal* efficiency. A statistically efficient algorithm is able to infer the desired quantity accurately while remaining robust in spite of the precense of noise. This can be quantified by error estimators such as the MSE (3.12). A computationnally efficient algorithm has a low complexity, i.e. it requires a number of fundamental operations to perform its task that scales "nicely" with the size of the input and output data. Lets precise this point by introducing the very basics of complexity theory, without the goal to be complete nor rigorous.

3.3.1 A quick detour in complexity theory and worst case analysis : P $\neq$ NP ?

Complexity theory aims at grouping into complexity classes the set of all problems that can be solved by computers. It exists a full zoology of such complexity classes. Let us formalize a

bit this idea. Let us denote a generic problem by Ψ. For example, considering the core of this thesis, the AWGN corrupted linear estimation problem, Ψ would be given by Ψ = "find $\mathbf{s}$ from the generic model (3.18)". ψ denotes a given *realization* or *instance* of Ψ which would be in the present case a particular realization of the random noise vector $\boldsymbol{\xi}$, of the sensing matrix $\mathbf{F}$ and of the measured signal $\mathbf{s}$ in (3.18).

Complexity theory is based on the notion of *worst case analysis* : a problem Ψ belongs to a given complexity class if its "*most difficult instance*" $\tilde{\psi}$ belongs to this class. The most difficult instance is basically the one that requires the largest number of operations to be solved among all the instances $\{\psi\}$ of Ψ. The main complexity classes of interest for us are the so-called polynomial time P and non-deterministic polynomial time NP classes. In order to define them properly, we would need to introduce concepts such as Turing machines which are out of the scope of the present thesis and can be found in any text book on complexity theory or computer science such as the very nice books [18, 45]. Let us define them in a more handwavy way.

We define a problem Ψ to belong to the P class if there exist an algorithm able to solve *any* of its instances $\{\psi\}$ performing a number of fundamental operations (such as additions, multiplications, etc) that scales as a polynomial in the size of the problem $O(N^k)$ where N is the number of variables of the solution we are looking for. We would say that a problem in P is "easy" i.e. can be solved efficiently. This idea of simplicity of problems in P is referred as the Cobham's thesis. Such problems include testing whether a number is prime, calculating the greatest common divisor or finding a maximum matching in a graph. In contrast, problems in NP are usually referred as "hard" problems. A problem is said to belong to the NP class if no algorithm is known *yet* to solve all of its instances (inluding the hardest one) in a number of operations bounded by a polynomial in the size of the solution, but if one is given an a priori solution to any instance, it can be checked efficiently if this is actually a solution as declared or not. The complexity of NP problems usually scale as an exponential in the problem size $O(e^N)$ which becomes *very* quickly intractable by brute force combinatorial methods.

From the moment an algorithm can provably solve efficiently a problem, it is known to be in P but it is really difficult to assert that a problem is *not* in P, or is in NP as it would require to prove that no efficient algorithm exist for the hardest instances of this problem. This emphasize the most fundamental question of complexity theory and computer science: Is $P \neq NP$? No one knows a proof or disproof of this assertion despite that most scientists think nowadays that there actually exists a fundamental difference between these two classes, i.e. there are problems that *are* really difficult, and will remain difficult in the future.

3.3.2 Complexity of sparse linear estimation and notion of typical complexity

All these are very general considerations, but what about the core problem of this thesis (3.18)? Unfortunately, this problem is thought to be NP which makes it a priori intractable. This is also what makes this problem interesting from a theoretical point of view, in addition of its

great applicability as we have seen in sec. 3.2.3. Fortunatly there exist efficient ways to solve such problems, and it is the subject of this thesis. As we said, complexity classes are based on the notion of worst cases, but what about *typical* cases? A typical case is an instance $\bar{\psi}$ that you would pick by selecting one at random in the set of all possible instances $\{\psi\}$ of the problem Ψ when N is large. It represents a kind of "average case". In many inference or combinatorial optimization problems, hard instances can be quite paradoxally very difficult to generate, most of the instances being easy: the hardest instances define the problem Ψ to be formally in NP but the typical instances place Ψ inside P in an effective way. And hopefully, many problems have lower typical complexity in proper regimes than what their worst case analyzes tell us. Further details about the typical complexity of the problem will be presented when discussing the phase diagram of the problem in sec. 5.1.

3.4 The convex optimization approach

How to solve a compressed sensing problem efficiently? The first methods were based on convex optimization methods. Let us review the very basics of this appoach that is *not* the kind of methods used and studied in this thesis but quickly presented here for sake of completeness. Very nice and complete reviews can be found such as [8, 46–48].

3.4.1 The LASSO regression for sparse linear estimation

The goal is thus to find the sparsest solution among a huge set of admissible solutions to (3.18). Mathematically one must be able to find the (hopefully) unique solution $\hat{\mathbf{x}}_0$ to (3.18) that has the smallest ℓ_0 norm which is defined as the number on non zero components of a vector. Unfortunately, this problem is NP and non convex: there exist no convex function such that its minimum is provably given by $\hat{\mathbf{x}}_0$ for any measurement rate $\alpha := M/N > \rho := K/N$, the information theoretical bound. To face this, the problem is relaxed by replacing the constraint of ℓ_0 minimization to an ℓ_1 minimization, which makes the problem convex. This problem is referred as the LASSO (Least Absolute Shrinkage and Selection Operator) regression:

$$\hat{\mathbf{x}}_1 = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{x}||_1 \text{ such that } \mathbf{y} = \mathbf{F}\mathbf{x} \tag{3.24}$$

which becomes in the noisy setting

$$\hat{\mathbf{x}}_1 = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{x}||_1 \text{ such that } ||\mathbf{y} - \mathbf{F}\mathbf{x}||_2^2 < \epsilon \tag{3.25}$$

$$\Leftrightarrow \hat{\mathbf{x}}_1 = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{F}\mathbf{x}||_2^2 + \lambda ||\mathbf{x}||_1 \tag{3.26}$$

where the second form is totally equivalent to the first one for a properly selected slack parameter λ, which is here to balance the relative weight of the sparsity constraint with the observations. ϵ is some small threshold relaxing the hard contraint of perfectly verifying the linear constraints, which takes into account the precense of noise. It appears that for a

measurement rate $\alpha > \alpha_{DT}(\rho) > \rho$, the solution to the LASSO problem is provably equal to the one of the intractable ℓ_0 optimization problem: $\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_0$. $\alpha_{DT}(\rho)$ is called the Donoho-Tanner transition [49] and defines the best performances one can reach asymptotically with convex optimization based methods, see sec. 5.1.1: polynomial-time optimization solvers or greedy algorithms can solve the problem from $O(K\log(N/K))$ measurements, not $O(K)$. Unfortunately this transition is far from the optimality $\alpha = \rho$ and this is why we need different methods to improve the results such as Bayesian inference (sec. 3.6) combined with spatial coupling (sec. 5.5) which is asymptotically optimal. If one knows some more information about the solution than the simple sparsity, more advanced models such as group sparsity, tree structures, etc can be constructed and solved by convex optimization [28]. This can reduce the number of required measurements but in the general setting of purely sparse signals, convex optimization based approaches are *not* optimal from an information theoretical point of view.

3.4.2 Why is the ℓ_1 norm a good norm ?

Why is the ℓ_1 norm the appropriate sparsity inducing one? Smaller p-norm are non convex but nothing prevents from taking a higher order norm such as the ℓ_2 one. In this case the convex problem to solve is:

$$\hat{\mathbf{x}}_2 = \underset{\mathbf{x}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{F}\mathbf{x}||_2^2 + \lambda ||\mathbf{x}||_2^2 \tag{3.27}$$

Solving (3.27) is refered as the *ridge regression* problem [8]. Let us try to understand why is the ℓ_1 norm minimization the proper choice, i.e. why LASSO overcomes ridge regression in solving sparse linear estimation problems. Let us define $\mathbf{s} = [1, \epsilon]$.We assume $\epsilon > 0$ without lose of generality. Its ℓ_1 and ℓ_2 norms are:

$$||\mathbf{s}||_1 = 1 + \epsilon \tag{3.28}$$
$$||\mathbf{s}||_2^2 = 1 + \epsilon^2 \tag{3.29}$$

What happens to these if we reduce one of its component by a small positive quantitiy δ such that $\delta < \epsilon \ll 1$.

$$||\mathbf{s} - [0, \delta]||_1 = 1 + \epsilon - \delta = ||\mathbf{s}||_1 - \delta \tag{3.30}$$
$$||\mathbf{s} - [\delta, 0]||_1 = 1 + \epsilon - \delta = ||\mathbf{s}||_1 - \delta \tag{3.31}$$
$$||\mathbf{s} - [0, \delta]||_2^2 = 1 + \epsilon^2 - 2\delta\epsilon + \delta^2 = ||\mathbf{s}||_2^2 - 2\delta\epsilon + \delta^2 \tag{3.32}$$
$$||\mathbf{s} - [\delta, 0]||_2^2 = 1 + \epsilon^2 - 2\delta + \delta^2 = ||\mathbf{s}||_2^2 - 2\delta + \delta^2 \tag{3.33}$$

So we now understand that reducing small components almost does not affect the ℓ_2 norm that tries in opposite to reduce larger ones and spread the power and thus would prevent sparsity whereas the ℓ_1 norm is affected in the same manner as small or large components are reduced, allowing to cancel easier components that should be. So it is more that higher order norms prevent sparsity, ℓ_1 optimization does not.

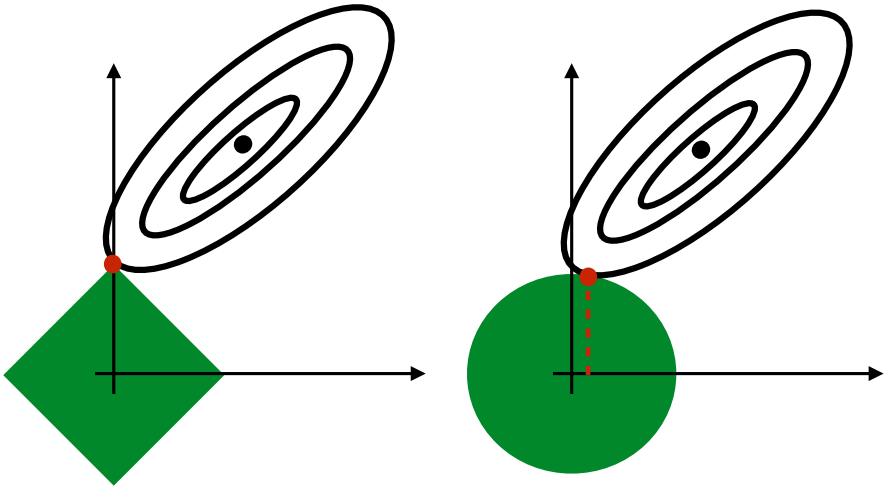

Figure 3.3 – Geometrical interpretation of why the ℓ_1 norm is the appropriate sparsity inducing one. The ellipses are iso-MSE lines. The green parts represents subspace with bounded ℓ_1 and ℓ_2 norm on the left and right plots respectively. The red point is the intersection between the further allowed iso-MSE line and the region bounding the norm of the vectors inside. We see that the intersection point which is the final estimate of the procedure cancels a components in the ℓ_1 optimization case whereas it would not choosing the ℓ_2 norm.

It can be understood in a more geometrical way as well. The optimization problem (3.25) has two parts. The first one enforces the linear observations to be fulfilled up to some error ϵ. This selects a subspace with smooth bounds (due to the ℓ_2 norm) such as the ellipses represented on Fig. 3.3 which extent is fixed by the relaxation parameter ϵ. Then one has to find the vector with smallest norm intersecting this region (the red point on the figure): it is the estimate $\hat{\mathbf{x}}$. Vectors with a bounded ℓ_2 norm belong to a ball (the disk on the figure), whereas with an ℓ_1 norm they belong to a polytop with sharp corners on the axis of the frame. So with high probability, the intersection point between the two regions will be on a axis in the ℓ_1 norm case and thus some components will be put to zero, whereas no components will be canceled in the ℓ_2 norm case due to its smooth nature. This phenomenon is even more pronounced in higher dimensions.

3.4.3 Advantages and disadvantages of convex optimization

In the present thesis, the methods that we will use are based on the Bayesian inference (sec. 3.6) due to its improved performances and phase transitions, see sec. 5.1.1. Nevertheless convex optimization methods are still used by many people and the field of research still very active. This is because despite not being optimal, convex optimization approaches have many advantages. The first one is the fact that optimization of the form (3.25) can be easily converted into linear programs and there exist a massive set of very efficient techniques and solvers combined with a well known theory of linear programming [46, 50]. Furthermore, by

the very definition of what a convex problem is, the problem to solve (3.25) always have a unique well defined solution. Modern solvers such as ℓ_1 magic [51] or NESTA [52] can solve compressed sensing instances in a quite fast way with convergence guarantees.

Another strong advantage of these methods are their robustness. Robust in the sense that a given solver can be used for a very large class of convex problems. "Details" of the instance such as the sensing matrix realization or the structure of the signal that can be more complicated than just sparse are mostly irrelevant: any convex optimization solver will do the job and output a solution, despite being usually not the best one in the sense that more advanced solvers, such as based on Bayesian inference, could have found a lower MSE solution for the same problem if properly used. The pay-off is that Bayesian inference requires the computation of problem dependent quantities, but it is what makes it more powerful as well: a Bayesian solver is specifically designed for a problem and thus performs generally better than a more general convex optimization solver.

3.5 The basics of information theory

Before to present the methodology of Bayesian inference, which is a statistical method, we present some fundamental concepts of information theory. The aim here is not completeness but really focus on the notions relevant for the present thesis and that are deeply connected to the statistical physics concepts. Very nice and complete books can be found as [7, 23] for a computer science and communications point of view and [24, 53] for an emphasize on the links with statistical physics. In this section, we use capital letters for the random variables, small letters for their realizations, or events.

3.5.1 Incertitude and information: the entropy

The fundamental object to quantify the information carried by some random variable $X \sim P_X(x|\boldsymbol{\theta})$, interpreted as a message sent to a receiver in the field of communication theory, is its *Shannon entropy* [54] or just entropy defined as:

$$H(X|\boldsymbol{\theta}) := \mathbb{E}_X\left(\log_2\left(1/P_X(x|\boldsymbol{\theta})\right)\right) = -\int dx P_X(x|\boldsymbol{\theta}) \log_2\left(P_X(x|\boldsymbol{\theta})\right) \tag{3.34}$$

The object $\log_2(1/P(x|\boldsymbol{\theta}))$ can be interpreted as the *suprise* of the event x: the less probable x is, the larger the surprise of observing it is. The informative content of X is its average surprise, the entropy. It can be interpreted the other way around: the entropy measures how much incertitude (i.e. lack of information) we have about X before its realization x is observed, thus how much information we gain once observed. We thus speak of information, uncertainty or incertitude in the same manner, defined as the entropy. Information is more appropriate when X has actually been observed, and incertitude when it has not yet. We could say that there is a fundamental conservation law linking information to incertitude: the information gained in observing some random variable realization compensates exactly the incertitude we

had about it before the observation, measured by the entropy. We can make a parallel with an isolated mechanical system. Its total energy is conserved, and its dynamic is the result of the conversion of potential energy into kinetic one. The incertitude can be interpreted as the potential energy and the information as the kinetic one.

A deterministic event has no entropy: we know everything about it, and thus gain nothing when observed. The other extreme case is the equidistributed random variable: we dont have any clue about what we will observe so our uncertainty about it reaches its maximum and thus we gain a maximum information observing its occurence. This is formalized by the second equality that verifies entropy in the next properties.

Why is this logarithm? It can be justified by rational arguments. Actually, the entropy is the *only* function verifying a number of necessary conditions for a coherent definition of what information is, including the previous remarks. We assume that X have n possible outcomes (we skeep the possible dependence on parameters):

$$H(X) \geq 0 \text{ with equality only if X is deterministic.} \tag{3.35}$$

$$H(X) \leq \log_2(n) = H(U),\ U \sim 1/n, \text{ the constant distribution over } n \text{ events.} \tag{3.36}$$

$$H(X,Y) = H(Y,X) \tag{3.37}$$

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{3.38}$$

$$H(X,Y) \leq H(X) + H(Y) \text{ with equality only if } P_{XY}(x,y) = P_X(x)P_Y(y) \tag{3.39}$$

$$H(Z|X,Y) \leq H(Z|X) \text{ with equality only if } P_{ZY}(z,y) = P_Z(z)P_Y(y) \tag{3.40}$$

The logarithm is in base 2 because the natural unit of information in communication is the bit as messages are usually coded in binary form. Thus an equiprobable binary random variable carries 1 bit of information by definition. This convention can also be justified interpreting the entropy as the number of dichotomic operations to perform (or minimum number of necessary yes/no questions to ask) to find the answer to a problem where all the answers are equiprobable, so in the worst case in a sense.

The first equality tells that there is no such thing as negative information: we cannot lose information from any new observation of a random variable realization, at worst we gain nothing in the case of a deterministic event. Let us demonstrate the fourth one for the example:

$$H(X,Y) = -\int dxdy P_{XY}(x,y)\log_2\left(P_{XY}(x,y)\right) \tag{3.41}$$

$$= -\int dxdy P_{XY}(x,y)\log_2\left(P_{X|Y}(x|y))\right) - \int dxdy P_{XY}(x,y)\log_2\left(P_Y(y)\right) \tag{3.42}$$

$$= \underbrace{-\int dy P_Y(y)\int dx P_{X|Y}(x|y)\log_2\left(P_{X|Y}(x|y))\right)}_{:=H(X|Y)} - \int dy P_Y(y)\log_2\left(P_Y(y)\right) \tag{3.43}$$

$$= H(X|Y) + H(Y) \tag{3.44}$$

where we skept the possible dependency in parameters. This equality tells that the total information or incertitude carried by a couple of random variables (X, Y) can be decomposed as the entropy of Y plus the conditional entropy $H(X|Y)$: the remaining incertitude about X once Y has been observed, or equivalently the additional information that the X observation would bring that has not already been obtained through the Y observation alone. The role of X and Y can be switched from the third equality. The fifth equality tells that the incertitude about a couple of random variable is maximum when they are totally independent, so knowing one's realization does not help to infer anything about the other one. Finally the last equality, that seems natural now we understand what information and incertitude means, tells that knowledge about more random variables realizations can only lower the incertitude about another one, and at worst does not bring any new information when they are independent. Equivalently it means that the information brought by observing a random variable realization having already observed two other ones cannot be larger than the information brought by observing it having already observed just one other random variable realization.

Links with the Boltzmann entropy $S = k_B \log(W)$ of statistical physics can be established (W is here the number of microstates of the system). From the information theoretical point of view, it can be interpreted as the number of bits (up to a multiplicative constant as the natural logarithm is used in physics) required to encode all the accessible microstates of the physical system with constant energy, i.e. in the microcanonical ensemble. The Boltzman constant is just here to fulfill the dimensionality requirement that the Boltzmann entropy times the temperature has the dimension of an energy. The entropy of a physical system can also be seen as its associated incertitude as it quantifies the information one would gain by measuring precisely its microstate, which is impossible in practice.

3.5.2 The mutual information

Now we have formalized the measure of information or incertitude carried by random variables, a natural question is the definition of a measure $I(X, Y)$ of the degree of correlation between random variables, i.e. how much information observing one of them brings about the other one: the *mutual information*. One could think that the previous definition of conditionnal entropy would do the job, but it is not symmetric with respect to X and Y.

Assume you first observe Y (the output of a noisy channel for example), then what is the remaining incertitude $H(X|Y)$ on X (the sent codeword)? It is the total incertitude about X, $H(X)$ minus the information gained about X (or equivalently minus the incertitude lost about X) from the observation of Y's realization, $I(X, Y)$. Thus:

$$
\begin{aligned}
H(X|Y) &= H(X) - I(X, Y) && (3.45)\\
\Rightarrow I(X, Y) &= H(X) - H(X|Y) && (3.46)\\
&= H(X) + H(Y) - H(X, Y) && (3.47)\\
&= H(Y) - H(Y|X) && (3.48)
\end{aligned}
$$

where the two last equalities were obtained using the property (3.38) of the entropy. We obtain a symmetric measure of the information contained in each variable about the other one, or equivalently how much the observation of one of the two variable reduce the incertitude on the other one. The equality (3.47) can be interpreted rewriting it as $H(X,Y) = H(X) + H(Y) - I(X,Y)$: the total information carried by the couple (X,Y) is the sum of the individual informations minus the information counted twice, the mutual information, due to the correlations between the two variables. The mutual information is null when the two variables are independent, so observing one does not bring any information about the second one. Thus mutual information can be seen as a measure of how much the correlated couple of variables with probability measure $P_{XY}(x,y)$ deviates from independent ones with measure $P_X(x)P_Y(y)$. Is there a way to formalize this notion? The answer is given by the Kullback-Leibler divergence.

3.5.3 The Kullback-Leibler divergence

The appropriate object for estimating "distances" between distributions $P(\mathbf{x})$ and $Q(\mathbf{x})$ is the Kullback-Leibler divergence. It measures how well the distribution Q describes the probabilistic structure of P. It is defined as:

$$KL(P||Q) := \mathbb{E}_P\left(\log_2\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right)\right) = \int d\mathbf{x} P(\mathbf{x}) \log_2\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \tag{3.49}$$

This is not really a distance as it is not symmetric nor verify the triangular inequality but still verifies the properties we are interested in: $KL(P||Q) \geq 0$ with equality only if $P = Q$. Now we can estimate how much P_{XY} differs from a factorizable measure $P_X P_Y$ like if (X,Y) were independent:

$$KL(P_{XY}||P_X P_Y) = \int dxdy P_{XY}(x,y) \log_2\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) \tag{3.50}$$

$$= \int dxdy P_{XY}(x,y) \log_2\left(\frac{P_{X|Y}(x|y)}{P_X(x)}\right) \tag{3.51}$$

$$= \int dxdy P_{X|Y}(x|y)P_Y(y) \log_2\left(P_{X|Y}(x|y)\right) - \int dxdy P_{XY}(x,y) \log_2\left(P_X(x)\right) \tag{3.52}$$

$$= \int dy P_Y(y) \int dx P_{X|Y}(x|y) \log_2\left(P_{X|Y}(x|y)\right) - \int dx P_X(x) \log_2\left(P_X(x)\right) \tag{3.53}$$

$$= -H(X|Y) + H(X) \tag{3.54}$$

$$= I(X,Y) \tag{3.55}$$

where we have used the marginalization property $\int dy P(x,y) = P(x)$. We find back the mutual information, the measure of how much random variables are correlated, i.e. how much their joint distribution is "far" from the factorized form: the greater the Kullback-Leibler divergence is between P_{XY} and $P_X P_Y$, the more correlated are X and Y and thus their mutual information is larger.

3.6 The Bayesian inference approach

Bayesian inference stands at the roots of today's most sophisticated inference algorithms and signal processing techniques for matrix reconstruction problems such as compressed sensing and dictionary learning, error correcting codes, artificial intelligence, statistical arbitrage and classification, large scale analysis of economical market or astrophysical data, bio-infomatics and phylogenetics algorithms, decision making in automated systems such as planes, pattern recognition, optimal control theory, decision helping for judges in courtroom or physicians with automated diagnostics. It is even used in philosophy and social sciences. The list could go on.

The strength of Bayesian inference resides in its very definition of being a general method for combining in a mathematical model all the observed data and the a priori information one have about the studied phenomenon or system. Let us formalize the basic principles of Bayesian inference, that are useful in the present context. We thus focus on the sparse linear estimation problem (3.18) but the method presented here is quite general. Many very nice references can be found for more details [7, 24]. In particular, [10] discusses the advantages and weaknesses of Bayesian inference with respect to the frequentist statistical methods.

3.6.1 The method applied to compressed sensing

Again, the problem is estimating $\mathbf{s}$ as accurately as possible from the knowledge of the finite data $\mathbf{y}$ generated from the linear relation (3.18) where the sensing matrix $\mathbf{F}$ is known too. Estimating the matrix as well can be of interest such as in matrix factorization or blind calibration problems [11, 55–60] but is out of the scope of the present thesis. This is done by optimizing some deterministic cost function (3.25) in the convex optimization approach, but in the Bayesian setting, we use a *probabilistic* point of view. To do so we define an intermediate vector $\mathbf{x}$ to represent the signal, with its associated probability distribution $P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ given the observed data and some parameters linked to the prior knowledge about $\mathbf{s}$ (the dependence on $\mathbf{F}$ is implicit). From the simple yet very powerful Bayes formula which form the core of the Bayesian methodology, this *posterior* distribution is obtained as the product of the *prior distribution* $P_0(\mathbf{x}|\boldsymbol{\theta})$ and the so called *likelihood* $P(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$:

$$P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = \frac{P_0(\mathbf{x}|\boldsymbol{\theta})P(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\int d\mathbf{x} P_0(\mathbf{x}|\boldsymbol{\theta})P(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})} = \frac{P_0(\mathbf{x}|\boldsymbol{\theta})P(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{P(\mathbf{y}|\boldsymbol{\theta})} \tag{3.56}$$

The likelihood is the probability of the observed data given that the input would have been $\mathbf{x}$ and the model parameters $\boldsymbol{\theta}$. It is obtained from the generating model knowledge: it enforces the signal estimate to verify the system (3.18), i.e. to give back the actual observations. In the AWGN case, the proper form is naturally given by a product of Gaussian densities: the measures are independent of each other from the random design of the sensing matrix and the authorized fluctuations of the estimated observations $\tilde{\mathbf{y}} := \mathbf{F}\mathbf{x}$ around the actual ones $\mathbf{y}$ are

Gaussian distributed due to the Gaussian nature of the noise:

$$P(\mathbf{y}|\mathbf{x},\Delta) = \prod_{\mu}^{M} \mathcal{N}(y_\mu|(\mathbf{Fx})_\mu,\Delta) \tag{3.57}$$

$$= \frac{1}{(2\pi\Delta)^{M/2}} \exp\left(-\frac{M}{2\Delta}||\mathbf{y}-\mathbf{Fx}||_2^2\right) \tag{3.58}$$

where we remind that $||\mathbf{y}||_p := 1/M\left(\sum_i^M |y_i|^p\right)^{1/p}$ is the rescaled ℓ_p norm. The prior allows to include assumed knowledge about the solution $\mathbf{s}$. From now on, the parameters $\boldsymbol{\theta}$ are considered fixed but we will see in sec. 4.3.8 how these can be learned efficiently in the Bayesian framework if unknown. So if one assumes sparsity about the signal and the fact that each of its entries have been generated independently from the same distribution, a proper factorizable prior would be of the form:

$$P_0(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i}^{N} \left[(1-\rho)\delta(x_i) + \rho\phi(x_i|\tilde{\boldsymbol{\theta}})\right] \tag{3.59}$$

where the parameters $\boldsymbol{\theta} = [\rho, \tilde{\boldsymbol{\theta}}]$ are the probability ρ for a component to be part of the support (sometimes referred as the density of the signal) and the parameters $\tilde{\boldsymbol{\theta}}$ that parametrize the distribution $\phi(\,|\tilde{\boldsymbol{\theta}})$ associated to the support components. This distribution can be shaped as desired. This is in part thanks to the flexibility of the prior that Bayesian inference overcomes convex optimization procedures, if used properly. This distribution is defined independently of the observations: using the data to estimate the prior would be a mistake as the information contained in the likelihood and the prior would be redundant.

Finally $P(\mathbf{y}|\boldsymbol{\theta}) = Z(\mathbf{y},\boldsymbol{\theta})$ is the unknown probability of the data independently of the signal (at fixed parameters $\boldsymbol{\theta}$), which can be interpreted as a partition function. $P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ is called the posterior because it is defined afterwards the data $\mathbf{y}$ has been obtained.

Convex ℓ_1 optimization procedures generally just "know" about sparsity of the solution, whereas way more information about the solution structure can be included in the Bayesian setting through the prior, such as how sparse the signal is (thanks to ρ) or what is the precise distribution of the support components. We could even design a prior enforcing hard constraints, i.e. that strongly correlates the components of $\mathbf{x}$, such as in the superposition codes studied later in this thesis sec. 9 or component-wise priors, assuming that all the signal components could have been generated from different distributions. This is for example studied in [21] but is out of the scope of this thesis, where we always consider the signal components to be i.i.d. This additionnal information (if matching well the true features of the signal) allows inference from less data than convex optimization requires. As we will see, Bayesian inference actually allows asymptotically to reach optimality in two distinct senses: *i*) It allows to solve the inference problem (3.18) from the *lowest possible* sampling rate $\alpha = \rho$ corresponding to as many samples as support components. This will be possible thanks to message-passing sec. 4.3 combined with *spatial coupling* sec. 5.5, a technique intensively

used in this thesis. ii) It can find the solution that has the lowest MSE among all approximate solutions, i.e. the minimum mean square error $MMSE$ estimator. This is possible only if the prior is the true distribution that has generated the random signal $\mathbf{s}$: we call this the *prior matching* or *Nishimori* condition where the second denomination comes from the statistical physics vocabulary. In this case, the estimator is said to be *Bayes optimal*: the reducible error is canceled and it remains only the irreducible and finite size errors, see sec. 3.1.6.

3.6.2 Different estimators for minimizing different risks: Bayesian decision theory

Let us assume that we are able in some way to obtain the true posterior distribution (3.56), which is actually a NP problem (see sec. 3.3). Actually this thesis is mainly about the approximate message-passing algorithm derived in sec. 4.3 which is able to efficiently solve it.

The question now is how can we actually use the posterior to perform inference and estimate $\mathbf{s}$? The answer resides in the *Bayesian decision theory*. Very nice courses on the subject are [7, 61, 62]. From this posterior, three different decisions seem naturals, each minimizing a different risk definition. We already defined the risk (3.13) in sec. 3.1.6 but its definition can be actually extended. We remain in the Bayesian framework considering the data $\mathbf{y}$ fixed as in sec. 3.1.6. In full generality, the risk associated to a *loss* $E(\hat{\mathbf{x}},\mathbf{x})$ (i.e. an error estimate) is its average with respect to the posterior distribution of the signal at fixed data:

$$R(\hat{\mathbf{x}}|\mathbf{y}) := \int d\mathbf{x} P(\mathbf{x}|\mathbf{y}) E(\hat{\mathbf{x}},\mathbf{x}) \tag{3.60}$$

It depends on the estimator $\hat{\mathbf{x}}$ and the data. The auxilliary vector $\mathbf{x}$ for which we have the posterior $P(\mathbf{x}|\mathbf{y})$ (3.56) represents the signal $\mathbf{s}$ that we don't know. The posterior can depend on parameters $\boldsymbol{\theta}$, but we drop this dependency for simplicity. (3.13) is thus (3.60) where the loss $E(\hat{\mathbf{x}},\mathbf{x})$ is taken to be the $MSE(\hat{\mathbf{x}},\mathbf{x})$ (3.12). The associated Bayes risk is (3.14).

A Bayesian decision is just an estimator that minimizes some risk. An important remark is that from (3.14), it is easy to see that if $\hat{\mathbf{x}}^*$ is an estimator minimizing the Bayes risk (3.14) then:

$$\begin{aligned}
\partial_{\hat{\mathbf{x}}} R(\hat{\mathbf{x}})\big|_{\hat{\mathbf{x}}^*} &= 0 && (3.61)\\
&= \left(\partial_{\hat{\mathbf{x}}} \int d\mathbf{y} P(\mathbf{y}) R(\hat{\mathbf{x}}|\mathbf{y})\right)\Big|_{\hat{\mathbf{x}}^*} && (3.62)\\
&= \int d\mathbf{y} P(\mathbf{y}) \partial_{\hat{\mathbf{x}}} R(\hat{\mathbf{x}}|\mathbf{y})\big|_{\hat{\mathbf{x}}^*} && (3.63)\\
\Rightarrow \partial_{\hat{\mathbf{x}}} R(\hat{\mathbf{x}}|\mathbf{y})\big|_{\hat{\mathbf{x}}^*} &= 0 && (3.64)\\
\Rightarrow \hat{\mathbf{x}}^* &= \underset{\hat{\mathbf{x}}}{\operatorname{argmin}}\, R(\hat{\mathbf{x}}) = \underset{\hat{\mathbf{x}}}{\operatorname{argmin}}\, R(\hat{\mathbf{x}}|\mathbf{y}) && (3.65)
\end{aligned}$$

Thus minimizing the risk or the Bayes risk to take a decision by defining an estimator $\hat{\mathbf{x}}^*$ is

actually perfectly equivalent.

The MAP estimator

One could think about taking the mode of the posterior, i.e. the signal that maximizes it. This is referred as the *maximum-a-posteriori MAP* estimator:

$$\hat{\mathbf{x}}_{MAP} = \underset{\mathbf{x}}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{y}) \tag{3.66}$$

The MAP estimator is appropriate in cases where the information resides in the overall state of the full vector, i.e. each individual or subset of components does not bring any information, only the full vector has a meaning. The risk minimized by the MAP estimator is associated to the following loss:

$$E_{MAP}(\hat{\mathbf{x}}, \mathbf{x}) = 1 - \delta_{\hat{\mathbf{x}},\mathbf{x}} \tag{3.67}$$

Indeed:

$$\underset{\hat{\mathbf{x}}}{\operatorname{argmin}} \int d\mathbf{x} P(\mathbf{x}|\mathbf{y})(1 - \delta_{\hat{\mathbf{x}},\mathbf{x}}) \tag{3.68}$$

$$= \underset{\hat{\mathbf{x}}}{\operatorname{argmin}} \left(1 - \int d\mathbf{x} P(\mathbf{x}|\mathbf{y}) \delta_{\hat{\mathbf{x}},\mathbf{x}}\right) \tag{3.69}$$

$$= \underset{\hat{\mathbf{x}}}{\operatorname{argmin}} \left(1 - P(\hat{\mathbf{x}}|\mathbf{y})\right) \tag{3.70}$$

$$= \underset{\hat{\mathbf{x}}}{\operatorname{argmax}}\, P(\hat{\mathbf{x}}|\mathbf{y}) \tag{3.71}$$

we find back the MAP estimator (3.66).

The MARG estimator

A related estimator is the minimal error assignments $MARG$ estimator:

$$\hat{\mathbf{x}}_{MARG} = \left[\underset{x_i}{\operatorname{argmax}}\, P(x_i|\mathbf{y})\right]_i^N \tag{3.72}$$

which is the component-wise MAP estimator, i.e. it the concatenation of the MAP estimates of the marginals $P(x_i|\mathbf{y})$ defined as:

$$P(x_i|\mathbf{y}) := \int d\mathbf{x}_{\setminus i} P(\mathbf{x}|\mathbf{y}) \tag{3.73}$$

We have that $\hat{\mathbf{x}}_{MAP} = \hat{\mathbf{x}}_{MARG}$ if the posterior is factorizable over the estimate components, i.e. $P(\mathbf{x}|\mathbf{y}) = \prod_i^N P(x_i|\mathbf{y})$. The $MARG$ estimator minimizes the risk associated to the number of

incorrectly infered components:

$$E_{MARG}(\hat{\mathbf{x}},\mathbf{x}) = \sum_{i}^{N}(1-\delta_{x_i,\hat{x}_i}) \tag{3.74}$$

This is well suited when the components are i.i.d discrete like in the binary channel models, that are classical noise models [23, 63] in communications.

The MMSE estimator

In the general model (3.18), we are interested in continuous signal and sensing matrix elements and thus the notion of "exactness" of the solution is not really meaningful nor essential. Anyway, as the precision of a computer is finite, we cannot hope to infer exactly a real value as opposed to discrete ones. A better suited loss in this case is the MSE (3.12). The associated estimate is denoted as the minimum mean square error $MMSE$ estimator. The $MSE(\hat{\mathbf{x}},\mathbf{s})$ of the $MMSE$ estimate $\hat{\mathbf{x}}$ can be interpreted as the empirical variance of a Gaussian distribution centered around the solution $\mathbf{s}$ that would have been sampled to generate the i.i.d components of $\hat{\mathbf{x}}$. It is a quite natural loss to use in the continuous framework and even more when the observations were corrupted by an AWGN such as in our case (3.18) because if we are performing inference under the prior matching condition, the $MSE(\hat{\mathbf{x}},\mathbf{s})$ becomes a measure correlated to the variance of the AWGN $\boldsymbol{\xi}$.

What is the expression of the i^{th} component of the $MMSE$ estimator when we observed $\mathbf{y}$? As before, we minimize the risk (3.13) that can be estimated from the knowledge of the posterior distribution. Differentiating it we obtain the estimator $\hat{x}_i(\mathbf{y})$:

$$\partial_{\hat{x}_i}\mathbb{E}_{\mathbf{x}|\mathbf{y}} < (\hat{\mathbf{x}}-\mathbf{x})^2 >= 2/N\,\mathbb{E}_{\mathbf{x}|\mathbf{y}}(\hat{x}_i - x_i) = 0 \tag{3.75}$$

$$\Rightarrow \hat{x}_i = \mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i) = \mathbb{E}_{x_i|\mathbf{y}}(x_i) \tag{3.76}$$

where $\mathbb{E}_{x_i|\mathbf{y}}$ denotes the average with respect to the posterior marginal distribution of x_i given $\mathbf{y}$ (3.73). Thus in order to perfom $MMSE$ estimation, we need the true posterior marginals. If the estimated marginals are equal to the true ones, we say that the estimation is Bayes optimal, i.e. it is the true $MMSE$ and no solution can statistically make a better estimate given the data.

3.6.3 Why is the minimum mean square error estimator the more appropriate ? A physics point of view

Now we can give a more fundamental justification for preferring the $MMSE$ estimator to the MAP or $MARG$ ones. The problems we are trying to solve are noisy, thus there exist an all set of possible solutions to (3.18), each weighted by its posterior probability. In such problems, the posterior can have a very complex shape which details depend on the observations $\mathbf{y}$. For example on Fig. 3.4, we show two different posterior distributions computed from two

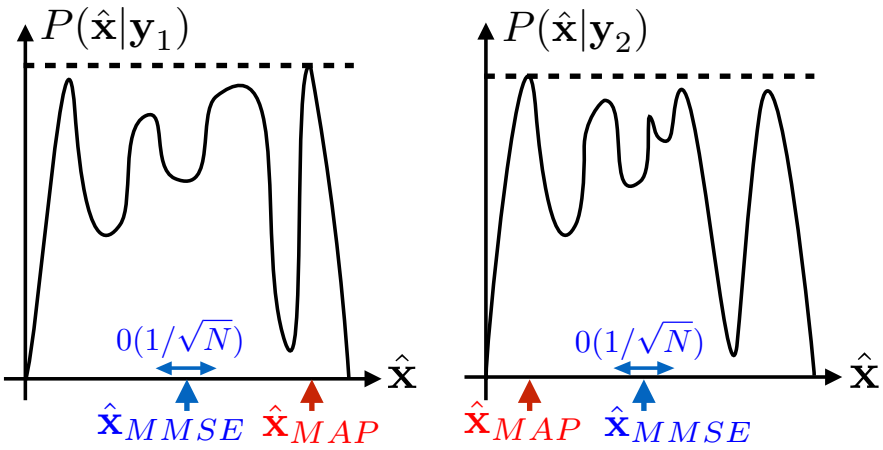

Figure 3.4 – These plots present two complex posterior distributions corresponding to two different observation vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ both related to the same signal $\mathbf{s}$, the difference coming from the noise $\boldsymbol{\xi}$ and measurement matrix $\mathbf{F}$ realizations. Are also represented the minimum mean square error $MMSE$ and maximum-a-posteriori MAP estimators in both cases. The MAP estimator corresponds to the mode of the distribution, meanwhile the $MMSE$ is the typical value, i.e. averaged with respect to the posterior. We observe that due to the fluctuations in the problem realization, small changes in the shape of the posterior can induce large changes in the MAP estimator meanwhile the fluctuations of the $MMSE$ one between different observation realizations are small $\in O(1/\sqrt{N})$.

different measurement vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ obtained from the *same* signal: the differences in the observations come from the noise and measurement matrix realizations. The point is that despite that these two measurements correspond to the same signal, their fluctuations modify the estimated posterior which mode can fluctuate a lot. It can be that there are many local maxima in the posterior with small differences but corresponding to totally different signal estimates and depending on the observations, the mode changes radically. In constrast, the $MMSE$ is robust to such fluctuations of the observations as it an averaged quantity, which thus cancels out the fluctuations in the thermodynamic limit. In statistical physics, we would say that the $MMSE$ estimate is *self-averaging*, like most of the thermodynamical quantities (average energy, average number of particules, total magnetization, etc). It means that its value in finite size problems converges to its asymptotic value as N increases (which is the same as its average value with respect to the disorder, here the noise and matrix realizations) and its relative fluctuations around this mean are $\in O\left(1/\sqrt{N}\right)$. Another way of seeing it is that the MAP estimator is a zero temperature quantity: it does not take into account the entropic contribution, i.e. there is no notion of average over the thermal fluctuations interpreted here as the various weighted solution estimates. Fig. 3.5 is a strange posterior distribution with the mode being a rare event: sampling this distribution, we would obtain the MAP estimate very rarely as opposed to many realizations that would be close to the $MMSE$ estimate as there are so many of them, despite being a bit less probable.

After this dicussion, we understand better why the physics community became interested in

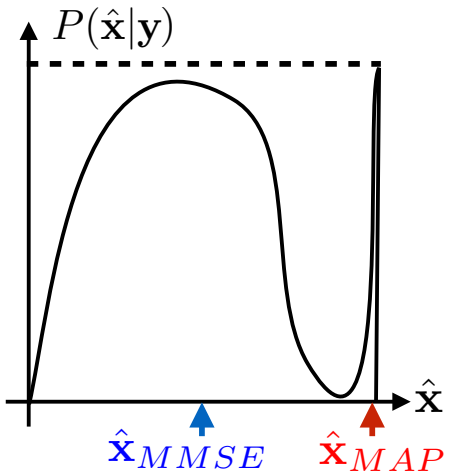

Figure 3.5 – This strange posterior illustrates that without taking the entropic constribution into account by selecting the MAP estimator, we would miss the essential contribution of the solutions weighted by the big bump of the distribution. Instead, the $MMSE$ estimator properly weights each solution and gives the typical estimate.

these topics: through the link with spin glass physics where the distributions are also very rough, like in constraint satisfaction problems as well [1, 24, 64]. Actually, compressed sensing itself can also be seen as a finite temperature constraint satisfaction problem or a densely connected (i.e. with infinite range interactions) spin glass model of continuous spins in an external field, where the interactions defined by the measurement matrix and observations enforce the state of the spins to verify the linear system (3.18) (up to noise, interpreted as the temperature) and the external field would be the prior in the Bayesian setting. The problem of inferring the signal that generated the measurements (or "planted" solution in the physics language) by estimating the $MMSE$ solution is equivalent to sampling from the Boltzmann measure of the appropriate spin glass model given by the Hamiltonian (4.26), whereas the maximum-a-posteriori estimate is given by its ground state (see [34, 35] for a more detailed discussion of the links between compressed sensing and spin glass physics).

Similar mappings can be established for many other computer science, inference and machine learning problems [1, 14, 57, 65, 66] where the typical phenomenology of spin glasses is observed: phase transitions and dynamical slowing down of the reconstruction algorithms near the critical "temperature" (the critical measurement rate in compressed sensing), see sec. 5.1.1. Furthermore, message-passing algorithms such as belief propagation presented in sec. 4.2.1 can be interpreted in terms of the cavity method used on single instances (see sec. 4.2.2), although the cavity method has been originally developed for computing thermodynamical quantities (i.e. averaged over the source of disorder) in spin glasses [24, 67].

So one needs to compute the marginal posterior distributions (3.73) in an efficient manner to perform $MMSE$ estimation. This is in general a very difficult problem, as it is equivalent to compute the normalization $P(\mathbf{y}|\boldsymbol{\theta})$ of the posterior which plays the role of the partition

function (5.2). Hopefully, a sub-field of research in inference is interested in the developement of efficient algorithms specifically designed for this task. Classical methods are based on monte carlo algorithms that directly sample the posterior for estimating it. But this distribution can have a very rough shape with many local minima like in Fig. 3.4 which can block this kind of dynamical algorithms (it must be understood that Fig. 3.4 is just a projection, in reality the distribution is defined over a very high dimensional space). Fortunatly it exists methods that are way more efficient such as the so-called message-passing algorithms that will be explained in great details and fully derived in this thesis.

3.6.4 Solving convex optimization problems with Bayesian inference

It is important to notice that the canonical convex optimization problems can be solved in the Bayesian framework as well. Looking at the LASSO regression (3.26) problem, we see that it is perfectly equivalent to find the mode, i.e. the MAP estimate associated to the posterior distribution $P(\mathbf{x}|\mathbf{y},\Delta,\lambda) \propto P(\mathbf{y}|\mathbf{x},\Delta)P_0(\mathbf{x}|\lambda)$:

$$\hat{\mathbf{x}}_1 = \underset{\mathbf{x}}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{y},\Delta,\lambda) = \underset{\mathbf{x}}{\operatorname{argmax}}\, \exp\left(-\frac{M}{2\Delta}||\mathbf{y}-\mathbf{F}\mathbf{x}||_2^2 - \lambda' N||\mathbf{x}||_1\right) \tag{3.77}$$

if we put $\Delta = M/2$ and $\lambda' = \lambda/N$, where we have used the likelihood (3.58) and the factorizable prior is $P_0(\mathbf{x}) \propto \exp\left(-\lambda' N||\mathbf{x}||_1\right)$. This prior is refered as the double exponential or *Laplace* prior. The same is true for the ridge regression (3.27) using an i.i.d Gaussian prior $P_0(\mathbf{x}) = \prod_i^N \mathcal{N}(x_i|0,1/(2\lambda))$. In the case of this Gaussian prior, the posterior mode is also the posterior mean of $P(\mathbf{x}|\mathbf{y},\Delta,\lambda)$, but it is not the case with the Laplace prior. In this case the posterior mode solves the LASSO regression and leads to a sparse solution whereas the posterior mean of $P(\mathbf{x}|\mathbf{y},\Delta,\lambda)$ is not sparse [8].

3.7 Error correction over the additive white Gaussian noise channel

We now present very briefly coding theory and error correcting codes. A very complete and comprehensive book on the subject is [23]. Error correcting codes are of interest in the statistical physics community for a long time as the decoding problem can be interpreted as computing magnetizations of disordered spins systems [65, 68–73]. Very nice and modern references emphasizing the numerous links between coding theory and diluted spin-glass models can be found in [63, 66].

The aim of coding theory is to find the "best way" to encode a message such that once sent to some receiver through a noisy channel, it can be decoded, i.e. recovered despite of the errors induced by the channel. What does the best way means? Three main considerations must be taken into account for a coding/decoding scheme: i) its robustness to the noise, ii) its symbolic cost and iii) the existence of an efficient decoder. A good scheme thus requires to be robust, in the sense that in regimes where the scheme should work (this notion of "favorable regime" will be discussed in details in sec. 5.1), its performances should decrease smoothly

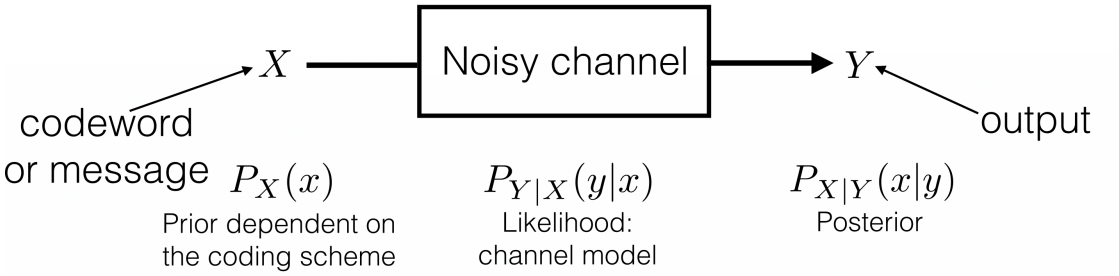

Figure 3.6 – Graphical representation of a generic noisy channel in the probabilistic framework: the codeword or input x is designed through its prior P_X, the channel by the likelihood $P_{Y|X}$ of the output y and the aim is to estimate the posterior distribution $P_{X|Y}$ of the input knowing the output in order to perform decoding through estimators.

as the noise influence increases: a scheme which performances are highly dependent on the precise noise level or which correction does not improve as the noise lowers is not reliable. Then must be considered its symbolic cost, i.e. what is the true quantity of information that one can send reliably thanks to this scheme each time a symbol is sent through the noisy channel. This is important in practical situations as each sent symbol has a cost in energy and time. If one could send symbols without any cost, the error correction would be easy: just send many times the same message (this is called a repetition code), then the receiver just makes a majority choice by selecting for each symbol of the message the one that has been received the most or by averaging over all the received noisy realizations of the symbols in the continuous setting. Finally, even these two conditions are optimized, it is useless if the receiver has no way to decode the message, i.e. it must exist a decoder which performs quickly, is itself robust to noise and has good finite size properties, see sec. 3.1.6.

3.7.1 The power constrained additive white Gaussian noise channel and its capacity

The communication channel model we are interested in is the i.i.d AWGN channel with zero mean and variance Δ, a classical model in communication also extensively studied by the physics community, as in [74, 75]. A generic noisy channel model in the probabilistic framework is represented on Fig. 3.6. It is modeled in the Bayesian framework through the likelihood of an output $\mathbf{y}$ given the input $\tilde{\mathbf{y}}$ of the channel, which is in the i.i.d AWGN case:

$$P(\mathbf{y}|\tilde{\mathbf{y}},\Delta) = \prod_{\mu}^{M} \mathcal{N}(y_\mu|\tilde{y}_\mu,\Delta) \tag{3.78}$$

A channel which likelihood is factorizable due to the independence assumption like here is referred as a memoryless channel [23]: all output symbols (the components of $\mathbf{y}$) are corrupted independently one of the other. The input is called the *codeword*, and is the encoded version of the original message $\mathbf{s}$ we want to send (the codeword can be the message itself, as in chap. 10). A natural question is wether there exist a maximum rate of information one can send reliably

through this commmunication channel, and this independently of the coding/decoding scheme. This notion called the *capacity* C has been formalized by Shannon in his celebrated paper [54] which started the field of communication and information theory. The noisy-channel coding theorem states that for any $\epsilon > 0$ and for any transmission rate $R < C$, there exist an encoding/decoding scheme transmitting data at rate R which error probability is less than ϵ, for a sufficiently large block length, the size of the codeword. Also, for any rate $R > C$, the probability of error at the receiver goes to one as the block length goes to infinity, for any coding/decoding scheme.

As discussed in sec. 3.5.2, the mutual information between the noisy ouput $\mathbf{y}$ and the message $\mathbf{s}$ represents the information gained about $\mathbf{s}$ when we observed $\mathbf{y}$. As long as it is positive, it means that some information about the message is accesible through the observation of the channel output: a good coding/decoding scheme increases this mutual information maintaining the existence of an associated efficient decoder, able to maximally exploit it. A capacity achieving scheme allows to communicate asymptotically until the capacity of the channel.

Of course, we always consider that the coding scheme $\tilde{\mathbf{y}} = f(\mathbf{s})$ is a bijection and thus finding back the codeword $\tilde{\mathbf{y}}$ is equivalent to decode $\mathbf{s}$ and vice-versa. The capacity is thus naturally defined as the maximum mutual information of the couple $(\mathbf{s}, \mathbf{y})$ or equivalently of the couple $(\tilde{\mathbf{y}}, \mathbf{y})$ as $\tilde{\mathbf{y}}$ is a deterministic function of $\mathbf{s}$. Maximum over what? As the likelihood is an inherent characteristic of the channel, the only degree of freedom is the codeword design $P_0(\tilde{\mathbf{y}})$ that directly follows from the message design $P_0(\mathbf{s})$ and coding scheme f. Thus the capacity of a communication channel is:

$$C := \max_{P_{\tilde{\mathbf{y}}}} I(\tilde{\mathbf{y}}, \mathbf{y}) = \max_{\{P_{\mathbf{s}}, f\}} I(\tilde{\mathbf{y}}(\mathbf{s}, f), \mathbf{y}) \tag{3.79}$$

The second equality underlines that it is perfectly equivalent to consider directly the design of the codeword or the design of the message and the coding scheme, which is the usual way. But in chapter chap. 10 the codeword is directly the signal, so the first equality is more appropriate. An important remark is that if one could send codewords with components of arbitrary amplitude through the channel, error correction would be useless as the relative noise (relative to the codeword amplitude) could be set to arbitrary small values. We thus always consider the codeword to be power constrained, i.e. we fix its power:

$$||\tilde{\mathbf{y}}||_2^2 = \int d\tilde{y} P(\tilde{y}) \tilde{y}^2 = P \tag{3.80}$$

In this way it can be compared to the noise variance to know its relative importance. The first equality in the power definition comes from the fact that the $\tilde{\mathbf{y}}$ are i.i.d. from the i.i.d assumption of the matrix elements in (3.18). A larger power requires more energy to input in the channel. The only relative parameter of interest is thus the so-called *signal to noise* ratio defined as $\text{snr} := P/\Delta$.

Computation of the capacity of the i.i.d additive white Gaussian noise power constrained channel

Let us compute the capacity of the power constrained AWGN channel. To do so, starting from the mutual information, we define a Lagragian to enforce the distribution of the codeword to be normalized and to have fixed power. We place ourselves in the scalar case, the codeword having i.i.d components, we just need to obtain the distribution of one component:

$$\mathscr{L} = I(\tilde{y}, y) + \lambda\left(\int d\tilde{y} P(\tilde{y})\tilde{y}^2 - P\right) + \gamma\left(\int d\tilde{y} P(\tilde{y}) - 1\right) \tag{3.81}$$

$$= \int d\tilde{y} dy P(\tilde{y}) P(y|\tilde{y}) \log_2\left(\frac{P(y|\tilde{y})}{P(y)}\right) + \lambda\left(\int d\tilde{y} P(\tilde{y})\tilde{y}^2 - P\right) + \gamma\left(\int d\tilde{y} P(\tilde{y}) - 1\right) \tag{3.82}$$

where we used the form (3.51) for the mutual information. Now we perform the functionnal derivative of the Lagrangian to find its optimum with respect to the codeword/input distribution, that we are looking for:

$$\frac{\delta\mathscr{L}}{\delta P(\tilde{y}^*)} = \int d\tilde{y}\delta(\tilde{y} - \tilde{y}^*)\left(\int dy P(y|\tilde{y}) \log_2\left(\frac{P(y|\tilde{y})}{P(y)}\right) + \lambda\tilde{y}^2 + \gamma\right)$$
$$- \int d\tilde{y} dy P(\tilde{y}) P(y|\tilde{y}) \frac{P(y|\tilde{y}^*)}{P(y)} = 0 \;\forall\; \tilde{y}^* \tag{3.83}$$

where the last term has been obtained using:

$$\frac{\delta P(y)}{\delta P(\tilde{y}^*)} = \frac{\delta}{\delta P(\tilde{y}^*)} \int d\tilde{y} P(\tilde{y}) P(y|\tilde{y}) = \int d\tilde{y}\delta(\tilde{y} - \tilde{y}^*) P(y|\tilde{y}) = P(y|\tilde{y}^*) \tag{3.84}$$

Now we notice that this last term of (3.83) is equal to -1. Plugging the fact that the likelihood of the channel ouput is Gaussian in it and after integrating over $\tilde{y}$, (3.83) simplifies to:

$$\int dy \mathscr{N}(y|\tilde{y}^*, \Delta) \log_2\left(P(y)\right) = \lambda(\tilde{y}^*)^2 + \tilde{\gamma} \tag{3.85}$$

where we used $\int dy \mathscr{N}(y|\tilde{y}^*, \Delta) \log_2\left(\mathscr{N}(y|\tilde{y}^*, \Delta)\right) = -1/2(1 + \log_2(2\pi\Delta))$ and we have put all the constants in $\tilde{\gamma}$. Using a Taylor expansion for $\log_2(P(y)) = a_0 + a_1 y + a_2 y^2 + a_3 y^3 + ...$, this last equality can be obtained only if the expansion is such that $a_i = 0 \;\forall\; i \neq \{0, 2\}$, thus the ouput distribution $P(y)$ is Gaussian with 0 mean and a unknown variance σ^2 to find. We re-write it using the likelihod:

$$P(y) = \mathscr{N}(y|0, \sigma^2) = \int d\tilde{y} P(y|\tilde{y}) P(\tilde{y}) = \int d\tilde{y} \mathscr{N}(y|\tilde{y}, \Delta) P(\tilde{y}) \tag{3.86}$$

The last equality can only be fulfilled if the codeword distribution $P(\tilde{y}) = \mathscr{N}(\tilde{y}|0, P)$ is a centered Gaussian, its variance being fixed by the power contraint (3.80). It implies that the channel ouput variance is the sum of the power and noise variance $\sigma^2 = P + \Delta$ as the noise and inputs are independent. Now we know the best codeword distribution, we can compute

the capacity from (3.79) and and (3.48):

$$C = -H(Y|\tilde{Y}) + H(Y) \tag{3.87}$$

$$= \int d\tilde{y} dy P(\tilde{y}) P(y|\tilde{y}) \log_2\left(P(y|\tilde{y})\right) - \int dy P(y) \log_2\left(P(y)\right) \tag{3.88}$$

$$= \int d\tilde{y} dy \mathcal{N}(\tilde{y}|0,P) \mathcal{N}(y|\tilde{y},\Delta) \log_2\left(\mathcal{N}(y|\tilde{y},\Delta)\right)$$

$$- \int dy \mathcal{N}(y|0,P+\Delta) \log_2\left(\mathcal{N}(y|0,P+\Delta)\right) \tag{3.89}$$

$$= -\frac{1}{2}\left(1+\log_2(2\pi\Delta)\right) + \frac{1}{2}\left(1+\log_2(2\pi(P+\Delta))\right) \tag{3.90}$$

$$= \frac{1}{2}\log_2(1+\text{snr}) \tag{3.91}$$

This is the maximum quantity of information in bits one can hope to transmit reliably per symbol sent through the i.i.d AWGN channel, and it increases with the snr as it should. One goal in communication theory over the AWGN channel is thus to find an encoder f for the message $\mathbf{s}$ such that the codeword is Gaussian distributed in order to get as close as possible to the capacity. Also one must derive an associated decoder to find back the message $\mathbf{s}$ from the observation of the noisy observation $\mathbf{y}$ of the codeword. Such a strategy, the sparse superposition codes and the associated message-passing decoder will be studied in this thesis in great details, see sec. 9.

3.7.2 Linear coding and the decoding problem

Now we have presented the AWGN channel and quantified the maximum rate for reliable communication on this channel, the question is how to reach it? Many coding strategies are possible, but of particular interest in this thesis is linear coding, $\tilde{\mathbf{y}} = f(\mathbf{s}) = \mathbf{F}\mathbf{s}$. The codeword is thus a linear combination of the basis $\mathbf{F}$ elements, which vector of coefficients is the message. This scheme is of interest for diverse reasons. First, the encoding procedure is trivial, it requires only a matrix multiplication which is of complexity $O(N^2)$ in the general case, but using structured operators, it can be reduced to $O(L\log(N))$. Second, this coding strategy has good *minimal distance* d. This notion is fundamental and allows for a geometrical interpretation of the error correction problem. First we define a code $\mathscr{C}$ (also referred as a codebook) as the ensemble of allowed codewords by the coding scheme. The minimal distance of a code is the minimal distance between two codewords of the code. The distance is expressed in Hamming distance in the discrete case, i.e. the number of different components between codewords but in the present continuous case, an appropriate distance is the ℓ_2 squared norm between two codewords:

$$d := \min_{\tilde{\mathbf{y}},\tilde{\mathbf{y}}' \in \mathscr{C}} ||\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'||_2^2 \tag{3.92}$$

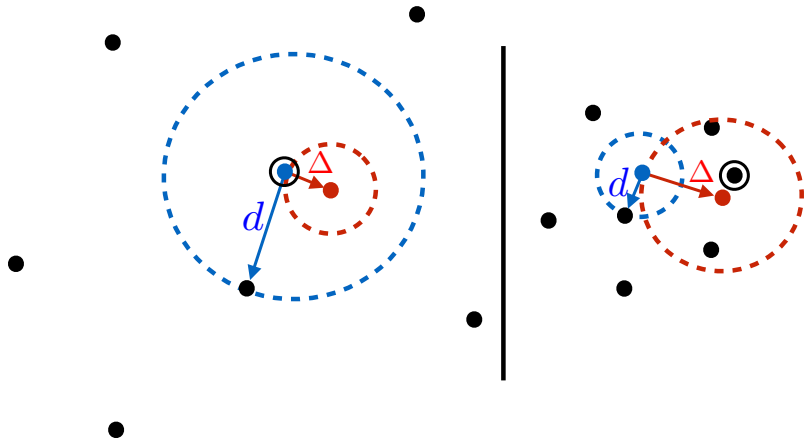

Figure 3.7 – Two different codes are represented, the codewords being the black dots. The code on the left is reliable as the minimal distance d between two codewords is larger than (two times) the noise variance which is the typical distance between the vector that outputs the AWGN channel (the red dot) and the actual sent codeword, the blue one. Here, the decoder that will output the closest neighboring codeword of the channel output selects the encircled codeword which is the transmitted one. On the opposite, the right code is not reliable, as there are many codewords closest to the channel output than the transmitted codeword as $d < \Delta$. The decoder will output the encircled codeword which is a mistake.

Basically, a decoder will output the closest neighboring codeword of the noisy channel output $\mathbf{y}$ in the codebook:

$$\hat{\mathbf{y}} = \underset{\tilde{\mathbf{y}} \in \mathscr{C}}{\operatorname{argmin}} \, ||\tilde{\mathbf{y}} - \mathbf{y}||_2^2 \tag{3.93}$$

Thus an error in decoding is highly probable if d is small compared to the noise variance as the noisy channel will typically ouput a vector at a distance $\in O(\Delta)$ of the transmitted codeword and thus if d is of the same order or smaller, there is no chance to distinguish between the good codeword and its closest neighbors, as represented on the right part of Fig. 3.7. In opposite, if $d > \Delta$ as on the left part of the figure, the code is reliable because the closest neighbor of the channel output is the transmitted codeword with high probability. The distance can be related to the snr noticing that:

$$d \leq ||\tilde{\mathbf{y}}||_2^2 + ||\tilde{\mathbf{y}}'||_2^2 \leq 2P \tag{3.94}$$

$$\Rightarrow d' := \frac{d}{\Delta} \leq 2 \, \text{snr} \tag{3.95}$$

where d' is a rescaled distance by the noise variance. Thus if the snr is too small, there is no way to have a minimal distance large enough to avoid wrong decoding. The quality of the error correction thus depends also strongly on the performances of the decoder and its ability to distinguish between codewords with small distance between them (but larger than Δ). Hopefully there exist very efficient algorithms such as the approximate message-passing

algorithm (see sec. 4.3) which relies on the linearity of the constraints, another reason for choosing linear coding.

A last remark follows from the geometrical understanding given by the Fig. 3.7 and answers the question: *Why not to directly send the message? Why is encoding necessary?* The coding strategy is here to project the messages in a higher dimensional space, such that the distances between the codewords bijectively associated to the messages are larger than the initial distances between messages, and in this way, the decoding of the codewords is more robust to the noise influence.

4 Mean field theory, graphical models and message-passing algorithms

This chapter is devoted to the mean field theory and message-passing algorithms for graphical models, the central algorithmic tools used in this thesis.

We start by introducing the mean field theory and the variationnal method used to compute approximations of the free energy. We will see that the free energy allows to recast inference problems as optimization ones. In addition, algorithms can be derived as fixed point equations associated with these approximated free energies. We will see how this kind of approximation is naturally justified when finite statistics is known about a system thanks to the maximum entropy criterion. We will apply the methodology to compressed sensing after having defined the Hamiltonian of the problem, and we will derive the mean field algorithm for compressed sensing. We then introduce the notion of factor graphs, a nice tool to represent complex statistical dependencies among variables and very useful to understand how message-passing works.

Then we will push further the variational method using a more advanced approximation that takes into account dependencies among variables, namely the tree graph or Bethe approximation. We will start by presenting the belief propagation algorithm, a very powerful inference tool for solving the marginalization problem, among others. We will see how the belief propagation equations can be derived when the probability measure to sample is assumed to be of the Bethe form that will be presented. We will then show that the Bethe free energy and belief propagation fixed points are the same. The Bethe free energy will be written in terms of the quantities computed iteratively by belief propagation.

We will realize that this algorithm cannot be straightforwardly applied to problems defined over continuous variables on dense graphs, and we will thus derive an appropriate algorithm for this case, the main tool of this thesis: the approximate message-passing algorithm. The derivation will be performed in two different ways, both starting from the belief propagation equations. We will present in a simple fashion how this algorithm works and give the building blocks necessary to construct an approximate message-passing algorithm for a given problem. Finally we will derive the asymptotic limit of the Bethe free energy on dense graphs with linear

constraints, which fixed points are the ones of the approximate message-passing and which can be expressed in terms of the quantities computed by the algorithm. This will be useful in order to derive learning equations for unknow parameters in the problem through the expectation maximization procedure. In this method, the paramaters to learn are updated in the direction that optimize this free energy, on in general, the cost function of the problem.

4.1 Bayesian inference as an optimization problem and graphical models

We have presented in sec. 3.5 the notion allowing for quantification of the infomation carried by a probability distribution or equivalently its uncertainty, the entropy, and defined the equivalent of a distance between distributions. Let us see how we can use these tools to perform inference by approximating the true posterior distribution (3.56) which is most of the time very hard to compute exactly but yet required to perform minimum mean square error estimation, see sec. 3.6.2: it selects among all the possible θ values the most probable one.

4.1.1 The variational method and Gibbs free energy

When the posterior $P(\mathbf{x}|\boldsymbol{\theta})$ (or any other distribution) is too complex to compute exactly, one need to approximate it in some way. To do so, we define an approximated distribution as $Q(\mathbf{x}|\boldsymbol{\theta}_Q)$ that can depend on some parameters $\boldsymbol{\theta}_Q$. Now the question is, how to choose them? We can use the natural idea of minimizing the "distance" between our approximated distribution and the true posterior: we will thus optimize the Kullback-Leibler divergence (3.49) between the two. We want to compute $KL(Q||P)$. Why not $KL(P||Q)$ instead, as the Kullback-Leibler divergence is not symmetric? Despite we know the formal expression of the posterior, we could not compute the required averages with respect to it (or the variationnal method would be useless) as it is equivalent to compute the partition function. In opposite, we can compute averages with respect to Q if it is simple enough. It is actually chosen in this purpose. Without loss of generality, we assume a Boltzmann form for the posterior $P(\mathbf{x}|\boldsymbol{\theta}) = \exp\left(-E(\mathbf{x}|\boldsymbol{\theta})\right)/Z(\boldsymbol{\theta})$. Forgetting about the log(2) basis in (3.49) as it does not change the fixed points of the free energy, we obtain:

$$KL(Q||P) = \int d\mathbf{x} Q(\mathbf{x}|\boldsymbol{\theta}_Q) \log\left(\frac{Q(\mathbf{x}|\boldsymbol{\theta}_Q)}{P(\mathbf{x}|\boldsymbol{\theta})}\right) \tag{4.1}$$

$$= \int d\mathbf{x} Q(\mathbf{x}|\boldsymbol{\theta}_Q) E(\mathbf{x}|\boldsymbol{\theta}) + \int d\mathbf{x} Q(\mathbf{x}|\boldsymbol{\theta}_Q) \log\big(Q(\mathbf{x}|\boldsymbol{\theta}_Q)\big) + \log\left(Z(\boldsymbol{\theta})\right) \tag{4.2}$$

$$= \mathbb{E}_{Q|\boldsymbol{\theta}_Q}\left(E(\mathbf{x}|\boldsymbol{\theta})\right) - H(Q|\boldsymbol{\theta}_Q) - F(\boldsymbol{\theta}) \tag{4.3}$$

where we recognized the entropy $H(Q|\boldsymbol{\theta}_Q)$ (3.34) and the *Helmotz free energy* (or just free energy) at fixed parameters $\boldsymbol{\theta}$, the true potential function of the problem:

$$F(\boldsymbol{\theta}) := -\log(Z(\boldsymbol{\theta})) \tag{4.4}$$

$$= \mathbb{E}_{P|\boldsymbol{\theta}}(E(\mathbf{x}|\boldsymbol{\theta})) - H(P|\boldsymbol{\theta}) \tag{4.5}$$

The second equality is obtained by writing the entropy (3.34) of the posterior P in its Boltzmann form. This free energy is not computable as its knowledge is equivalent to the computation of the true partition function $Z(\boldsymbol{\theta})$, or equivalently of the posterior. But this form suggests the definition of a *variational free energy* or *Gibbs free energy* associated to the approximate distribution Q:

$$F_Q(\boldsymbol{\theta}_Q) := \mathbb{E}_{Q|\boldsymbol{\theta}_Q}(E(\mathbf{x}|\boldsymbol{\theta})) - H(Q|\boldsymbol{\theta}_Q) \tag{4.6}$$

From this and (4.3) we obtain the following important equality that stands at the roots of the variational method:

$$F_Q(\boldsymbol{\theta}_Q) = F(\boldsymbol{\theta}) + KL(Q||P) \tag{4.7}$$

and as $KL(Q||P) \geq 0$ with equality only if $Q = P$, we have that $F_Q(\boldsymbol{\theta}_Q) \geq F(\boldsymbol{\theta})$. This validates a posteriori that the best parameters for Q are given by those minimizing $KL(Q||P)$ as it correponds to the ones that minimize the variational free energy, lower bounded by the true one. The advantage with the variational formalism is that the Gibbs free energy (4.6) can be quite easy to compute for an appropriate Q and the free parameters $\boldsymbol{\theta}_Q$ optimal values are computed by optimizing it with respect to them.

4.1.2 The mean field approximation

The most natural approximation for Q in the variational method is the so-called *mean field* approximation, where one assumes that Q is factorizable over subsets of variables $\{\mathbf{x}^a := [x_1^a, \ldots, x_{n_a}^a]\}_a^G$ that can overlap, where n_a is the number of variables in the subset a:

$$Q(\mathbf{x}|\boldsymbol{\theta}_Q) = \frac{1}{Z(\boldsymbol{\theta}_Q)} \prod_a^G \psi_a(\mathbf{x}_a|\boldsymbol{\theta}_Q) \tag{4.8}$$

$$= \frac{1}{Z(\boldsymbol{\theta}_Q)} \exp\left(-\sum_a^G E_a(\mathbf{x}_a|\boldsymbol{\theta}_Q)\right) \tag{4.9}$$

$$= \frac{1}{Z(\boldsymbol{\theta}_Q)} \exp\left(-E(\mathbf{x}|\boldsymbol{\theta}_Q)\right) \tag{4.10}$$

where $\psi_a(\mathbf{x}_a|\boldsymbol{\theta}_Q)$ is some function of the subset $\mathbf{x}_a$. We call it the *compatibility function*, *constraint* or *factor a*. The second form is the associated Boltzmann form, where we define

the "energy" assocated to the compatibility function a:

$$E_a\left(\mathbf{x}_a|\boldsymbol{\theta}_Q\right) := -\log\left(\psi_a(\mathbf{x}_a|\boldsymbol{\theta}_Q)\right) \tag{4.11}$$

$$\Rightarrow E\left(\mathbf{x}|\boldsymbol{\theta}_Q\right) = \sum_a^G E_a\left(\mathbf{x}_a|\boldsymbol{\theta}_Q\right) \tag{4.12}$$

$E\left(\mathbf{x}|\boldsymbol{\theta}_Q\right)$ is the total energy of the system i.e. the Hamiltonian, referred as the *cost function* in statistical inference and computer science. The easiest mean field approximation to deal with, sometimes referred as the *naive mean field approximation*, corresponds to consider the subsets as being the individual variables, so to write Q as a fully factorizable distribution over the signal components, considered as independent:

$$Q(\mathbf{x}|\boldsymbol{h}) = \prod_i^N Q_i(x_i|h_i) \tag{4.13}$$

where Q_i is the approximate marginal distribution of x_i (Q_i can be trivially computed by normalizing any factor ψ_i, in this way Q is already normalized). The denomination of mean field approximation comes from the interpretation of the parameters $\{h_i\}_i^N$ as local fields felt by the variables that summarize the interactions with the other ones.

4.1.3 Justification of mean field approximations by maximum entropy criterion

Assume that you have some partial knowledge about some complex system made of the interacting variables $\mathbf{x}$, such as the first and second order statistics of $\mathbf{x}$. What is the best mean field approximation Q you can do of the true unknown distribution P (here even its formal expression can be unknown)? A possible answer resides in the *maximum entropy criterion*. It is a kind of formalization of the Occam's razor: if one have some knowledge about a system, he should use a model that is in agreement with it, but does not assume anything additional. So in a sense, one should use the minimal model that fits the assumptions or the knowledge about the system. As we speak about statistical models represented by distributions, the natural object to quantify how much we constrain a distribution is through its entropy (3.34).

As we will derive, the Ising model is the mean field approximation of maximum entropy when only the first and second order statistics of the variables $\{x_i\}_i^N$ are known, at least empirically. To see that, we use the method of Lagrange multipliers and define the Lagrangian of a distribution Q starting from its entropy and defining Lagrange multipliers $\{h_i\}_i^N$ that fix its marginals $\{m_i := \int dx_i x_i Q_i(x_i)\}_i^N$ and $\{J_{ij}\}_{ij}^{N,N}$ for the second moments

$\{\Sigma_{ij} := \int dx_i dx_j x_i x_j Q_{ij}(x_i, x_j)\}_{i,j}^{N,N}$ in addition of Γ for its normalization:

$$\mathscr{L}(Q, \mathbf{h}, \mathbf{J}, \Gamma) = H(Q) + \sum_i^N h_i \left(\int dx_i x_i Q_i(x_i) - m_i \right) + \sum_{i,j}^{N,N} J_{ij} \left(\int dx_i dx_j x_i x_j Q_{ij}(x_i, x_j) - \Sigma_{ij} \right) + \Gamma \left(\int d\mathbf{x} Q(\mathbf{x}) - 1 \right) \tag{4.14}$$

Now to find the extremum of this object, we weakly perturbe the distribution $Q \to Q + \delta Q$ and compute the new Lagrangian which is:

$$\mathscr{L}(Q + \delta Q, \mathbf{h}, \mathbf{J}, \Gamma) = \mathscr{L}(Q, \mathbf{h}, \mathbf{J}, \Gamma) \tag{4.15}$$

$$+ \int d\mathbf{x} \delta Q(\mathbf{x}) \left[-\log(Q(\mathbf{x})) - 1 + \sum_i^N h_i x_i + \sum_{i,j}^{N,N} J_{ij} x_i x_j + \Gamma \right] + O(\delta Q^2) \tag{4.16}$$

At the maximum, the first order must cancel out for any $\mathbf{x}$, thus the integrand must always be zero, giving the shape of the distribution Q:

$$0 = -\log(Q(\mathbf{x})) - 1 + \sum_i^N h_i x_i + \sum_{i,j}^{N,N} J_{ij} x_i x_j + \Gamma \tag{4.17}$$

$$\Rightarrow Q(\mathbf{x}|\mathbf{h}, \mathbf{J}) = \frac{1}{Z_Q(\mathbf{h}, \mathbf{J})} \exp\left(\sum_i^N h_i x_i + \sum_{i,j}^{N,N} J_{ij} x_i x_j \right) \tag{4.18}$$

$$= \frac{1}{Z_Q(\mathbf{h}, \mathbf{J})} \prod_i^N \psi_i(x_i|h_i) \prod_{i,j}^{N,N} \psi_{ij}(x_i, x_j|J_{ij}) \tag{4.19}$$

where we have put all the $\mathbf{x}$ independent terms into the normalization constant $Z_Q(\mathbf{h}, \mathbf{J})$. We find back a mean field approximation of the form (4.8), and thus understand now that it corresponds to the minimal model with fixed finite statistics. From this we interpret the lagrange multipliers $\{h_i\}_i^N$ as external fields and $\{J_{ij}\}_{i,j}^{N,N}$ as two points interactions between the variables. Taking into account higher order statistics would lead to more complex models, but in most of the practical situations, higher than second order statistics are difficult to extract from finite data because the number of samples required increases quickly with the order of the moment we want to compute. This follows from the fact that the larger the moment, the larger the amplitude of the fluctuations of its empirical estimate around its true value.

A remark is that interpreting the constraints we want to enforce for the distribution Q through (4.14) as the average energy part in the variational free energy (4.6), the Lagrangian could be interpreted as a negative variational free energy. But the method is different in the sense that when we associate a Gibbs free energy F_Q to a distribution, we *assume its form* that can depend on parameters and the optimization of F_Q gives the best parameters that one should use in conjunction with this particular form of distribution. In the method of the maximum entropy, one *assumes constraints* that must verify the distribution, but *not its form*, and the optimization of the Lagrangian gives the form of the distribution one should use. Then, in

order to find the best parameters introduced during the derivation to enforce the contraints, the Lagrange multipliers, one can a posteriori use the variational method. The two methods are thus complementary in a sense.

4.1.4 The maximum entropy criterion finds the most probable model

Let us give a more precise sense to the maximum entropy principle when used to find the value of a parameter of some variational distribution depending on it. Assume you have access to some data $\mathbf{y}$ that you assume generated by $P(\mathbf{y}|\theta)$ parametrized by a parameter θ. The maximum entropy criterion states that its "best" value θ^* is the one maximizing the entropy of its posterior $P(\theta|\mathbf{y}) \propto P_0(\theta)P(\mathbf{y}|\theta)$:

$$\theta^* = \underset{\theta}{\text{argmax}} \left[-\int d\mathbf{y} P(\theta|\mathbf{y}) \log_2 \left(P(\theta|\mathbf{y}) \right) \right] \tag{4.20}$$

$$\Rightarrow 0 = -\int d\mathbf{y} \frac{\partial P(\theta|\mathbf{y})}{\partial \theta} \left(\log_2 \left(P(\theta|\mathbf{y}) \right) + 1 \right) \tag{4.21}$$

$$\Rightarrow 0 = \frac{\partial P(\theta|\mathbf{y})}{\partial \theta} \tag{4.22}$$

The last equality shows that the maximum entropy criterion for choosing θ is thus equivalent to take the maximum of the posterior of θ, or just its likelihood $P(\mathbf{y}|\theta)$ if no prior is assumed. The maximum entropy criterion is thus perfectly equivalent to the maximum-a-posteriori MAP principle, discussed in sec. 3.6.2.

4.1.5 Factor graphs

Let us see now how we can define graphical representations of complex functions such as posterior distributions in the Bayesian framework. The appropriate tool are the so called *factor graphs*. In full generality, they are used to represent the dependency structure among variables encoded through a factorizable function $Q(\mathbf{x})$, that can depends on many parameters, so of the form (4.8) (the partition function Z can be forgotten for a generic function, but must be present if we want Q to be a probability distribution).

A factor graph is a bipartite graph $G = \{\mathcal{V}, \mathcal{F}, \mathcal{E}\}$ where en edge $e \in \mathcal{E}$ is present between a variable node $v \in \mathcal{V}$ and a factor node $f \in \mathcal{F}$ only if the factor $\psi_f(\mathbf{x}_{f\setminus v}, x_v)$ present in the factorized form of Q depends on x_v. The circle nodes are associated to the variables while squares represent the factors. For example, if we want to associate a factor graph to the posterior distribution of the compressed sensing problem, we write it in its factorized form. Using (3.56) combined with (3.58) and (3.59), we see that the posterior decomposes as a product of functions over the single components due to the prior part and over the ensemble of components due to the likelihood. The associated factor graph is given by Fig. 4.1, and the factorizable structure of the posterior becomes clear. On the graph are representd objects that will be defined in sec. 4.2.1, namely the cavity messages.

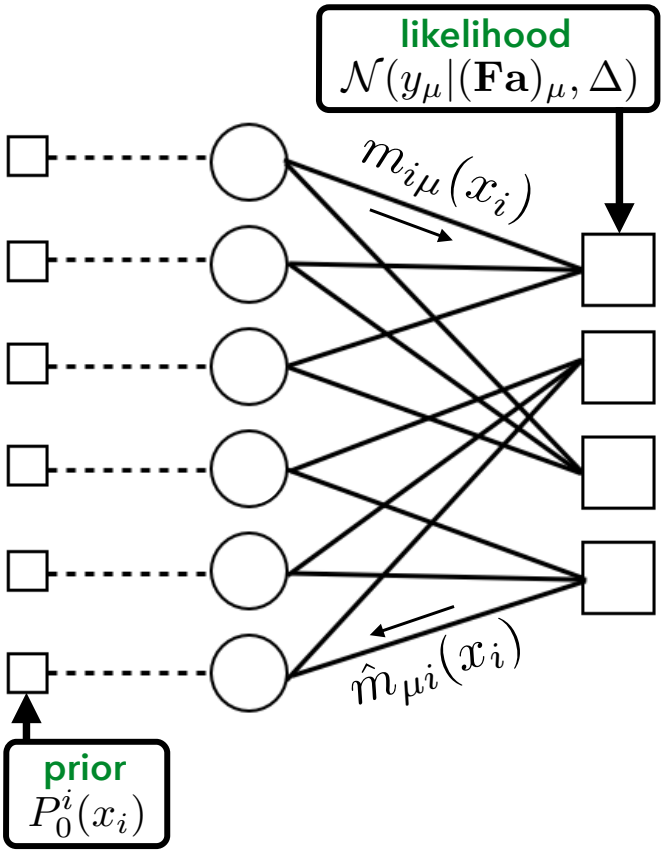

Figure 4.1 – Factor graph associated with a linear estimation problem under i.i.d AWGN corruption. The squares are the factors or constraints. Here they are associated to the prior and likelihood terms. The factors are connected to the variables (circle nodes) they depend on in their functionnal representation. The node-to-factor $m_{i\mu}(x_i)$ and factor-to-node $\hat{m}_{\mu i}(x_i)$ cavity messages are represented. They should stand on the same edge as they share same indices but we put them on different ones for sake of readibility.

4.1.6 The Hamiltonian of linear estimation problems and compressed sensing

In the Bayesian framework, we can associate an Hamiltonian to the linear estimation problem (and thus to compressed sensing). We start by rewriting the posterior distribution in the Boltzmann form:

$$P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp\left(-E(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})\right) \tag{4.23}$$

$$\Rightarrow E(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = -\log\left(P_0(\mathbf{x}|\boldsymbol{\theta})\right) - \log\left(P(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\right) \tag{4.24}$$

From (3.56) combined with (3.58) and (3.59) we get the Hamiltonian, denoted by E not to confuse it with the entropy H (the notation generally used in physics for the Hamiltonian, the

entropy symbol being S):

$$E(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = -\sum_i^N \log\left(P_0^i(x_i)\right) + \frac{1}{2\Delta}\sum_\mu^M \left(y_\mu - (\mathbf{Fx})_\mu\right)^2 + \frac{M}{2}\log(2\pi\Delta) \tag{4.25}$$

$$= -\sum_i^N \log[(1-\rho)\delta(x_i) + \rho\phi(x_i)] + \frac{1}{2\Delta}\sum_\mu^M \left(y_\mu - (\mathbf{Fx})_\mu\right)^2 + \frac{M}{2}\log(2\pi\Delta) \tag{4.26}$$

4.1.7 Naive mean field estimation for compressed sensing

Let us now derive the naive mean field solution, approximating the true posterior as fully factorizable (4.13). In order to find what are the best parameters of the approximate distribution, we first need to write the variational free energy for the present problem. Using the variational free energy definition (4.7) we obtain:

$$F_Q(\boldsymbol{\theta}_Q) = -\sum_i^N \mathbb{E}_{Q_i}\left(\log(P_0^i(x_i)) - \log(Q_i(x_i))\right)$$

$$+ \frac{1}{2\Delta}\sum_\mu^M \mathbb{E}_Q\left(\left(y_\mu - (\mathbf{Fx})_\mu\right)^2\right) + \frac{M}{2}\log(2\pi\Delta) \tag{4.27}$$

$$= \sum_i^N KL(Q_i||P_0^i) + \frac{M}{2}\log(2\pi\Delta) \tag{4.28}$$

$$+ \frac{1}{2\Delta}\sum_\mu^M \left(y_\mu^2 - 2y_\mu(\mathbf{Fa})_\mu + \sum_{i,j\neq i}^{N,N} F_{\mu i}F_{\mu j}a_i a_j + \sum_i^N F_{\mu i}^2(v_i + a_i^2)\right) \tag{4.29}$$

$$= \sum_i^N KL(Q_i||P_0^i) + \frac{M}{2}\log(2\pi\Delta) + \frac{1}{2\Delta}\sum_\mu^M \left([y_\mu - (\mathbf{Fa})_\mu]^2 + (\mathbf{F}^2\mathbf{v})_\mu\right) \tag{4.30}$$

where we have used the Kullback-Leibler divergence definition (4.1), the additivity property of the entropy for independent variables (3.39) together with the definition of the marginal means and variances of Q:

$$a_i := \int dx_i x_i Q_i(x_i) \tag{4.31}$$

$$v_i := \int dx_i x_i^2 Q_i(x_i) - a_i^2 \tag{4.32}$$

Now that we have the naive mean field Gibbs free energy, we can figure out the expression of the marginals $\{Q_i(x_i)\}_i^N$. We perturbe one of the marginals $Q_i \to Q_i + \delta Q_i$, the associated perturbed distribution is denoted as $\tilde{Q}$. The perturbation term of the Gibbs free energy must

cancel at the minimum for any $\mathbf{x}$:

$$\begin{aligned}
&F_Q(\boldsymbol{\theta}_Q) - F_{\tilde{Q}}(\boldsymbol{\theta}_Q) = 0 \\
\Rightarrow &\int dx_i \delta Q_i(x_i) \left[\frac{1}{2\Delta} \sum_\mu^M \left\{ x_i 2F_{\mu i} \left(\sum_{j\neq i}^N F_{\mu j} a_j - y_\mu \right) + x_i^2 F_{\mu i}^2 \right\} + 1 + \log\left(\frac{Q_i(x_i)}{P_0^i(x_i)} \right) \right] = 0 \\
\Rightarrow &Q_i(x_i|\mathbf{F},\mathbf{a}) = \frac{1}{Z_i(\mathbf{F},\mathbf{a})} P_0^i(x_i) \exp\left(\frac{x_i}{\Delta} \sum_\mu^M F_{\mu i} \left(y_\mu - \sum_j^N F_{\mu j} a_j + F_{\mu i} a_i \right) - \frac{x_i^2}{2\Delta} \sum_\mu^M F_{\mu i}^2 \right) \quad (4.33)
\end{aligned}$$

Defining the following quantities:

$$\Sigma_i^2 := \frac{\Delta}{\sum_\mu^M F_{\mu i}^2} \quad (4.34)$$

$$R_i := a_i + \frac{\Sigma_i^2}{\Delta} \sum_\mu^M F_{\mu i} \left(y_\mu - (\mathbf{Fa})_\mu \right) \quad (4.35)$$

we obtain after simplification of (4.33) the following form of the marginal distributions for compressed sensing under the naive mean field approximation: a product of the prior and a Gaussian mean field that summarize the influence of all the other variables on the i^{th} one:

$$Q_i(x_i|R_i,\Sigma_i^2) = \frac{1}{Z_i(R_i,\Sigma_i^2)} P_0^i(x_i) \exp\left(-\frac{(x_i - R_i)^2}{2\Sigma_i^2} \right) \quad (4.36)$$

From this we can compute the Kullback-Leibler divergence (3.49) appearing in the previous mean field free energy (4.30):

$$\begin{aligned}
\sum_i^N KL(Q_i||P_0^i) &= \sum_i^N \int dx_i Q_i(x_i) \log\left(\frac{\exp\left(-\frac{(x_i-R_i)^2}{2\Sigma_i^2} \right)}{Z_i} \right) &(4.37)\\
&= \sum_i^N \int dx_i Q_i(x_i) \left[-\frac{(x_i - R_i)^2}{2\Sigma_i^2} - \log(Z_i) \right] &(4.38)\\
&= -\sum_i^N \left(\log(Z_i) + \frac{v_i + (a_i - R_i)^2}{2\Sigma_i^2} \right) &(4.39)
\end{aligned}$$

Adding a time index to these equations gives us the mean field algorithm for compressed sensing Fig. 4.2, where we use the non linear thresholding functions:

$$f_a(\Sigma_i^2, R_i) := \frac{1}{Z(R_i,\Sigma_i^2)} \int dx_i \; x_i \; P_0^i(x_i) \exp\left(-\frac{(x_i - R_i)^2}{2\Sigma_i^2} \right) \quad (4.40)$$

$$f_c(\Sigma_i^2, R_i) := \frac{1}{Z(R_i,\Sigma_i^2)} \int dx_i \; x_i^2 \; P_0^i(x_i) \exp\left(-\frac{(x_i - R_i)^2}{2\Sigma_i^2} \right) - f_a(\Sigma_i^2, R_i)^2 \quad (4.41)$$

$t \leftarrow 0$
$\delta \leftarrow \epsilon + 1$
$\Sigma_i^2 \leftarrow \frac{\Delta}{\sum_\mu^M F_{\mu i}^2} \; \forall \; i$
while $t < t_{max}$ **and** $\delta_{max} > \epsilon$ **do**
    $R_i^{t+1} \leftarrow a_i^t + \frac{\Sigma_i^2}{\Delta} \sum_\mu^M F_{\mu i} \left(y_\mu - (\mathbf{F}\mathbf{a}^t)_\mu \right)$
    $a_i^{t+1} \leftarrow f_a \left((\Sigma_i)^2, R_i^{t+1} \right)$
    $t \leftarrow t + 1$
    $\delta \leftarrow ||\mathbf{a}^t - \mathbf{a}^{t-1}||_2^2$
end while
return $\mathbf{a}^t$

Figure 4.2 – The mean field (or iterative thresholding) algorithm for compressed sensing. ϵ is the accuracy for convergence and t_{max} the maximum number of iterations. A suitable initialization for the quantities is $(a_i^{t=0} = \mathbb{E}_{P_0}(x_i))$. Once the algorithm has converged, i.e. the quantities do not change anymore from iteration to iteration, the estimate of the i^{th} signal component is a_i^t. The nonlinear thresholding function f_a take into account the prior distribution $P_0(\mathbf{x})$.

with normalization

$$Z(R_i, \Sigma_i^2) := \int dx_i P_0^i(x_i) \exp\left(-\frac{(x_i - R_i)^2}{2\Sigma_i^2} \right) \tag{4.42}$$

This algorithm directly computes the estimates $\mathbf{a}$ of the signal components. The only non linear part in the mean field algorithm is the component-wise computations of $\mathbf{a}$ and $\mathbf{v}$ through these thresholding functions, the rest is linear and can be written in a efficient way with matrix operations which implies a parallel updating scheme of the estimates. It is possible to think about a randomized updating scheme as well which can sometimes help the convergence in message-passing algorithms as discussed in [76, 77] but the payoff is a slowing down of the algorithm as matrix operations are greatly optimized. In all this thesis, we will always consider a parallel updates scheme, but it must be kept in mind that other strategies are possible with their own paybacks and advantages. The study of the perfomances of this algorithm written in the present form and the comparisons with the approximate message-passing algorithm can be found in [78] and it appears that despite good results in compressed sensing, the approximate message-passing algorithm that will be derived in sec. 4.3 has a greater potential. The mean field approximation though can be asymptotically exact in models where the variance of the mean field felt by the variables goes to zero in the thermodynamic limit, for example in spin models such as the Ising fully connected ferromagnet [79]. It also worth noticing that this mean field algorithm Fig. 4.2 is exactly equivalent to the iterative thresholding algorithm [78, 80] if $\mathbf{F}$ is properly rescaled such that $\sum_\mu F_{\mu i}^2 = 1 \; \forall \; i$ because defining the residual $\mathbf{z}^t = \mathbf{y} - \mathbf{F}\mathbf{a}^t$, the iterations become:

$$\mathbf{z}^t = \mathbf{y} - \mathbf{F}\mathbf{a}^t \tag{4.43}$$

$$\mathbf{a}^{t+1} = \eta_\Delta(\mathbf{F}^\intercal \mathbf{z}^t + \mathbf{a}^t) \tag{4.44}$$

where $\eta_\Delta(x) = f_a(\Delta, x)$, which is the classical form of the iterative thresholding algorithm.

4.2 Belief propagation and cavities

Let us now present a more advanced mean field algorithm, the so called belief propagation algorithm BP, which allows to reach way better performances than the naive mean field approximation in many problems. The assumption behind it is that the factor graph associated to the distribution we want to sample has a tree structure: a tree is a graph such that there exist a *unique* path between two variables in the graph. In this case BP is exact [24], as we will show in sec. 4.2.5. But more generally, BP is justified (but not strictly exact) when the distribution to sample is a Bethe measure as we will see in sec. 4.2.2. From now on, we use $m_i(x_i)$ for the BP estimates of the true marginals $P_i(x_i)$.

4.2.1 The canonical belief propagation equations

The canonical BP equations, that allow to estimate the marginals $\{P_i(x_i)\}_i^N$ of a factorized distribution of the form (4.8) with an associated tree-like factor graph, i.e. a graph such that locally its structure is a tree despite the existence of "long" (of extensive size) loops in the graph, are given by:

$$\hat{m}_{ai}^t(x_i) = \frac{1}{\hat{z}_{ai}^t} \int d\mathbf{x}_{a\setminus i}\, \psi_a(\mathbf{x}_{a\setminus i}, x_i) \prod_{j\in\partial a\setminus i} m_{ja}^t(x_j) \tag{4.45}$$

$$m_{ia}^{t+1}(x_i) = \frac{1}{z_{ia}^{t+1}} \prod_{b\in\partial i\setminus a} \hat{m}_{bi}^t(x_i) \tag{4.46}$$

These quantities allow for the estimation of the marginals at time step t through:

$$m_i^t(x_i) = \frac{1}{z_i^t} \prod_{a\in\partial i} \hat{m}_{ai}^t(x_i) \tag{4.47}$$

$$= \frac{1}{z_{ia}'^t} \hat{m}_{ai}^t(x_i) m_{ia}^{t+1}(x_i) \quad \text{for any } a \in \partial i \tag{4.48}$$

$$m_a^t(\mathbf{x}_a) = \frac{\psi_a(\mathbf{x}_a)}{z_a^t} \prod_{i\in\partial a} m_{ia}^t \tag{4.49}$$

A graphical representation of the BP equations is depicted on Fig. 4.3. The distributions (4.45) and (4.46) that are iteratively computed are the so called *cavity messages* (or simply messages), this vocabulary coming from the cavity method of statistical physics [24, 67]. This is because BP can be thought as the replica symmetric cavity equations on a single graph, or equivalently the replica symmetric cavity equations are the BP equations on an infinitely large graph or averaged over an infinite number of finite random graphs, each representing a random instance of the problem of interest. This last remark is true only if the problem is replica symmetric, i.e. the number of fixed points of the BP equations on an instance of the problem is sub-exponential in N. In the present thesis, the BP fixed point is unique when

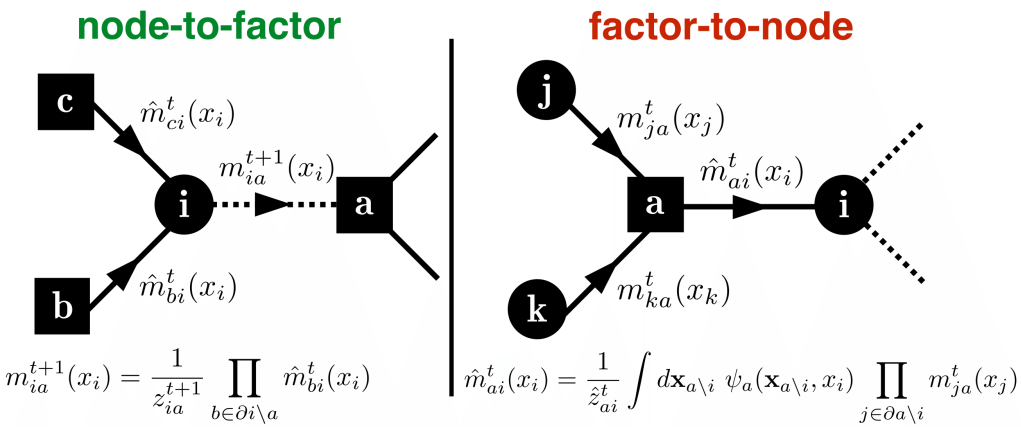

Figure 4.3 – Graphical representation of the belief propagation equations, with the associated equations below. The dashed lines are edges present in the original factor graph which are considered not present in the cavity graph in which the cavity message is computed. On the left is the node-to-factor cavity message, the probability distribution of the x_i variable in a cavity graph where the edge (i,a) has been removed, i.e. a graph where x_i is connected to all its neighbor factors except the factor a. On the right is the factor-to-node cavity message, the probability distribution of the x_i variable in a cavity graph where all the edges connected to x_i are removed except (a,i), i.e. a graph where x_i is only connected the factor a, not its other neighbors in the original factor graph.

starting from a random initial condition of the messages. This does not imply the uniqueness of the BP fixed point, there can be another one but that require the messages to be initialized "close" to it in order to see the messages converge to this other fixed point as we will see in sec. 5.1.1: there is a regime where the message-passing always converges to a wrong solution despite the existence of another fixed point corresponding to the true solution of the problem. Another replica symmetric example with multiple fixed points is the Ising ferromagnet case where there exist two fixed points below the critical temperature corresponding to positive and negative average magnetizations. In the replica symmetric case, the messages (or any other thermodynamical quantity) are self-averaging: their fixed points values do not depend on the graph details (the measurement matrix $\mathbf{F}$ realization in linear estimation) if large enough, the fluctuations being $\in O(1\sqrt{N})$. This is why the $MMSE$ estimator discussed in sec. 3.6.2 is self averaging as well: the approximate message-passing algorithm (see sec. 4.3) used to compute it is directly derived from BP and applied to inference problems, which are replica symmetric under proper conditions (at least under the prior matching condition [57], see sec. 3.6).

4.2.2 Understanding belief propagation in terms of cavity graphs

The factor-to-node cavity message (4.45) is interpreted as the (approximate) probability distribution of x_i in a modified graphical model, refered as a *cavity graph*, where x_i is only connected to a, not anymore to its other factor neighbors in the original graph, see right part on Fig. 4.3. This distribution carries the information on how strongly the factor a depends

on the variable x_i, so its influence on x_i. The node-to-factor cavity message (4.46) is the probability of x_i in another cavity graph where x_i is connected to all its neighbors except a, the edge between them being removed, see left part on Fig. 4.3. This message carries the information on the influence of all the neighbors factors on x_i except a.

With this interpretation, the BP equations can be understood easily. As the graph is assumed to be a tree, the set of factor-to-node messages coming on x_i are conditionally independent, thus they just multiply. The node-to-factor message $m_{ia}(x_i)$ of x_i with the edge (i,a) removed is thus naturally given as the product of all the factor-to-node messages $\{\hat{m}_{bi}(x_i)\}_{b\in\partial i\setminus a}$ except the one coming from a (4.46). Now the factor-to-node message $\hat{m}_{ai}(x_i)$: we consider the cavity graph where x_i is only connected to the factor a, which is equivalent to assume $m_{ia}(x_i) = C$ in this cavity graph, i.e. the distribution of x_i in this cavity graph where we additionally removed the edge (i,a) is uniform (x_i does not feel any constraint). Now, the joint distribution of all the neighbors of the factor a is (up to a normalization) the product of their cavity distributions $\prod_{j\in\partial a} m_{ja}(x_j) \propto \prod_{j\in\partial a\setminus i} m_{ja}(x_j)$, i.e. their joint distribution in a graph where the factor a is not here, that we multiply by the compatibility function $\psi_a(\mathbf{x}_a)$ to include back its influence. Thus the cavity message $\hat{m}_{ai}(x_i)$ is the marginalization of this joint distribution with respect to the neighbors other than x_i that we must normalize, from which we get the equation (4.45). Finally the marginal (4.47) is given as the product of all the individual influences of the factors neighbors to x_i.

4.2.3 Derivation of belief propagation from cavities and the assumption of Bethe measure

We gave here arguments justifying a posteriori the BP equations thanks to cavity graphs, but is there a way to directly derive the BP equations starting from cavity graphs? The answer is given by a very insightful exercise (exercise 19.1) extracted from [24]. Let us assume we have a distribution $P(\mathbf{x})$ associated to a factor graph $G = (\mathcal{V},\mathcal{F},\mathcal{E})$ (see sec. 4.1.5 for the definition of factor graphs). A *cavity* $C = (\mathcal{V}_C,\mathcal{F}_C,\mathcal{E}_C)$ of the graph G is a sub-graph of G such that if any factor a is included in C, then all its neigboring variable nodes ∂a are in C as well:

$$a \in \mathcal{F}_C \Rightarrow \partial a \in \mathcal{V}_C \tag{4.50}$$

The boundary ∂C of the cavity is the set of edges $\{(a,i)\}_{a\notin\mathcal{F}_C, i\in\mathcal{V}_C}$ connecting variable nodes inside the cavity to factors outside of it, see Fig. 4.4. The probability distribution $P(\mathbf{x})$ is a *Bethe measure* if there exist a set of cavity messages $\{\hat{m}_{ai}(x_i)\}$ such that if $P(\mathbf{x})$ is restricted to any non extensive cavity C, $P(\mathbf{x}_C)$ can be expressed (up to a small error that goes to zero in the thermodynamic limit) as a local bulk term and a boundary contribution obtained from the messages:

$$P(\mathbf{x}_C) \approx \frac{1}{z_C} \prod_{a\in\mathcal{F}_C} \psi_a(\mathbf{x}_a) \prod_{(a,i)\in\partial C} \hat{m}_{ai}(x_i) \tag{4.51}$$

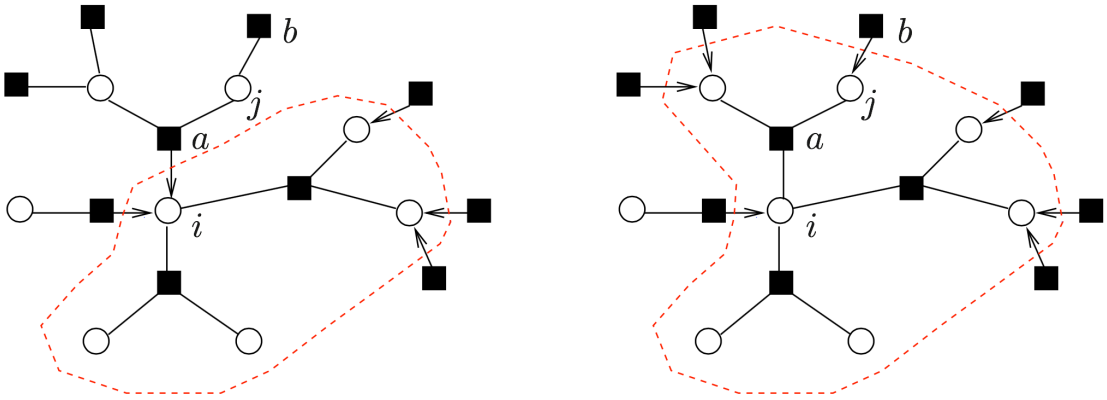

Figure 4.4 – Figure taken from [24]. Two examples of cavity graphs. A cavity graph must include all the variable nodes neighbors to any factor in it. The cavity on the left is extended to include one more factor, and thus the two additional variable nodes i and j as well for it to remain a cavity. The consistency constraints of the Bethe measure for these two cavities imply the belief propagation equations for the messages sent from the boundary of the first cavity.

Now let us assume that we have defined some cavity C, for example the left one on Fig. 4.4 and decide to extend it including the factor a in it, and thus its two other neighboring variable nodes that were not in C as well in order to maintain the cavity definition. We obtain the new cavity $\tilde{C} = (\mathcal{V}_{\tilde{C}} = \mathcal{V}_C \cup \partial a, \mathcal{F}_{\tilde{C}} = \mathcal{F}_C \cup a, \mathcal{E}_{\tilde{C}} = \mathcal{E}_C \cup \{(a,i)\}_{i\in\partial a})$, the right one on Fig. 4.4. As $P(\mathbf{x})$ is assumed to be a Bethe measure, the distribution of this new cavity can be expressed as the previous cavity distribution (4.51) times the new factor included and the new boundary contribution. But one must be careful to divide the result by the boundary contribution $\hat{m}_{ai}(x_i)$ in $P(\mathbf{x}_C)$ that is now included in the bulk and thus overcounted:

$$\tilde{P}(\mathbf{x}_{\tilde{C}}) = \tilde{P}(\mathbf{x}_C, \mathbf{x}_{a\setminus i}) \approx \frac{1}{z_{\tilde{C}}} P(\mathbf{x}_C) \frac{\psi_a(\mathbf{x}_{a\setminus i}, x_i)}{\hat{m}_{ai}(x_i)} \prod_{j\in\partial a\setminus i} \prod_{b\in\partial j\setminus a} \hat{m}_{bj}(x_j) \tag{4.52}$$

For these distributions to be coherent, the marginalization constraint must be verified:

$$P(\mathbf{x}_C) = \int \tilde{P}(\mathbf{x}_C, \mathbf{x}_{a\setminus i}) d\mathbf{x}_{a\setminus i} \tag{4.53}$$

$$= \frac{1}{z_{\tilde{C}}} P(\mathbf{x}_C) \frac{1}{\hat{m}_{ai}(x_i)} \int d\mathbf{x}_{a\setminus i} \psi_a(\mathbf{x}_{a\setminus i}, x_i) \prod_{j\in\partial a\setminus i} \prod_{b\in\partial j\setminus a} \hat{m}_{bj}(x_j) \tag{4.54}$$

$$\Rightarrow \hat{m}_{ai}(x_i) = \frac{1}{z_{\tilde{C}}} \int d\mathbf{x}_{a\setminus i} \psi_a(\mathbf{x}_{a\setminus i}, x_i) \prod_{j\in\partial a\setminus i} \prod_{b\in\partial j\setminus a} \hat{m}_{bj}(x_j) \tag{4.55}$$

$$= \frac{1}{\hat{z}_{ai}} \int d\mathbf{x}_{a\setminus i} \psi_a(\mathbf{x}_{a\setminus i}, x_i) \prod_{j\in\partial a\setminus i} m_{ja}(x_j) \tag{4.56}$$

$$m_{ja}(x_j) = \frac{1}{z_{ja}} \prod_{b\in\partial j\setminus a} \hat{m}_{bj}(x_j) \tag{4.57}$$

where we have been careful to always normalize distributions. We find back the belief propagation equations (4.45), (4.46) at their fixed point, which are thus implied by the assumption

that the distribution is a Bethe measure.

4.2.4 When does belief propagation work

Despite that BP is exact only on trees, it can be a very good approximation for any tree-like graph or more generaly as long as the distribution to sample is a Bethe measure as shown in the previous section. When used on such graph, the algorithm is referred as loopy-BP and if the algorithm converges, the BP marginals (4.47) are estimates of the true ones under the tree approximation. These loops can a priori induce correlations not taken into account by the BP equations but if they are "long enough", the induced correlations decrease so fast to zero that BP becomes quickly almost exact [24, 81, 81]. When does this situation occur? Hopefully, such locally tree-like graphs appear naturally in many applications: combinatorial optimization problems, modern coding theory, neural networks, artificial intelligence, etc. The tree-like property of the factor graphs in all these fields is a consequence of a common feature: these are all sparse random graphs, i.e. these are randomly generated graphs with a fixed average connectivity which does not scale with N, the number of random variables in the problem. It can be shown [24, 82] that sparse random graphs have loops which size typically grows as $O(\log(N))$, thus the correlations in such graphs decay fast enough for BP to converge and accuractly approximate the marginal distributions.

Of course, these generic considerations are not always true. In many graphical models, typically in combinatorial optimization problems, this assumption of small correlations can break down in certain parameters regimes despite the sparsity of the graph and BP cannot sample anymore the marginals as the space of solution splits into an exponential (in N) number of disconnected clusters of solutions. This scenario is referred as replica symmetry breaking [1, 24, 67, 81, 82] but is out of the scope of this thesis: inference problems always have a solution by definition, and this prevents the replica symmetry breaking phenomenon to occur, at least in the case of the prior matching condition when the true generating model of the signal is know [57]. All the theoretical analisies in this thesis will assume this condition, see sec. 4.3 and sec. 9.6.

There exist different alternatives to include part of the correlations induced by the loops in the graph, not taken into accout by BP in its canonical form. A very general and popular one is refered as generalized belief propagation algorithms [83], another technique is the loop corrected belief propagation [84]. But in problems with a glassy phenomenology where anyway the system is deep in its replica symmetry broken phase where its measure really splits into exponentially many ones, this phenomenon *must* be taken into account and more advanced algorithms such as survey propagation should be considered [1, 64, 82, 85–87].

4.2.5 Belief propagation and the Bethe free energy

We previously assumed that the BP algorithm, that can be thought as a dynamical process on a graph was exact on trees, i.e. its fixed point marginals were the exact ones $m_i(x_i) = \int d\mathbf{x}_{\setminus i} P(\mathbf{x})$. Let us prove that it is actually the case. In the case of a tree factor graph, the probability distribution of the variables can be written exactly as a factorized distribution over the single variable marginals $\{P_i(x_i)\}_i^N$ and the "factor marginals" $\{P_a(\mathbf{x}_a)\}_a^G$:

$$P(\mathbf{x}) = \frac{1}{Z}\prod_a^G \psi_a(\mathbf{x}_a) = \prod_a^G P_a(\mathbf{x}_a)\prod_i^N P_i(x_i)^{1-c_i} \tag{4.58}$$

where c_i denotes the connectivity of the node i, the number of factors to which the variable x_i is connected to. The inductive proof can be found in [24]. From this we can compute the associated Gibbs free energy using (4.6). We skeep the possible dependencies in parameters. The average energy part is found using (4.11), (4.12):

$$\mathbb{E}_P\left(\sum_a^G E_a(\mathbf{x}_a)\right) = -\sum_a^G \mathbb{E}_{P_a}\left(\log\left(\psi(\mathbf{x}_a)\right)\right) \tag{4.59}$$

Then the entropy part using (3.34) with (4.58):

$$H(P) = -\sum_a^G \mathbb{E}_{P_a}\left(\log\left(P_a(\mathbf{x}_a)\right)\right) - \sum_i^N (1-c_i)\mathbb{E}_{P_i}\left(\log\left(P_i(x_i)\right)\right) \tag{4.60}$$

Thus the Gibbs free energy for a tree (which is also its true Helmotz free energy as (4.58) is exact for a tree) is given by:

$$\begin{aligned} F_{Bethe}(\{P_a,\psi_a\}_a^G,\{P_i\}_i^N) = &-\sum_a^G \int d\mathbf{x}_a P_a(\mathbf{x}_a)\log\left(\frac{\psi_a(\mathbf{x}_a)}{P_a(\mathbf{x}_a)}\right) \\ &-\sum_i^N (c_i-1)\int dx_i P_i(x_i)\log\left(P_i(x_i)\right) \end{aligned} \tag{4.61}$$

This free energy is also referred as the Bethe free energy. Let us compute the marginals, used in the parametrization (4.58). As discussed in sec. 4.1.1, we thus minimize the Bethe free energy to find their expressions, but we also need to be careful to enforce their normalization and the marginalization conditions to get coherent definitions of probability distributions. These two conditions together imply the normalization of the full distribution, we dont need a additional Lagrange multiplier to enforce it. We thus create a Lagrangian from the Bethe free energy:

$$\begin{aligned} &\mathscr{L}_{Bethe}(\{P_a,\psi_a\}_a^G,\{P_i\}_i^N) = F_{Bethe}(\{P_a,\psi_a\}_a^G,\{P_i\}_i^N) + \sum_i^N \gamma_i\left(\int dx_i P_i(x_i) - 1\right) \\ &+\sum_a^G \sum_{i\in\partial a}\int dx_i \lambda_{ai}(x_i)\left(\int d\mathbf{x}_{a\setminus i} P_a(\mathbf{x}_a) - P_i(x_i)\right) \end{aligned} \tag{4.62}$$

Now we optimize it with respect to the one point marginal by functionnal derivation:

$$\frac{\delta \mathscr{L}_{Bethe}}{\delta P_i(x_i^*)} = -(c_i - 1)\left[\log\left(P_i(x_i^*)\right) + 1\right] + \gamma_i - \sum_{a \in \partial i} \lambda_{ai}(x_i^*) = 0 \tag{4.63}$$

$$\Rightarrow P_i(x_i^*) = \frac{1}{z_i} \exp\left(-\frac{1}{c_i - 1} \sum_{a \in \partial i} \lambda_{ai}(x_i^*)\right) \tag{4.64}$$

where we put all the constants in the normalization. And now the factor marginals:

$$\frac{\delta \mathscr{L}_{Bethe}}{\delta P_a(\mathbf{x}_a^*)} = \log\left(\frac{P_a(\mathbf{x}_a^*)}{\psi_a(\mathbf{x}_a^*)}\right) + 1 + \sum_{i \in \partial a} \lambda_{ai}(x_i^*) = 0 \tag{4.65}$$

$$\Rightarrow P_a(\mathbf{x}_a^*) = \frac{\psi_a(\mathbf{x}_a^*)}{z_a'} \exp\left(-\sum_{i \in \partial a} \lambda_{ai}(x_i^*)\right) \tag{4.66}$$

Now using the following reparametrization in (4.64) we get:

$$\lambda_{ai}(x_i^*) = -\sum_{b \in \partial i \setminus a} \log\left(\hat{m}_{bi}(x_i^*)\right) \tag{4.67}$$

$$\Rightarrow \sum_{a \in \partial i} \lambda_{ai}(x_i^*) = -(c_i - 1) \sum_{a \in \partial i} \log\left(\hat{m}_{ai}(x_i^*)\right) \tag{4.68}$$

$$\Rightarrow P_i(x_i) = \frac{1}{z_i} \prod_{a \in \partial i} \hat{m}_{ai}(x_i) \tag{4.69}$$

$$= m_i(x_i) \tag{4.70}$$

where we recognized the BP marginal expression (4.47). And now applying the same for the factor marginals (4.66):

$$\Rightarrow P_a(\mathbf{x}_a) = \frac{\psi_a(\mathbf{x}_a)}{z_a'} \prod_{i \in \partial a} \prod_{b \in \partial i \setminus a} \hat{m}_{bi}(x_i) \tag{4.71}$$

$$= \frac{\psi_a(\mathbf{x}_a)}{z_a} \prod_{i \in \partial a} m_{ia}(x_i) \tag{4.72}$$

$$= m_a(\mathbf{x}_a) \tag{4.73}$$

where we have used (4.46) and (4.49) *at the fixed point* of BP, i.e. droping the time index. We thus realize that the fixed point marginals $\{P_i\}_i^N$ and $\{P_a\}_a^G$ of the Bethe free energy (4.64) gives back the BP marginals $\{m_i\}_i^N$ and $\{m_a\}_a^G$. A very nice review on graphical models and the links between belief propagation and the Bethe free energy is [88].

Belief propagation can be generalized to optimize more complex variational free energies, that take into acount more complex probabilistic models. Some classical mean field models include the Kikuchi and junction tree approximations. These message-passing algorithms are refered as generalized belief propagation algorithms [83]. In this framework, the Bethe free energy is a particular choice of parametrization (4.58) for the distribution of the model. But in the present thesis, dense graphical models with linear constraints are studied, and in this case the Bethe free energy is asymptotically exact as we will see in sec. 4.3.

4.2.6 The Bethe free energy in terms of cavity messages

The Bethe free energy (4.61) is a formal expression useful to show the equivalence with the BP fixed points as done in the previous section, but is not very practical from the computational point of view. We now derive another equivalent expression of it, expressed in terms of the cavity messages at their fixed point. This will be useful when deriving expectation maximization learning equations for example. We start from the expression (4.61). Using the expression of the marginals as a function of the cavity messages (4.69), (4.72) the free energy becomes:

$$
\begin{aligned}
F_{Bethe} &= -\sum_a^G \int d\mathbf{x}_a P_a(\mathbf{x}_a) \log\left(\frac{z_a}{\prod_{i\in\partial a} m_{ia}(x_i)}\right) \\
&-\sum_i^N (c_i-1)\int dx_i P_i(x_i)\log\left(\frac{\prod_{a\in\partial i}\hat{m}_{ai}(x_i)}{z_i}\right) \qquad (4.74)\\
&= \underbrace{-\sum_a^G \log(z_a) - \sum_i^N \log(z_i)}_{:=\tilde{F}} + \sum_i^N c_i \log(z_i) \\
&+\sum_a^G \sum_{i\in\partial a}\int d\mathbf{x}_a P_a(\mathbf{x}_a)\log(m_{ia}(x_i)) - \sum_i^N (c_i-1)\int dx_i P_i(x_i)\log\left(\prod_{a\in\partial i}\hat{m}_{ai}(x_i)\right) \qquad (4.75)
\end{aligned}
$$

Now we use the following identity that is a direct consequence of the definition of the cavity messages:

$$
\prod_{b\in\partial i}\hat{m}_{bi}(x_i) = m_{ia}(x_i)\hat{m}_{ai}(x_i)z_{ia} \text{ for any } a\in\partial i \qquad (4.76)
$$

where z_{ia} is the normalization of $m_{ia}(x_i)$. Now using the fact that $\sum_i^N c_i f_i = \sum_i^N \sum_{a\in\partial i} f_i$ for a generic objecf f_i that depends on the variable index (the connectivity can be replaced by an additional sum over the neighbors factor indices) and using the marginalization property of $P_a(\mathbf{x}_a)$, we deduce:

$$
\begin{aligned}
F_{Bethe} &= \tilde{F} + \sum_a^G \sum_{i\in\partial a}\int P_i(x_i)\log(m_{ia}(x_i)) + \sum_i^N \sum_{a\in\partial i}\log(z_i) \\
&-\sum_i^N \sum_{a\in\partial i}\left[\int dx_i P_i(x_i)\left(\log(m_{ia}(x_i))+\log(\hat{m}_{ai}(x_i))\right)+\log(z_{ia})\right] \\
&+\sum_i^N \left[\int dx_i P_i(x_i)\left(\log(m_{ia}(x_i))+\log(\hat{m}_{ai}(x_i))\right)+\log(z_{ia})\right] \qquad (4.77)
\end{aligned}
$$

Now we use that the sums $\sum_a^G \sum_{i\in\partial a} f_{ia} = \sum_i^N \sum_{a\in\partial i} f_{ia}$ are equal (f_{ia} is any function depending on the variables and factors indices) and the identity:

$$
\begin{aligned}
P_i(x_i) &= \frac{z_{ia}}{z_i} m_{ia}(x_i)\hat{m}_{ai}(x_i) \qquad (4.78)\\
\Rightarrow z_i &= z_{ia}\int dx_i m_{ia}(x_i)\hat{m}_{ai}(x_i) \qquad (4.79)
\end{aligned}
$$

We thus obtain:

$$
\begin{aligned}
F_{Bethe} &= \tilde{F} + \sum_{i}^{N} \sum_{a \in \partial i} \left[\log(z_{ia}) + \log\left(\int dx_i m_{ia}(x_i) \hat{m}_{ai}(x_i) \right) \right] \\
&- \sum_{i}^{N} \sum_{a \in \partial i} \left[\int dx_i P_i(x_i) \log(\hat{m}_{ai}(x_i)) + \log(z_{ia}) \right] \\
&+ \sum_{i}^{N} \left[\int dx_i P_i(x_i) \left(\log(m_{ia}(x_i)) + \log(\hat{m}_{ai}(x_i)) \right) + \log(z_{ia}) \right] \qquad (4.80) \\
&= \tilde{F} + \sum_{i}^{N} \sum_{a \in \partial i} \log\left(\int dx_i m_{ia}(x_i) \hat{m}_{ai}(x_i) \right) - \sum_{i}^{N} \sum_{a \in \partial i} \int dx_i P_i(x_i) \log(\hat{m}_{ai}(x_i)) \\
&+ \sum_{i}^{N} \left[\int dx_i P_i(x_i) \left(\log\left(\frac{1}{z_{ia}} \prod_{b \in \partial i \setminus a} \hat{m}_{bi}(x_i) \right) + \log(\hat{m}_{ai}(x_i)) \right) + \log(z_{ia}) \right] \qquad (4.81) \\
&= \tilde{F} + \sum_{i}^{N} \sum_{a \in \partial i} \log\left(\int dx_i m_{ia}(x_i) \hat{m}_{ai}(x_i) \right) \qquad (4.82)
\end{aligned}
$$

So the final expression of the Bethe free energy in terms of the cavity messages fixed point is:

$$
\begin{aligned}
F_{Bethe} &= -\sum_{a}^{G} \log(z_a) - \sum_{i}^{N} \log(z_i) + \sum_{i}^{N} \sum_{a \in \partial i} \log(\tilde{z}_{ia}) \qquad (4.83) \\
z_a &= \int d\mathbf{x}_a \psi_a(\mathbf{x}_a) \prod_{i \in \partial a} m_{ia}(x_i) \qquad (4.84) \\
z_i &= \int dx_i \prod_{a \in \partial i} \hat{m}_{ai}(x_i) \qquad (4.85) \\
\tilde{z}_{ia} &:= \int dx_i m_{ia}(x_i) \hat{m}_{ai}(x_i) \qquad (4.86)
\end{aligned}
$$

This expression is only true at the fixed points of the messages, but at any time step t of the algorithm, an approximated free energy can be computed plugging the messages at this time in this expression. This formula can be understood in the following way: the total free energy on a tree graphical model is the sum of the contributions of each factors and their associated neighborhood (edges and variables), of the individual variable and their adjacent edges contributions but as each edges has been overcounted, we remove each edge contribution once.

4.2.7 Derivation of belief propagation from the Bethe free energy

This form of the Bethe free energy is more practical than the (4.61) because the belief propagation equations can be trivially derived as fixed point equations for this potential. Let us show it for sake of completeness. Starting from (4.83) and performing the functional derivative with

respect to the cavity messages, we obtain at the fixed point:

$$\frac{\delta F_{Bethe}}{\delta \hat{m}_{ai}(x_i^*)} = -\frac{1}{z_i}\int dx_i \delta(x_i - x_i^*) \prod_{b\in\partial i\setminus a} \hat{m}_{bi}(x_i) + \frac{1}{\tilde{z}_{ia}}\int dx_i \delta(x_i - x_i^*) m_{ia}(x_i) = 0 \tag{4.87}$$

$$\Rightarrow m_{ia}(x_i^*) = \frac{1}{z_{ia}} \prod_{b\in\partial i\setminus a} \hat{m}_{bi}(x_i^*) \tag{4.88}$$

$$z_{ia} = \frac{z_i}{\tilde{z}_{ia}} \tag{4.89}$$

In the similar way:

$$\frac{\delta F_{Bethe}}{\delta m_{ia}(x_i^*)} = -\frac{1}{z_a}\int dx_i d\mathbf{x}_{a\setminus i} \psi_a(x_i, \mathbf{x}_{a\setminus i}) \delta(x_i - x_i^*) \prod_{j\in\partial a\setminus i} m_{ja}(x_j) + \frac{1}{\tilde{z}_{ia}}\int dx_i \delta(x_i - x_i^*) \hat{m}_{ai}(x_i) = 0 \tag{4.90}$$

$$\Rightarrow \hat{m}_{ai}(x_i^*) = \frac{1}{\hat{z}_{ai}}\int d\mathbf{x}_{a\setminus i} \psi_a(x_i^*, \mathbf{x}_{a\setminus i}) \prod_{j\in\partial a\setminus i} m_{ja}(x_j) \tag{4.91}$$

$$\hat{z}_{ai} = \frac{z_a}{\tilde{z}_{ia}} \tag{4.92}$$

which are exactly the BP equations (4.45), (4.46) at their fixed point.

4.3 The approximate message-passing algorithm

If the only available information about the signal is the matrix $\mathbf{F}$ and the vector of measurements $\mathbf{y}$ in (3.18), then the information-theoretically best possible estimate of each signal component is computed as a weighted average over all solutions of the linear system, where the weight of each solution is given by the prior. Of course, the undetermined linear system (3.18) has exponentially many (in N) solutions and hence computing exactly the above weighted average is in general intractable. The corresponding expectation to perform inference can be, however, approximated efficiently via the approximate message-passing algorithm [34, 35, 89, 90] that we will present now. But before, let us expose why belief propagation is not the right tool to use in the present context.

In order to be as general as possible, we now consider that the components of the signal we want to infer are B-dimensional vectors $\mathbf{x}_l = [x_i]_{i\in l}$ where l denotes both the vector variable index and the set of indices $\{i \in l\}$ of the scalar components of $\mathbf{x}$ concatenated to form the new vector variable $\mathbf{x}_l$. These new variables are called *sections*. We define $\mathbf{F}_{\mu l} := [F_{\mu i}]_{i\in l}$ as a vector of elements of the matrix $\mathbf{F}$ that act on the section $\mathbf{x}_l$ (see Fig. 4.5). Working with vectors is useful as we will work in this setting for the sparse superposition codes sec. 9 and the scalar equations can be recovered taking $B = 1$ in the final equations. All the previous derivations (the BP algorithm, the Bethe free energy, etc.) would have been the same with vector variables

so the obtained results remain valid. We will denote by L the number of sections $\{\mathbf{x}_l\}_l^L$, in order to keep the notation $N = LB$ for the number of 1-d components of the signal. We take a generic factorizable prior over the sections. Fig. 4.5 shows how a linear estimation problem with a scalar components signal which prior constrain groups of B non overlaping components can be interpreted as an equivalent problem where now the components are B-d sections. The scalar matrix elements are concatenated in B-d vectors as well, and are applied to the vectorial signal components using the usual scalar product between vectors. This construction changes nothing to the scalar measurements, nor to the fact that the noise is i.i.d applied on the 1-d components of the signal.

4.3.1 Why is the canonical belief propagation not an option for dense linear systems over reals?

Belief propagation is a very powerful inference algorithm but it has caveats. We will write the BP equations for the linear estimation problem and understand why the equations are intractable.

The Hamiltonian we consider is thus a direct extension of (4.26) to the vectorial case and we consider that the measurement matrix is full:

$$E(\mathbf{x}) = -\sum_l^L \log\left(P_0^l(\mathbf{x}_l)\right) + \frac{1}{2\Delta}\sum_\mu^M \left(y_\mu - (\mathbf{F}\mathbf{x})_\mu\right)^2 \tag{4.93}$$

A remark is that when there are factors $\{\phi(x_i)\}$ depending on single variables, the node-to-factor BP message (4.46) and the marginal (4.47) can be rewritten as:

$$m_{ia}^{t+1}(x_i) = \frac{1}{z_{ia}^{t+1}}\phi(x_i)\prod_{b\in\partial i\setminus a}\hat{m}_{bi}^t(x_i) \tag{4.94}$$

$$m_i^t(x_i) = \frac{1}{z_i^t}\phi(x_i)\prod_{a\in\partial i}\hat{m}_{ai}^t(x_i) \tag{4.95}$$

$$= \frac{1}{z_{ia}'^t}\phi(x_i)\hat{m}_{ai}^t(x_i)m_{ia}^{t+1}(x_i) \text{ for any } a\in\partial i \tag{4.96}$$

which is a perfectly equivalent form as these single variable factors can be integrated in the set of previous factors $\{\psi_a\}$ considering that they do not receive any node-to-factor messages (i.e. the product of messages in (4.45) is 1 and there is no marginalization to perform as they are only connected to one variable). This form will be more practical to use. Now using that the constraints are given by the likelihood of the observations (3.58) and that the prior is

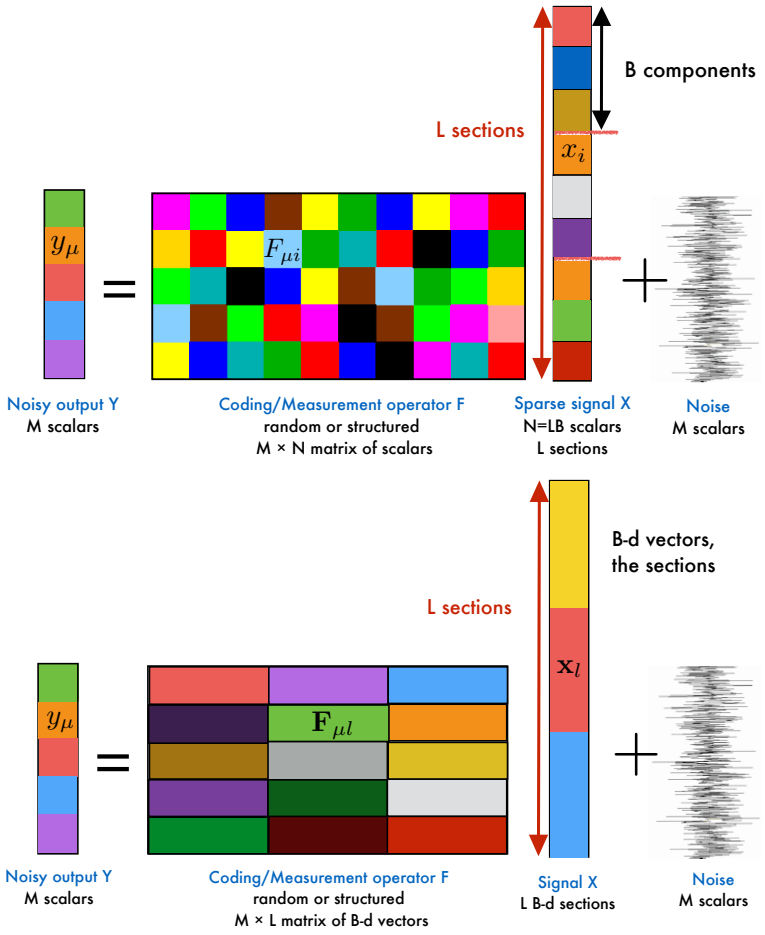

Figure 4.5 – **Up** : Representation of the linear estimation problem over the i.i.d AWGN channel in terms of a signal and matrix with scalar components. The prior on this signal is factorizable over non overlaping groups of B components, denoted as the sections. **Down** : Reinterpreting the same problem in terms of B-d variables. Now, the matrix elements are concatenated to form B-d vectors that are applied (using the usual scalar product for vectors) on the associated B-d vectors representing the new components of the signal, the sections. In this new setting, all the signal sections are uncorrelated by the prior.

factorizable over the signal sections, we get the BP equations for vectorial linear estimation:

$$\hat{m}^t_{\mu l}(\mathbf{x}_l) = \frac{1}{\hat{z}^t_{\mu l}} \int \left[\prod_{k \neq l}^{L-1} d\mathbf{x}_k m^t_{k\mu}(\mathbf{x}_k) \right] e^{-\frac{\text{snr}}{2}\left(\sum_{k\neq l}^{L-1} \mathbf{F}^\intercal_{\mu k}\mathbf{x}_k + \mathbf{F}^\intercal_{\mu l}\mathbf{x}_l - y_\mu\right)^2} \tag{4.97}$$

$$m^{t+1}_{l\mu}(\mathbf{x}_l) = \frac{1}{z^{t+1}_{l\mu}} P^l_0(\mathbf{x}_l) \prod_{\gamma\neq\mu}^{M-1} \hat{m}^t_{\gamma l}(\mathbf{x}_l) \tag{4.98}$$

$$m^t_l(\mathbf{x}_l) = \frac{1}{z^t_l} P^l_0(\mathbf{x}_l) \prod_{\gamma}^{M} \hat{m}^t_{\gamma l}(\mathbf{x}_l) \tag{4.99}$$

Working with the noise variance or the snr $= 1/\Delta$ is the same as we fix the power to be 1. What

are the problems with these equations? There are essentially three:

• The factor graph associated to the linear estimation problem problem Fig. 4.1 (where the 1-d x_i variables are replace by the B-d $\mathbf{x}_l$ ones) is densely connected. The likelihood constraints enforce all the variable nodes to be connected to all the likelihood factor nodes when the measurement matrix is full. This implies that the number of messages to store in the memory and exchange at each time step is $2ML \in O(L^2)$ (2 per edges) which is way too many and scales badly with the problem size.

• Furthermore, these messages are probability distributions over real variables. Such objects are really difficult to store on a computer as in general they have no analytical form that could be decomposed as simple functions. It would require to store a discretized version of each message, a histogram which discretization step is small enough to have an high numerical accuracy. This is impossible as the number of message is large and anyway, it would lower greatly the efficiency of the algorithm.

• As we are working in the continuous framework $\mathbf{x} \in \mathbb{R}^{BL}$, $\mathbf{F} \in \mathbb{R}^{M,BL}$ the computations of the factor-to-node cavity messages (4.97) require very high dimensional integrals to be performed ($(L-1)B$ integrals in the full matrix case) which are non analytic and thus would have to be computed numerically.

We understand that BP in this form is not an option. BP is useful when the factor graph is sparse and when the variables have few discrete states. In this case the integrals become sums over the states of few neighbors which is tractable. Furthermore, there are not too many messages to store due to the graph sparsity (this number scales as the number of variables) and each message can be easily stored as a small vector giving the probability of each discrete state.

Belief propagation based reconstruction algorithms were introduced in compressed sensing by [91]. The authors used sparse measurement matrices to reduce the number of messages and make the graph locally tree-like and then treated the BP messages as probabilities over real numbers, that were represented by a histogram, one of the three major problems of BP discussed before, and that will face the approximate message-passing algorithm.

4.3.2 Why message-passing works on dense graphs?

Let us assume that we have an infinitely powerful computer with infinite memory, such that all the previous problems are not of concern anymore. Is the BP algorithm a good inference algorithm anyway for such linear estimation problems, where the factor graph is dense? We explained in sec. 4.2.5 that BP finds the fixed point marginals of the Bethe free energy which is exact for trees, and suggests that BP is an accurate approximation in the case of sparse graphs because of their locally tree-like structure, see sec. 4.2.4. But here it is not the case at all. It is even the opposite extreme case: the graph is *full* of loops. But actually, such very dense graphs share with tree-like ones the important common feature that makes BP the algorithm

of choice for inference: the correlations between variables are very weak. Let us detail a bit more this notion.

These systems (tree-like and dense graphical models) are equivalent to infinite dimensional systems. This can be seen from the following fact: starting from any node in the system and moving with no coming back (without passing two times by the same edge), it is impossible to return at the starting point. It is trivial for a tree as it is the very definition of what a tree graph is and thus it becomes true with high probability for infinitely large tree-like graphs. On densely connected graphs, it becomes true with high probability as well as the graph size increases, because the number of paths becomes so large that taking one that luckily comes back to the initial point becomes infinitesimaly probable, which is not the case on a 2-d graph like a grid for example.

This is why message-passing works on such graphs: the independence between neighboring variables assumed for computing the cavity messages is asymptotically valid, as the only possible paths that could correlate these variables in the cavity graphs have lenghts that diverge in the random sparse graphs case, and the variables are anyway almost independent in the densely connected case as each variable is connected to all other ones making the influence of each single one asymptotically null.

4.3.3 Derivation of the approximate message-passing algorithm from belief propagation

In order to get an algorithm capable of dealing with continuous variables and dense graphs with linear constraints, the approximate message-passing algorithm, we will start from BP and then perform two principal steps: i) In order to face the problem of storing distributions over real variables, we will parametrize the cavity messages thanks to their first and second moments, i.e. project them on Gaussians. This step is exact in the large signal limit $L \to \infty$ as we will see, and is validated by "law of large numbers like" arguments as the number of incoming message on each factor is very large. ii) We will expand the cavity quantities that depend on factors and variables indices around marginal quantities that depend only on the variables indices. The correction around these, the so called *Onsager reaction term* in statistical physics will be essential for the algorithm performance, and makes all the difference with the naive mean field approximation, see sec. 4.1.7. The resulting algorithm, referred as the Thouless-Palmer-Anderson TAP equations in statistical physics, derived to deal with spin glasses [79, 92], will be obtained for linear estimation: it is the approximate message-passing algorithm. Before the apparition of AMP, message-passing algorithms for dense graphs were already studied [93] but the strategy adopted here is different.

Gaussian parametrization

The cavity messages in BP (4.97), (4.98) can be represented only by their mean and variance as done by [94,95]. Let us see how to derive iterative equations on these moments. For the rest of this derivation, we skeep the time index for sake of readibility, these will be added back at the end and justified in sec. 4.3.4.

In order to fix the power to 1 in (3.18), we use the following scaling for the matrix elements: $F_{\mu i} \in O(1/\sqrt{L}) \ll 1$, as we assume in the derivation that $L \gg 1$. Using this scaling, we will Taylor expand in the matrix elements the exponential appearing in the factor-to-node cavity message (4.97). But before that, after developing the square in this exponential, we need to apply the Hubbard-Stratanovitch transform to $w(\mathbf{x}_{\setminus l}) := \sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^\intercal \mathbf{x}_k$ to simplify the resulting expression, the aim being to linearize all the $\mathbf{x}_l$ independent terms in the exponential so that the integrals become independent:

$$e^{-\frac{w^2 \mathrm{snr}}{2}} = \sqrt{\frac{\mathrm{snr}}{2\pi}} \int d\lambda e^{-\frac{\lambda^2 \mathrm{snr}}{2} + i\mathrm{snr}\lambda w} \tag{4.100}$$

$$\Rightarrow \hat{m}_{\mu l}(\mathbf{x}_l) = \frac{\sqrt{\mathrm{snr}}}{\sqrt{2\pi}\hat{z}_{\mu l}} e^{-\frac{\mathrm{snr}}{2}\left(\mathbf{F}_{\mu l}^\intercal \mathbf{x}_l - y_\mu\right)^2} \int d\lambda e^{-\frac{\lambda^2 \mathrm{snr}}{2}} \prod_{k\neq l}^{L-1} \underbrace{\left[\int d\mathbf{x}_k m_{k\mu}(\mathbf{x}_k) e^{\mathrm{snr}\mathbf{F}_{\mu k}^\intercal \mathbf{x}_k \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)}\right]}_{:=u_k} \tag{4.101}$$

To define the approximate messages Gaussian parametrization, we need the following vectorial objects:

$$\mathbf{a}_\square := \int \mathbf{x}\, m_\square(\mathbf{x})\, d\mathbf{x} \tag{4.102}$$

$$\mathbf{v}_\square := \int \mathbf{x}^2 m_\square(\mathbf{x})\, d\mathbf{x} - \mathbf{a}_\square^2 \tag{4.103}$$

where the square $\bullet^2$ is an elementwise operation as the inverse operation $\bullet^{-1}$ used later on. Expanding in $F_{\mu k}$ the u_k appearing in (4.101) and using the two previous definitions, the integral u_k can be written as:

$$\begin{aligned}
u_k &\approx \int d\mathbf{x}_k m_{k\mu}(\mathbf{x}_k) \left(1 + \mathrm{snr}\mathbf{F}_{\mu k}^\intercal \mathbf{x}_k \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right) + \frac{1}{2}\left[\mathrm{snr}\mathbf{F}_{\mu k}^\intercal \mathbf{x}_k \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)\right]^2\right) \\
&= \left(1 + \mathrm{snr}\mathbf{F}_{\mu k}^\intercal \mathbf{a}_{k\mu} \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right) + \frac{1}{2}(\mathbf{v}_{k\mu} + \mathbf{a}_{k\mu}^2)\left[\mathrm{snr}\mathbf{F}_{\mu k}^\intercal \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)\right]^2\right) \\
&\approx \left(1 + \mathrm{snr}\mathbf{F}_{\mu k}^\intercal \mathbf{a}_{k\mu} \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right) + \frac{\mathbf{a}_{k\mu}^2}{2}\left[\mathrm{snr}\mathbf{F}_{\mu k}^\intercal \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)\right]^2\right) \\
&\quad \left(1 + \frac{\mathbf{v}_{k\mu}}{2}\left[\mathrm{snr}\mathbf{F}_{\mu k}^\intercal \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)\right]^2\right) \\
&\approx e^{\mathbf{a}_{k\mu}^\intercal \mathbf{F}_{\mu k} \mathrm{snr}\left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right) + \frac{\mathrm{snr}^2}{2}\mathbf{v}_{k\mu}^\intercal \mathbf{F}_{\mu k}^2 \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)^2}
\end{aligned} \tag{4.104}$$

where we kept only the terms up to $O(1/L)$. This allows us to write the cavity factor-to-node message as:

$$\Rightarrow \hat{m}_{\mu l}(\mathbf{x}_l) = \frac{\sqrt{\text{snr}}}{\sqrt{2\pi}\hat{z}_{\mu l}} e^{-\frac{\text{snr}}{2}\left(\mathbf{F}_{\mu l}^\intercal \mathbf{x}_l - y_\mu\right)^2} \int d\lambda e^{-\frac{\text{snr}\lambda^2}{2}} \prod_{k\neq l}^{L-1} \left[e^{\mathbf{a}_{k\mu}^\intercal \mathbf{F}_{\mu k} \text{snr}\left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right) + \frac{\text{snr}^2}{2}\mathbf{v}_{k\mu}^\intercal \mathbf{F}_{\mu k}^2 \left(y_\mu - \mathbf{F}_{\mu l}^\intercal \mathbf{x}_l + i\lambda\right)^2} \right] \tag{4.105}$$

The Gaussian integral over λ can now be performed easily, and putting all the $\mathbf{x}_l$ independent terms in the normalization constant $\hat{z}_{\mu l}$ we obtain:

$$\hat{m}_{\mu l}(\mathbf{x}_l) = \frac{1}{\hat{z}_{\mu l}} e^{-\frac{1}{2}\mathbf{A}_{\mu l}^\intercal \mathbf{x}_l^2 + \mathbf{B}_{\mu l}^\intercal \mathbf{x}_l} \tag{4.106}$$

$$\hat{z}_{\mu l} = \prod_{i\in l}^{B} \sqrt{\frac{2\pi}{A_{\mu i}}} e^{\frac{B_{\mu i}^2}{2A_{\mu i}}} \tag{4.107}$$

$$\mathbf{A}_{\mu l} := \frac{\mathbf{F}_{\mu l}^2}{1/\text{snr} + \sum_{k\neq l}^{L-1} \mathbf{v}_{k\mu}^\intercal \mathbf{F}_{\mu k}^2} \tag{4.108}$$

$$\mathbf{B}_{\mu l} := \frac{\mathbf{F}_{\mu l}\left(y_\mu - \sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^\intercal \mathbf{a}_{k\mu}\right)}{1/\text{snr} + \sum_{k\neq l}^{L-1} \mathbf{v}_{k\mu}^\intercal \mathbf{F}_{\mu k}^2} \tag{4.109}$$

We deduce the node-to-factor cavity message expression from (4.94):

$$m_{l\mu}(\mathbf{x}_l) = \frac{1}{z_{l\mu}} P_0^l(\mathbf{x}_l) e^{-\frac{1}{2}(\mathbf{x}_l^2)^\intercal \sum_{\gamma\neq\mu}^{M-1} \mathbf{A}_{\gamma l} + \mathbf{x}_l^\intercal \sum_{\gamma\neq\mu}^{M-1} \mathbf{B}_{\gamma l}} \tag{4.110}$$

$$z_{l\mu} = \int d\mathbf{x}_l P_0^l(\mathbf{x}_l) e^{-\frac{1}{2}(\mathbf{x}_l^2)^\intercal \sum_{\gamma\neq\mu}^{M-1} \mathbf{A}_{\gamma l} + \mathbf{x}_l^\intercal \sum_{\gamma\neq\mu}^{M-1} \mathbf{B}_{\gamma l}} \tag{4.111}$$

where sums of the form $\sum_\gamma \mathbf{\Gamma}_{\gamma l} := \left[\sum_\gamma \Gamma_{\gamma i}\right]_{i\in l}$ are vectors of size B. We now have projected the set of cavity messages onto Gaussian distributions, fully parametrized by their first and second moments.

We define l_i as the B-d section index (or the set of indices, depending on the context) to which the i^{th} 1-d signal component belongs to. We can now define a probability measure over the section l: $m_B((\mathbf{\Sigma}_l)^2, \mathbf{R}_l, \mathbf{x}_l)$ and the corresponding 1-d components marginals, the marginals of the 1-d variables in the section: $\{m_i((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}, x_i)\}_{i\in l}$. We also define the associated vector

of averages $\mathbf{f}_{a_l}$ and variances $\mathbf{f}_{c_l}$ over these marginals:

$$m_B((\mathbf{\Sigma}_l)^2, \mathbf{R}_l, \mathbf{x}_l) := \frac{1}{z((\mathbf{\Sigma}_l)^2, \mathbf{R}_l)} P_0^l(\mathbf{x}_l) e^{-([\mathbf{x}_l - \mathbf{R}_l]^2)^\intercal (2\mathbf{\Sigma}_l^2)^{-1}} \tag{4.112}$$

$$m_i((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}, x_i) := \int d\mathbf{x}_{l_i \setminus i} m_B((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}, \mathbf{x}_{l_i}) \tag{4.113}$$

$$z((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) = z((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) = \int d\mathbf{x}_l P_0^l(\mathbf{x}_l) e^{-([\mathbf{x}_l - \mathbf{R}_l]^2)^\intercal (2\mathbf{\Sigma}_l^2)^{-1}} \tag{4.114}$$

$$f_{a_i}((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) := \int dx_i m_i((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}, x_i) x_i \tag{4.115}$$

$$f_{c_i}((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) := \int dx_i m_i((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}, x_i) x_i^2 - f_{a_i}((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i})^2 \tag{4.116}$$

$$\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) := \left[f_{a_i}((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) \right]_{i \in l} \tag{4.117}$$

$$\mathbf{f}_{c_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) := \left[f_{c_i}((\mathbf{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) \right]_{i \in l} \tag{4.118}$$

Using (4.110) together with these definitions and (4.102), (4.103) we get the second order BP iterations:

$$\mathbf{a}_{l\mu} = \mathbf{f}_{a_l}\left(\frac{1}{\sum_{\gamma \neq \mu}^{M-1} \mathbf{A}_{\gamma l}}, \frac{\sum_{\gamma \neq \mu}^{M-1} \mathbf{B}_{\gamma l}}{\sum_{\gamma \neq \mu}^{M-1} \mathbf{A}_{\gamma l}} \right) \tag{4.119}$$

$$\mathbf{v}_{l\mu} = \mathbf{f}_{c_l}\left(\frac{1}{\sum_{\gamma \neq \mu}^{M-1} \mathbf{A}_{\gamma l}}, \frac{\sum_{\gamma \neq \mu}^{M-1} \mathbf{B}_{\gamma l}}{\sum_{\gamma \neq \mu}^{M-1} \mathbf{A}_{\gamma l}} \right) \tag{4.120}$$

$$\mathbf{a}_l = \mathbf{f}_{a_l}\left(\frac{1}{\sum_{\mu}^{M} \mathbf{A}_{\mu l}}, \frac{\sum_{\mu}^{M} \mathbf{B}_{\mu l}}{\sum_{\mu}^{M} \mathbf{A}_{\mu l}} \right) \tag{4.121}$$

$$\mathbf{v}_l = \mathbf{f}_{c_l}\left(\frac{1}{\sum_{\mu}^{M} \mathbf{A}_{\mu l}}, \frac{\sum_{\mu}^{M} \mathbf{B}_{\mu l}}{\sum_{\mu}^{M} \mathbf{A}_{\mu l}} \right) \tag{4.122}$$

where the two last equations are the marginal mean (4.47) and associated variance, that takes into account all factors. At this stage, after indexing with the time, the algorithm defined by the set of equations (4.108), (4.109) and (4.119), (4.120) together with the definitions (4.113), (4.117) and (4.118) is usually referred as relaxed-BP [34, 35, 94, 96], which is exact for linear estimation as the number of sections $L \to \infty$. After convergence, the final estimates are obtained through (4.121). This first step thus solves the problem of storing the messages, as now each message is parametrized by just two numbers, its mean $\mathbf{a}_{l\mu}$ and variance $\mathbf{v}_{l\mu}$.

Reduction of the number of messages: the TAP equations

We can simplify further the equations, going from an algorithm where $2ML$ messages are exchanged at each time step to one with only $M + L$ messages per time step [97]. The following expansion, the Thouless-Anderson-Palmer approximation in statistical physics of spin glasses [92] is exact in the large signal size limit, as the previous Gaussian parametrization. It starts by noticing that in the $L \to \infty$ limit (and thus the number M of factors diverges as well such that

the measurement rate is constant), the cavity quantities (4.119), (4.120), (4.108) and (4.109) become almost independent of the μ index (which is equivalent to say that each factor's influence becomes infinitely weak as there are so many). We can thus re-write these objects as marginal quantities (that depend on single variable indices) keeping the proper first order correction in $F_{\mu i}$, the Onsager reaction term, essential for the efficiency of the AMP algorithm.

We first define new useful quantities (again, all the operations such as $1/\bullet$ or the dot product $\mathbf{uv}$ applied to vectors are elementwise, as opposed to $\mathbf{u}^\intercal\mathbf{v}$ which is the usual scalar product between vectors):

$$w_\mu := \sum_k^L \mathbf{F}_{\mu k}^\intercal \mathbf{a}_{k\mu} \tag{4.123}$$

$$\Theta_\mu := \sum_k^L (\mathbf{F}_{\mu k}^2)^\intercal \mathbf{v}_{k\mu} \tag{4.124}$$

$$(\mathbf{\Sigma}_k)^2 := \frac{1}{\sum_\mu^M \mathbf{A}_{\mu k}} \tag{4.125}$$

$$\mathbf{R}_k := \frac{\sum_\mu^M \mathbf{B}_{\mu k}}{\sum_\mu^M \mathbf{A}_{\mu k}} \tag{4.126}$$

$$(\mathbf{\Sigma}_{k\mu})^2 := \frac{1}{\sum_{\gamma\neq\mu}^{M-1} \mathbf{A}_{\gamma k}} \tag{4.127}$$

$$\mathbf{R}_{k\mu} := \frac{\sum_{\gamma\neq\mu}^{M-1} \mathbf{B}_{\gamma k}}{\sum_{\gamma\neq\mu}^{M-1} \mathbf{A}_{\gamma k}} \tag{4.128}$$

We always remain in the limit $B \in O(1)$ for the derivation, so if B terms $\in O(1/\sqrt{L})$ are summed, the result is still $\in O(1/\sqrt{L})$. We now expand the cavity quantities (4.119), (4.120) of the relaxed-BP algorithm considering the μ's factor influence weak. Let us start by the cavity averages:

$$\mathbf{a}_{l\mu} = \mathbf{f}_{a_l}((\mathbf{\Sigma}_{l\mu})^2, \mathbf{R}_{l\mu}) \tag{4.129}$$

$$\begin{aligned} &\approx \mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) + ((\mathbf{\Sigma}_{l\mu})^2 - (\mathbf{\Sigma}_l)^2)\nabla_{(\mathbf{\Sigma}_l)^2}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) \\ &\quad + (\mathbf{R}_{l\mu} - \mathbf{R}_l)\nabla_{\mathbf{R}_l}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) \end{aligned} \tag{4.130}$$

$$\begin{aligned} &= \mathbf{a}_l + \left[\frac{\mathbf{A}_{\mu l}}{(\sum_\gamma^M \mathbf{A}_{\gamma l})(\sum_\gamma^M \mathbf{A}_{\gamma l} - \mathbf{A}_{\mu l})}\right]\nabla_{(\mathbf{\Sigma}_l)^2}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) \\ &\quad + \left[\frac{\mathbf{B}_{\mu l}(\sum_\gamma^M \mathbf{A}_{\gamma l}) - \mathbf{A}_{\mu l}(\sum_\gamma^M \mathbf{B}_{\gamma l})}{(\sum_\gamma^M \mathbf{A}_{\gamma l})(\sum_\gamma^M \mathbf{A}_{\gamma l} - \mathbf{A}_{\mu l})}\right]\nabla_{\mathbf{R}_l}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) + o(1/\sqrt{L}) \end{aligned} \tag{4.131}$$

where we have used (4.121), (4.125), (4.126), (4.127) and (4.128). The gradient operator outputs a vector, for example: $\nabla_{\mathbf{R}_l}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l) := \left[\partial_{R_i}\mathbf{f}_{a_l}((\mathbf{\Sigma}_l)^2, \mathbf{R}_l)\right]_{i\in l}$. Now we use the fact that the $\mathbf{A}_{\gamma l} \in O(1/L)$ is a strictly positive term and $\mathbf{B}_{\gamma l} \in O(1/\sqrt{L})$ can be of both signs thus $\sum_\gamma \mathbf{A}_{\gamma l}$ and $\sum_\gamma \mathbf{B}_{\gamma l}$ are both $\in O(1)$. Furthermore, using $(\mathbf{\Sigma}_{l\mu})^2 = (\mathbf{\Sigma}_l)^2 + O(1/L)$ we obtain the first order

corrections to $\mathbf{a}_l$ and $\mathbf{v}_l$ (following the same computation):

$$\mathbf{a}_{l\mu} = \mathbf{a}_l - \underbrace{(\boldsymbol{\Sigma}_l)^2 \mathbf{B}_{\mu l} \nabla_{\mathbf{R}_l} \mathbf{f}_{a_l}((\boldsymbol{\Sigma}_l)^2, \mathbf{R}_l)}_{:=\boldsymbol{\epsilon}_{\mathbf{a}_{l\mu}} \in O(1/\sqrt{L})} + o(1/\sqrt{L}) \tag{4.132}$$

$$\mathbf{v}_{l\mu} = \mathbf{v}_l - \underbrace{(\boldsymbol{\Sigma}_l)^2 \mathbf{B}_{\mu l} \nabla_{\mathbf{R}_l} \mathbf{f}_{c_l}((\boldsymbol{\Sigma}_l)^2, \mathbf{R}_l)}_{:=\boldsymbol{\epsilon}_{\mathbf{v}_{l\mu}} \in O(1/\sqrt{L})} + o(1/\sqrt{L}) \tag{4.133}$$

$\boldsymbol{\epsilon}_{\mathbf{a}_{l\mu}/\mathbf{v}_{l\mu}} := \left[\epsilon_{a_{i\mu}/v_{i\mu}}\right]_{i\in l}$ is the (positive or negative) vector of corrections $\in O(1/\sqrt{L})$ linking $\mathbf{a}_{l\mu}/\mathbf{v}_{l\mu}$ to $\mathbf{a}_l/\mathbf{v}_l$. We need to express all the cavity quantities appearing in these corrections in terms of marginal quantities. In order to do so, we start by expanding (4.125) and (4.126) in the $\mathbf{F}$ elements and keeping only the $O(1)$ terms, we get:

$$\begin{aligned}
(\boldsymbol{\Sigma}_l)^2 &= \left[\sum_\mu^M \frac{\mathbf{F}_{\mu l}^2}{1/\mathrm{snr} + \Theta_\mu - \mathbf{v}_{l\mu}^\intercal \mathbf{F}_{\mu l}^2}\right]^{-1} \\
&\approx \left[\sum_\mu^M \frac{\mathbf{F}_{\mu l}^2}{1/\mathrm{snr} + \Theta_\mu}\right]^{-1}
\end{aligned} \tag{4.134}$$

$$\begin{aligned}
\mathbf{R}_l &= (\boldsymbol{\Sigma}_l)^2 \left[\sum_\mu^M \frac{\mathbf{F}_{\mu l}(y_\mu - w_\mu + \mathbf{F}_{\mu l}\mathbf{a}_{l\mu}^\intercal)}{1/\mathrm{snr} + \Theta_\mu - \mathbf{v}_{l\mu}^\intercal \mathbf{F}_{\mu l}^2}\right] \\
&\approx (\boldsymbol{\Sigma}_l)^2 \left[\sum_\mu^M \frac{\mathbf{F}_{\mu l}\left(y_\mu - w_\mu\right)}{1/\mathrm{snr} + \Theta_\mu} + \underbrace{\sum_\mu^M \mathbf{F}_{\mu l} \frac{(\mathbf{F}_{\mu l})^\intercal \mathbf{a}_l}{1/\mathrm{snr} + \Theta_\mu}}_{\boldsymbol{\Sigma}_l^{-2}\mathbf{a}_l + O(1/\sqrt{L})} - \underbrace{\sum_\mu^M \mathbf{F}_{\mu l} \frac{(\mathbf{F}_{\mu l})^\intercal \boldsymbol{\epsilon}_{\mathbf{a}_{l\mu}}}{1/\mathrm{snr} + \Theta_\mu}}_{\in O(1/L)}\right] \\
&\approx \mathbf{a}_l + (\boldsymbol{\Sigma}_l)^2 \sum_\mu^M \frac{\mathbf{F}_{\mu l}(y_\mu - w_\mu)}{1/\mathrm{snr} + \Theta_\mu}
\end{aligned} \tag{4.135}$$

These depend on the previously defined quantities (4.123) and (4.124) that we thus need to

expand keeping only the terms in $O(1)$ as well to get marginal quantities:

$$\begin{aligned}
\Theta_\mu &= \sum_k^L (\mathbf{F}_{\mu k}^2)^\intercal \mathbf{v}_k - \underbrace{\sum_k^L (\mathbf{F}_{\mu k}^2)^\intercal \boldsymbol{\epsilon}_{\mathbf{v}_{k\mu}}}_{\in O(1/L)} \\
&\approx \sum_k^L (\mathbf{F}_{\mu k}^2)^\intercal \mathbf{v}_k \qquad (4.136) \\
w_\mu &= \sum_k^L \mathbf{F}_{\mu k}^\intercal \mathbf{a}_k - \sum_k^L \mathbf{F}_{\mu k}^\intercal \boldsymbol{\epsilon}_{\mathbf{a}_{k\mu}} \\
&\approx \sum_k^L \mathbf{F}_{\mu k}^\intercal \mathbf{a}_k - \sum_k^L \mathbf{F}_{\mu k}^\intercal \left[\frac{\mathbf{F}_{\mu k}\left(y_\mu - \sum_{k'}^L \mathbf{F}_{\mu k'}^\intercal \mathbf{a}_{k'\mu} + \mathbf{F}_{\mu k}^\intercal \mathbf{a}_{k\mu}\right)}{1/\mathrm{snr} + \sum_{k'}^L \mathbf{v}_{k'\mu}^\intercal \mathbf{F}_{\mu k'}^2 - \mathbf{v}_{k\mu}^\intercal \mathbf{F}_{\mu k}^2} \underbrace{(\boldsymbol{\Sigma}_k)^2 \nabla_{\mathbf{R}_k} \mathbf{f}_{a_k}((\boldsymbol{\Sigma}_k)^2, \mathbf{R}_k)}_{=\mathbf{v}_k} \right] \\
&\approx \sum_k^L \mathbf{F}_{\mu k}^\intercal \mathbf{a}_k - \frac{y_\mu - w_\mu}{1/\mathrm{snr} + \Theta_\mu} \sum_k^L (\mathbf{F}_{\mu k}^2)^\intercal \mathbf{v}_k \qquad (4.137)
\end{aligned}$$

The last equality is obtained using (4.132), (4.109), neglecting $o(1)$ terms and noticing that:

$$\begin{aligned}
\partial_{R_i} f_{a_i}((\boldsymbol{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) &= \partial_{R_i} \left(\frac{\int d\mathbf{x} P_0(\mathbf{x}) x_i \exp\left((\mathbf{x} - \mathbf{R}_{l_i})^\intercal (2(\boldsymbol{\Sigma}_{l_i})^2)^{-1}\right)}{\int d\mathbf{x} P_0(\mathbf{x}) \exp\left((\mathbf{x} - \mathbf{R}_{l_i})^\intercal (2(\boldsymbol{\Sigma}_{l_i})^2)^{-1}\right)} \right) \qquad (4.138) \\
&= \frac{1}{(\Sigma_i)^2} \left(\mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i^2) - R\, \mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i) - \mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i)(\mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i) - R) \right) \qquad (4.139) \\
&= \frac{1}{(\Sigma_i)^2} \left(\mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i^2) - \mathbb{E}_{\mathbf{x}|\mathbf{y}}(x_i)^2 \right) \qquad (4.140) \\
&= \frac{1}{(\Sigma_i)^2} f_{c_i}((\boldsymbol{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) \qquad (4.141) \\
\Rightarrow f_{c_i}((\boldsymbol{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) &= v_i = (\Sigma_i)^2 \partial_{R_i} f_{a_i}((\boldsymbol{\Sigma}_{l_i})^2, \mathbf{R}_{l_i}) \qquad (4.142) \\
\Rightarrow \mathbf{f}_{c_l}((\boldsymbol{\Sigma}_l)^2, \mathbf{R}_l) &= \mathbf{v}_l = (\boldsymbol{\Sigma}_l)^2 \nabla_{\mathbf{R}_l} \mathbf{f}_{a_l}((\boldsymbol{\Sigma}_l)^2, \mathbf{R}_l) \qquad (4.143)
\end{aligned}$$

Adding back the time indices and rewriting the previous set of equations in terms of their single components, we get the AMP algorithm, see Fig. 4.6. Here $\mathbb{E}_{\mathbf{x}|\mathbf{y}}$ is the posterior average estimation by the AMP algorithm. The only problem-dependent objects in the AMP are the denoising functions $\{f_{a_i}, f_{c_i}\}$ that depend on the assumed prior, see the next section. The updating schedule of Fig. 4.6 is what we would have obtained keeping the time index when starting the derivation from the usual parallel BP updates (4.97), (4.98) as we did. On Fig. 4.6, we added a damping scheme controled by the parameter α: at $\alpha = 0$, we recover the derived AMP algorithm but for finite values $0 < \alpha < 1$, the algorithm can converge easier in situations where it experiences some troubles such as numerical oscillations. This scheme is validated empirically, but other ones could be tried. It must be understood that when $\alpha \neq 0$, the algorithm does not follow anymore the state evolution asymptotic predictions that we will derive in sec. 5.3. In all the theoretical analyzes of this thesis, we thus always consider $\alpha = 0$ but in practical situations, a small $\alpha \approx 0.1$ can help.

The AMP algorithm has been generalized to any noise model in [89, 90] and called GAMP

$t \leftarrow 0$
$\delta \leftarrow \epsilon + 1$
while $t < t_{max}$ **and** $\delta > \epsilon$ **do**
  $\tilde{\Theta}_\mu^{t+1} \leftarrow \sum_i^N F_{\mu i}^2 v_i^t$
  $w_\mu^{t+1} \leftarrow \alpha w_\mu^t + (1-\alpha)\left(\sum_i^N F_{\mu i} a_i^t - \tilde{\Theta}_\mu^{t+1} \frac{y_\mu - w_\mu^t}{1/\mathrm{snr}+\Theta_\mu^t}\right)$
  $\Theta_\mu^{t+1} \leftarrow \alpha \Theta_\mu^t + (1-\alpha)\tilde{\Theta}_\mu^{t+1}$
  $\Sigma_i^{t+1} \leftarrow \left[\sum_\mu^M \frac{F_{\mu i}^2}{1/\mathrm{snr}+\Theta_\mu^{t+1}}\right]^{-1/2}$
  $R_i^{t+1} \leftarrow a_i^t + (\Sigma_i^{t+1})^2 \sum_\mu^M F_{\mu i} \frac{y_\mu - w_\mu^{t+1}}{1/\mathrm{snr}+\Theta_\mu^{t+1}}$
  $v_i^{t+1} \leftarrow f_{c_i}\left((\mathbf{\Sigma}_{l_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1}\right)$
  $a_i^{t+1} \leftarrow f_{a_i}\left((\mathbf{\Sigma}_{l_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1}\right)$
  $t \leftarrow t+1$
  $\delta \leftarrow ||\mathbf{a}^{t+1} - \mathbf{a}^t||_2^2$
end while
return $\mathbf{a}^t$

Figure 4.6 – The approximate message-passing algorithm with damping controled by $0 \leq \alpha < 1$. l_i is the index of the section to which the i^{th} 1-d variable belongs to. In the scalar components case $B = 1$, the sections are just the components. ϵ is the accuracy for convergence and t_{max} the maximum number of iterations. A suitable initialization for the quantities is ($a_i^{t=0} = \mathbb{E}_{P_0}(x_i)$, $v_i^{t=0} = \mathrm{Var}_{P_0}(x_i)$, $w_\mu^{t=0} = y_\mu$). Once the algorithm has converged, i.e. the quantities do not change anymore from iteration to iteration, the estimate of the l^{th} signal section is $\mathbf{a}_l^t$. At $\alpha = 0$, the usual approximate message-passing algorithm is recovered. $\{\tilde{\Theta}_\mu\}_\mu^M$ are auxilliary variables for the damping scheme.

for "generalized AMP". The algorithm presented here is equivalent to GAMP for the i.i.d AWGN channel. It is important to note that the name "approximate" message-passing is a little misleading since as we have shown through the derivation, for dense i.i.d random measurement matrices the AMP is asymptotically equivalent to BP, i.e. all the leading terms in N are included in AMP.

4.3.4 Alternative "simpler" derivation of the approximate message-passing algorithm

The previous derivation did not assume anything and proves the asymptotic equivalence between AMP and BP but is a bit long. We present here an alternative derivation of the AMP algorithm, which will use directly the assumption that the node-to-factor messages can be represented as Gaussians. This derivation is very close to the one of the non-parametric BP algorithm [98] in the special case where only one Gaussian in used in the messages parametrization. Like the previous derivation, it starts from the factor-to-node cavity message (4.97). We first define $\mathbf{\Gamma}_{\mu l} := \sum_{k \neq l}^{L-1} \mathbf{F}_{\mu k}^\intercal \mathbf{x}_k$, the variable that appears in the exponential in (4.97). Now we assume that the cavity node-to-factor messages are Gaussians. In a sense we place ourselves

directly at the step (4.110) of the previous derivation:

$$m^t_{k\mu}(\mathbf{x}_k) = \mathcal{N}\left(\mathbf{x}_k \middle| \mathbf{a}^t_{k\mu}, \mathbf{v}^t_{k\mu}\right) \tag{4.144}$$

As discussed in sec. 4.2.2, when computing the factor-to-node messages we assume that the node-to-factor messages arriving on the factor of interest are conditionally independent (due to the tree approximation behind BP). This implies that $\mathbf{\Gamma}^t_{\mu l} \sim \mathcal{N}\left(\mathbf{\Gamma}^t_{\mu l} \middle| \bar{\mathbf{\Gamma}}^t_{\mu l}, \mathbf{v}^t_{\mathbf{\Gamma}_{\mu l}}\right)$ in (4.97) is also Gaussian distributed with moments given by:

$$\bar{\mathbf{\Gamma}}^t_{\mu l} = \sum_{k \neq l} \mathbf{F}^\intercal_{\mu k} \mathbf{a}^t_{k\mu} \tag{4.145}$$

$$\mathbf{v}^t_{\mathbf{\Gamma}_{\mu l}} = \sum_{k \neq l} (\mathbf{F}^2_{\mu k})^\intercal \mathbf{v}^t_{k\mu} \tag{4.146}$$

We thus obtain the factor-to-node message $\hat{m}^{t+1}_{\mu l}(\mathbf{x}_l)$ expression from (4.97):

$$\hat{m}^t_{\mu l}(\mathbf{x}_l) = \frac{1}{\hat{z}_{\mu l}\, t} \int \mathcal{N}\left(\mathbf{\Gamma}_{\mu l} \middle| y_\mu - \mathbf{F}^\intercal_{\mu l} \mathbf{x}_l, 1/\mathrm{snr}\right) \mathcal{N}\left(\mathbf{\Gamma}_{\mu l} \middle| \bar{\mathbf{\Gamma}}^t_{\mu l}, \mathbf{v}^t_{\mathbf{\Gamma}_{\mu l}}\right) d\mathbf{\Gamma}_{\mu l} \tag{4.147}$$

$$= \mathcal{N}\left(\mathbf{x}_l \middle| \mathbf{a}^t_{\mu l}, \mathbf{v}^t_{\mu l}\right) \tag{4.148}$$

where the first Gaussian is the likelihood part. The moments are given by:

$$\mathbf{a}^t_{\mu l} = \frac{y_\mu - \sum_{k \neq l} \mathbf{F}^\intercal_{\mu k} \mathbf{a}^t_{k\mu}}{\mathbf{F}_{\mu l}} = \frac{\mathbf{B}^t_{\mu l}}{\mathbf{A}^t_{\mu l}} \tag{4.149}$$

$$\mathbf{v}^t_{\mu l} = \frac{1/\mathrm{snr} + \sum_{k \neq l} (\mathbf{F}^2_{\mu k})^\intercal \mathbf{v}^t_{k\mu}}{\mathbf{F}^2_{\mu l}} = \frac{1}{\mathbf{A}^t_{\mu l}} \tag{4.150}$$

where we recognized the expressions (4.108), (4.109) obtained in the previous derivation, thus it is coherent. Now in order to compute the node-to-factor message from (4.98), we need the fact that a product of K Gaussians distributions over the same variable with respective means and variances $\{r_k, v_k\}_k^K$ gives a new Gaussian distribution with moments $\{r, v\}$ given by:

$$v = \left(\sum_k^K v_k^{-1}\right)^{-1} \tag{4.151}$$

$$r = v \sum_k^K r_k v_k^{-1} \tag{4.152}$$

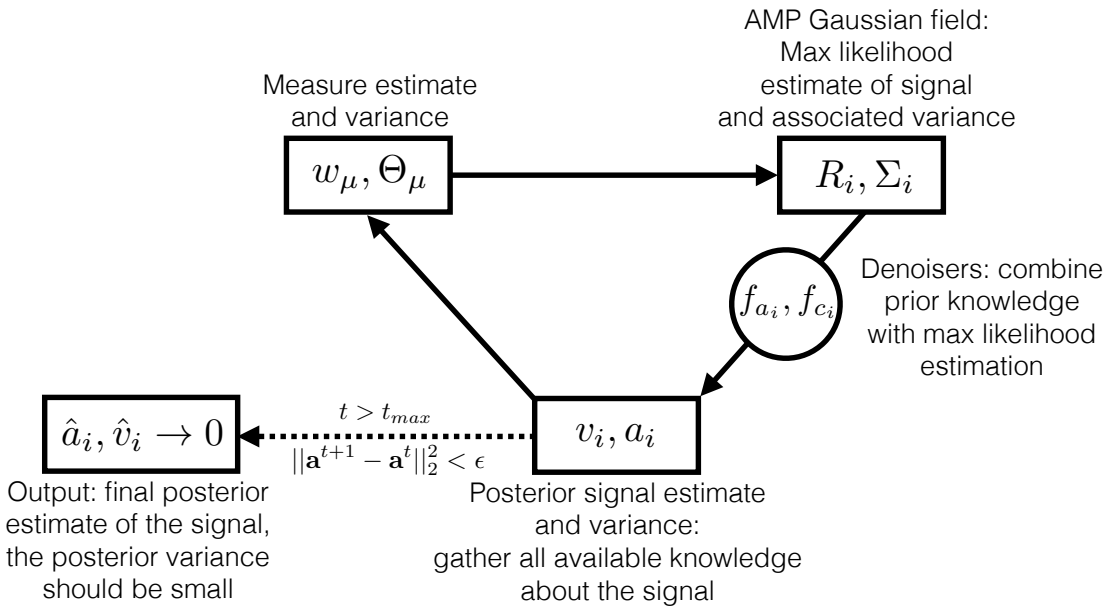

Figure 4.7 – Graphical representation of the approximate message-passing algorithm. It iteratively repeats three main steps until convergence: i) computation of an estimate of the measurement ans its variance, ii) computation of the maximum likelihood estimate of the signal and its variance, the AMP fields and then iii) combine the previous maximum likelihood estimate with the prior through the denoisers to get the posterior estimate and variance.

From this combined with (4.98), we close the equations on the cavity node-to-factor means and variances:

$$m_{l\mu}^{t+1}(\mathbf{x}_l) = \frac{1}{z_{l\mu}^{t+1}} P_0^l(\mathbf{x}_l) \prod_{\nu\neq\mu} \hat{m}_{\nu l}^t(\mathbf{x}_l) \tag{4.153}$$

$$= \frac{1}{z_{l\mu}^{t+1}} P_0^l(\mathbf{x}_l) \mathcal{N}(\mathbf{x}_l | \tilde{\mathbf{a}}_{l\mu}^{t+1}, \tilde{\mathbf{v}}_{l\mu}^{t+1}) \tag{4.154}$$

$$\tilde{\mathbf{v}}_{l\mu}^{t+1} = \left[\sum_{\nu\neq\mu} \left(\frac{\mathbf{F}_{\nu l}^2}{1/\mathrm{snr} + \sum_{k\neq l} (\mathbf{F}_{\nu k}^2)^\intercal \mathbf{v}_{k\nu}^t} \right) \right]^{-1} = \frac{1}{\sum_{\nu\neq\mu} \mathbf{A}_{\nu l}^t} \tag{4.155}$$

$$\tilde{\mathbf{a}}_{l\mu}^{t+1} = \mathbf{v}_{l\mu}^{t+1} \sum_{\nu\neq\mu} \frac{\mathbf{F}_{\nu l} \left(y_\mu - \sum_{k\neq l} \mathbf{F}_{\nu k}^\intercal \mathbf{a}_{k\nu}^t \right)}{1/\mathrm{snr} + \sum_{k\neq l} (\mathbf{F}_{\nu k}^2)^\intercal \mathbf{v}_{k\nu}^t} = \frac{\sum_{\nu\neq\mu} \mathbf{B}_{\nu l}^t}{\sum_{\nu\neq\mu} \mathbf{A}_{\nu l}^t} \tag{4.156}$$

using again (4.108), (4.109). The last step is to compute the mean $\mathbf{a}_{l\mu}^{t+1}$ and variance $\mathbf{v}_{l\mu}^{t+1}$ of the cavity message (4.154): it gives exactly the relaxed-BP equations (4.119), (4.120) and thus the AMP algorithm after the TAP step like in the previous section.

4.3.5 Understanding the approximate message-passing algorithm

Let us explain what AMP is doing. It is an iterative algorithm that iteratively repeat three basic steps, each time computing an estimate of some quantity and its associated variance. This is represented on Fig. 4.7:

First step : From the actual knowledge of the posterior signal estimate and variance, AMP computes the estimate of the measurement vector and its variance. This estimate $\mathbf{w}^{t+1}$ (see Fig. 4.6) is a sum of two terms: the first one is the measure one would get if the true signal was given by the actual posterior estimate $\mathbf{a}^t$, the second one is a "gradient like" term that tends to *increase* the difference between the measure and its new estimate. This increase is proportional to the measurement variance and the difference. It can seem a priori strange to amplify this difference, but it is actually understandable from the use of this estimate in the second step.

Second step : AMP computes what we call the AMP fields: $\mathbf{R}^{t+1}$ and associated variances $(\mathbf{\Sigma}^{t+1})^2$, the means and variances of a Gaussian mean field on each variable that summarizes the influence of all the likelihood factors, i.e. this Gaussian field tends to maximize the likelihood of the signal by enforcing its estimate to match the measurements. $\mathbf{R}^{t+1}$ (see Fig. 4.6) is the sum of the previous posterior estimate, the best estimate at the previous time step plus another "gradient like" term which is proportional to the AMP field variance and the most recent difference between the measure estimate and the true measurement vector: now we understand that the amplification of this difference in the previous step actually leads to a stronger shift of the AMP fields $\mathbf{R}^{t+1}$ in the proper direction to reduce the difference with the measurement vector, thus to increase the likelihood.

Third step : The last step takes as input the new AMP fields and combine them with the prior distribution of the signal to get the new posterior estimate that gather all the actual information about the signal. This is performed thanks to the denoiser f_a: this function averages over all the possible signal estimates properly weighted by their actual posterior distribution, given by the product of the AMP Gaussian field and the prior. f_c computes the associated posterior variance that should converge to 0 as the algorithm converges to its fixed point, hopefully given by the true posterior $MMSE$ estimate (under the prior matching condition, and above the BP transition if no spatial coupling is used, see sec. 5.1.1 and sec. 5.5).

These steps are repeated until convergence or the maximum number of iterations is reached, and the estimator is the last posterior estimate of the signal.

4.3.6 How to quickly cook an approximate message-passing algorithm for your linear estimation problem

In the generic AMP algorithm, the only problem dependent part are the denoising functions $\{f_{a_i}, f_{c_i}\}$, but once adapted the algorithm Fig. 4.6 can be applied to a large class of linear estimation problems. We give here "blocks" for constructing such denoising functions. For a

factorizable prior $P_0(\mathbf{x}) = \prod_i^N P_0^i(x_i)$, we need the posterior partition function $z(\Sigma^2, R)$, the first $u(\Sigma^2, R)$ and second $v(\Sigma^2, R)$ non normalized moments. We consider that the prior $P_0^i(x)$ is a linear combination of different distributions $p_u(x)$:

$$P_0^i(x) \propto \sum_u w_u p_u(x) \tag{4.157}$$

where w_u is the weight of the distribution p_u in the prior. These weights dont need to be normalized in the present construction. From this we can construct the posterior normalization:

$$\begin{aligned} z(\Sigma^2, R) &:= \int dx\, P_0^i(x)\, \mathcal{N}(x|R, \Sigma^2) \\ &= \sum_u w_u \int dx\, p_u(x)\, \mathcal{N}(x|R, \Sigma^2) \\ &= \sum_u w_u z_u(\Sigma^2, R) \end{aligned} \tag{4.158}$$

In the same way we construct the first and second non normalized moments:

$$\begin{aligned} \gamma(\Sigma^2, R) &:= \int dx\, P_0^i(x)\, \mathcal{N}(x|R, \Sigma^2) x \\ &= \sum_u w_u \int dx\, p_u(x)\, \mathcal{N}(x|R, \Sigma^2) x \\ &= \sum_u w_u \gamma_u(\Sigma^2, R) \end{aligned} \tag{4.159}$$

$$\begin{aligned} \tau(\Sigma^2, R) &:= \int dx\, P_0^i(x)\, \mathcal{N}(x|R, \Sigma^2) x^2 \\ &= \sum_u w_u \int dx\, p_u(x)\, \mathcal{N}(x|R, \Sigma^2) x^2 \\ &= \sum_u w_u \tau_u(\Sigma^2, R) \end{aligned} \tag{4.160}$$

From this contruction, we define the denoisers as:

$$f_{a_i}(\Sigma^2, R) := \frac{\gamma(\Sigma^2, R)}{z(\Sigma^2, R)} \tag{4.161}$$

$$f_{c_i}(\Sigma^2, R) := \frac{\tau(\Sigma^2, R)}{z(\Sigma^2, R)} - f_a(\Sigma^2, R)^2 \tag{4.162}$$

see the Tab. 4.1 for possible triplets (z_u, γ_u, τ_u) to construct denoisers. Numerically, it is careful to take for the variance denoiser $f_{c_i}(\Sigma^2, R) = \max\left(f_{c_i}(\Sigma^2, R), \epsilon\right)$ where ϵ is a very small constant, like 10^{-20}. This avoids possible negative variances that can appear at the very beginning of the convergence.

prior term $p_u(x_i)$	$z_u(\Sigma^2, R)$
Dirac $\delta(x_i - m)$	$\mathcal{N}\left(m\|R, \Sigma^2\right)$
Gauss $\mathcal{N}(x_i\|m, v)$	$\mathcal{N}\left(m\|R, \Sigma^2 + v\right)$
Exponential $\lambda e^{-\lambda x_i}\mathbb{I}(x_i > 0)$	$\frac{\lambda}{2}\exp\left(\frac{\lambda}{2}(\lambda\Sigma^2 - 2R)\right)\mathrm{erfc}\left[\frac{\lambda\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right]$
Laplace $\frac{\beta}{2}e^{-\beta\|x_i\|}$	$\frac{\beta e^{\beta^2\Sigma^2/2}}{4}\left(e^{-\beta R}\mathrm{erfc}\left(\frac{\beta\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right) + e^{\beta R}\mathrm{erfc}\left(\frac{\beta\Sigma^2 + R}{\sqrt{2\Sigma^2}}\right)\right)$

prior term $p_u(x_i)$	$\gamma_u(\Sigma^2, R)$
Dirac $\delta(x_i - m)$	$m\mathcal{N}\left(m\|R, \Sigma^2\right)$
Gauss $\mathcal{N}(x_i\|m, v)$	$\mathcal{N}\left(m\|R, \Sigma^2 + v\right)\frac{m\Sigma^2 + Rv}{\Sigma^2 + v}$
Exponential $\lambda e^{-\lambda x_i}\mathbb{I}(x_i > 0)$	$\frac{\lambda e^{-R^2/(2\Sigma^2)}}{2\sqrt{\pi}}\left(\sqrt{2\Sigma^2} + \sqrt{\pi}(R - \lambda\Sigma^2)e^{(R-\lambda\Sigma^2)^2/(2\Sigma^2)}\mathrm{erfc}\left[\frac{\lambda\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right]\right)$
Laplace $\frac{\beta}{2}e^{-\beta\|x_i\|}$	$\frac{\beta e^{\beta^2\Sigma^2/2}}{4}\Big(e^{-\beta R}(R - \beta\Sigma^2)\mathrm{erfc}\left(\frac{\beta\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right)$ $+e^{\beta R}(R + \beta\Sigma^2)\mathrm{erfc}\left(\frac{\beta\Sigma^2 + R}{\sqrt{2\Sigma^2}}\right)\Big)$

prior term $p_u(x_i)$	$\tau_u(\Sigma^2, R)$
Dirac $\delta(x_i - m)$	$m^2\mathcal{N}\left(m\|R, \Sigma^2\right)$
Gauss $\mathcal{N}(x_i\|m, v)$	$\mathcal{N}\left(m\|R, \Sigma^2 + v\right)\frac{m^2\Sigma^4 + \Sigma^2(2mR + \Sigma^2)v + (R^2 + \Sigma^2)v^2}{(\Sigma^2 + v)^2}$
Exponential $\lambda e^{-\lambda x_i}\mathbb{I}(x_i > 0)$	$\frac{\lambda e^{-R^2/(2\Sigma^2)}}{2\sqrt{2\pi}}\big\{2\Sigma(R - \lambda\Sigma^2)$ $+\sqrt{2\pi}\left((R - \lambda\Sigma^2)^2 + \Sigma^2\right)e^{(R-\lambda\Sigma^2)^2/(2\Sigma^2)}\mathrm{erfc}\left[\frac{\lambda\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right]\big\}$
Laplace $\frac{\beta}{2}e^{-\beta\|x_i\|}$	$\frac{\beta e^{\beta^2\Sigma^2/2}}{4}\Big(-4\beta\Sigma^3 e^{-R^2/(2\Sigma^2)} + e^{-\beta R}(\Sigma^2 + (\beta\Sigma^2 - R)^2)\mathrm{erfc}\left(\frac{\beta\Sigma^2 - R}{\sqrt{2\Sigma^2}}\right)$ $+e^{\beta R}(\Sigma^2 + (\beta\Sigma^2 + R)^2)\mathrm{erfc}\left(\frac{\beta\Sigma^2 + R}{\sqrt{2\Sigma^2}}\right)\Big)$

Table 4.1 – Examples of functions for the prior construction

4.3.7 The Bethe free energy for large dense graphs with i.i.d additive white Gaussian noise

For completeness, we now show how to derive an expression of the Bethe free energy that depends on the quantities appearing in the AMP algorithm. The resulting expression is only valid at the fixed point of the algorithm and is true asymptotically on dense graphs as the derivation uses the same assumptions that for AMP, see sec. 4.3.3. We consider that the measurement matrix is homogeneous.

The factors contributions

We start from the free energy expression (4.83). Let us start by computing the term (4.84). Using the likelihood of the compressed sensing with AWGN (3.58), we get:

$$z_\mu = \int d\mathbf{x} \prod_i^N m_{i\mu}(x_i) \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{(y_\mu - \sum_i^N F_{\mu i} x_i)^2}{2\Delta} \right) \tag{4.163}$$

which is almost (up to one additional variable) the (scalar) normalization of (4.97) at the fixed point. Thus the "clean" derivation is exactly the same as in sec. 4.3.3: i) apply the Stratanovitch transformation to linearize the squared sum that appears in the exponent. In this way there are no crossed terms between variables and the integrals can be perfomed independently. ii) Expand the exponential up to second order using the fact the the **F** elements are small, iii) integrate the independent integrals and use the same trick as (4.104) to get a similar expression, except that the signal dependent term inside the parenthesis are not present anymore as all variables have been integrated and finally iv) perform the Gaussian integral over λ, the auxiliary parameter introduced for the Stratanovitch transform and simplify the result.

Now the fastest way is to remind ourselves that the Bethe approximation is equivalent to assume the tree property of the graph and thus as discussed in sec. 4.2.2, when a cavity is dug in the graph by removing a factor, the neighboring variables (i.e. all of them in the present dense case) are considered independent. Thus the central limit theorem implies that $\gamma := \sum_i^N F_{\mu i} x_i$ appearing in the exponent in (4.163) is a Gaussian random variable with mean $w_\mu := \sum_i^N F_{\mu i} a_{i\mu}$ and variance $\Theta_\mu := \sum_i^N F_{\mu i}^2 v_{i\mu}$, the same as (4.123), (4.124) in the scalar case where $a_{i\mu}$ and $v_{i\mu}$ are the cavity mean and variance associated to $m_{i\mu}(x_i)$ and computed respectively through (4.119) and (4.120) in the AMP framework. Finally the result is:

$$z_\mu = \int d\gamma \mathcal{N}(\gamma | y_\mu, \Delta) \mathcal{N}(\gamma | w_\mu, \Theta_\mu) \tag{4.164}$$

$$= \mathcal{N}\left(w_\mu | y_\mu, \Delta + \Theta_\mu\right) \tag{4.165}$$

The nodes and edges contributions

We now compute (4.85). From (4.106) that we got in the AMP derivation and as each variable is connected to each likelihood factor in the homogeneous measurement matrix case that we consider here, we have:

$$z_i = \left[\prod_\mu^M \hat{z}_{\mu i} \right]^{-1} \int dx_i P_0(x_i) \exp\left(-\frac{x_i^2}{2} \sum_\mu^M A_{\mu i} + x_i \sum_\mu^M B_{\mu i} \right) \tag{4.166}$$

$$= \left[\prod_\mu^M \hat{z}_{\mu i} \right]^{-1} \tilde{z}_i \tag{4.167}$$

where the prior factor must not be forgotten in (4.85) and we define:

$$\tilde{z}_i := \int dx_i P_0(x_i) \exp\left(-\frac{x_i^2}{2\Sigma_i^2} + x_i \frac{R_i}{\Sigma_i^2}\right) \tag{4.168}$$

where we have used the definition of the AMP fields (4.126), (4.125). Now we notice that using (4.86) with (4.46), we get at the fixed point that:

$$\tilde{z}_{i\mu} = \frac{\tilde{z}_i}{z_{i\mu}\hat{z}_{\mu i}} \tag{4.169}$$

It allows with (4.167) to re-write the nodes and edges contributions to the free energy (4.83) as:

$$-\sum_i^N \log(z_i) + \sum_i^N \sum_\mu^M \log\left(\tilde{z}_{i\mu}\right) \tag{4.170}$$

$$= -\sum_i^N \left(\log(\tilde{z}_i) - \sum_\mu^M \log\left(\hat{z}_{\mu i}\right) - \sum_\mu^M \left(\log\left(\frac{\tilde{z}_i}{z_{i\mu}}\right) - \log\left(\hat{z}_{\mu i}\right)\right)\right) \tag{4.171}$$

$$= -\sum_i^N \left(\log(\tilde{z}_i) + \sum_\mu^M \log\left(\frac{z_{i\mu}}{\tilde{z}_i}\right)\right) \tag{4.172}$$

It is easy to verify from (4.126), (4.125), (4.128), (4.127) that at first order:

$$\frac{1}{\Sigma_{i\mu}^2} := \sum_{\gamma\neq\mu}^M A_{\gamma i} \tag{4.173}$$

$$\approx \frac{1}{\Sigma_i^2}\left(1 - \Sigma_i^2 A_{\mu i}\right) \tag{4.174}$$

$$\Rightarrow R_{i\mu} := \frac{\sum_{\gamma\neq\mu}^M B_{\gamma i}}{\sum_{\gamma\neq\mu}^M A_{\gamma i}} \tag{4.175}$$

$$= R_i\left(1 + \Sigma_i^2 A_{\mu i}\right) - \Sigma_i^2 B_{\mu i} \tag{4.176}$$

which implies also at first order:

$$\frac{R_{i\mu}}{\Sigma_{i\mu}^2} \approx \frac{R_i}{\Sigma_i^2} - B_{\mu i} \tag{4.177}$$

reminding that $A_{\mu i} \in O(1/N)$ and $B_{\mu i} \in O(1/\sqrt{N})$. From these results, we can simplify the node-to-factor partition function (4.111):

$$z_{i\mu} = \int dx_i P_0(x_i) \exp\left(-\frac{x_i^2}{2\Sigma_{i\mu}^2} + x_i \frac{R_{i\mu}}{\Sigma_{i\mu}^2}\right) \tag{4.178}$$

$$\approx \int dx_i P_0(x_i) \exp\left(-\frac{x_i^2}{2\Sigma_i^2} + x_i \frac{R_i}{\Sigma_i^2}\right)\left(1 - x_i B_{\mu i} + \frac{x_i^2}{2}\left(B_{\mu i}^2 + A_{\mu i}\right)\right) \tag{4.179}$$

$$= \tilde{z}_i + \int dx_i P_0(x_i) \exp\left(-\frac{x_i^2}{2\Sigma_i^2} + x_i \frac{R_i}{\Sigma_i^2}\right)\left(-x_i B_{\mu i} + \frac{x_i^2}{2}\left(B_{\mu i}^2 + A_{\mu i}\right)\right) \tag{4.180}$$

The last equality allows to simplify the second sum appearing in (4.172):

$$\sum_\mu^M \log\left(\frac{z_{i\mu}}{\tilde{z}_i}\right) = \sum_\mu^M \log\left(1 + \frac{1}{\tilde{z}_i}\int dx_i P_0(x_i) e^{-\frac{x_i^2}{2\Sigma_i^2} + x_i\frac{R_i}{\Sigma_i^2}}\left(-x_i B_{\mu i} + \frac{x_i^2}{2}\left(B_{\mu i}^2 + A_{\mu i}\right)\right)\right) \tag{4.181}$$

$$\approx \frac{1}{\tilde{z}_i}\int dx_i P_0(x_i) e^{-\frac{x_i^2}{2\Sigma_i^2} + x_i\frac{R_i}{\Sigma_i^2}} \sum_\mu^M \left(-x_i B_{\mu i} + \frac{x_i^2}{2}\left(B_{\mu i}^2 + A_{\mu i}\right)\right)$$
$$- \frac{1}{2}\sum_\mu^M B_{\mu i}^2 \Bigg(\underbrace{\frac{1}{\tilde{z}_i}\int dx_i P_0(x_i) e^{-\frac{x_i^2}{2\Sigma_i^2} + x_i\frac{R_i}{\Sigma_i^2}} x_i}_{=a_i}\Bigg)^2 \tag{4.182}$$

$$\approx \frac{1}{\tilde{z}_i}\int dx_i P_0(x_i) e^{-\frac{x_i^2}{2\Sigma_i^2} + x_i\frac{R_i}{\Sigma_i^2}} \left(-x_i \frac{R_i}{\Sigma_i^2} + \frac{x_i^2}{2}\left(\sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} + \frac{1}{\Sigma_i^2}\right)\right)$$
$$- \frac{a_i^2}{2}\sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} \tag{4.183}$$

$$= -a_i \frac{R_i}{\Sigma_i^2} + \frac{v_i + a_i^2}{2}\left(\sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} + \frac{1}{\Sigma_i^2}\right) - \frac{a_i^2}{2}\sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} \tag{4.184}$$

$$= -a_i \frac{R_i}{\Sigma_i^2} + \frac{v_i + a_i^2}{2\Sigma_i^2} + \frac{v_i}{2}\left(\sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2}\right) \tag{4.185}$$

where we have used (4.124), (4.123) inside (4.109) which gives at first order:

$$\sum_\mu^M B_{\mu i}^2 \approx \sum_\mu^M \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} \tag{4.186}$$

and where we have recognized the expressions of the marginal posterior mean a_i and variance v_i ((4.102), (4.103), (4.112)). Finally, as we want to express everything in terms of quantities computed by the AMP algorithm, we notice that we can re-write $\tilde{z}_i$ in terms of the AMP variable partition function (4.114) that we call here $z_i^{AMP} = \tilde{z}_i \exp\left(-R_i^2/(2\Sigma_i^2)\right)$. Finally the free

energy reads:

$$F \approx -\sum_{\mu}^{M} \log(z_\mu) - \sum_{i}^{N} \left[\log\left(z_i^{AMP}\right) + \frac{(R_i - a_i)^2 + v_i}{2\Sigma_i^2} + \frac{v_i}{2} \sum_{\mu}^{M} \frac{F_{\mu i}^2 (y_\mu - w_\mu)^2}{(\Delta + \Theta_\mu)^2} \right] \tag{4.187}$$

$$\approx -\sum_{\mu}^{M} \log(z_\mu) - \sum_{i}^{N} \left[\log\left(z_i^{AMP}\right) + \frac{(R_i - a_i)^2 + v_i}{2\Sigma_i^2} \right] - \sum_{\mu}^{M} \Theta_\mu \frac{(y_\mu - w_\mu)^2}{2(\Delta + \Theta_\mu)^2} \tag{4.188}$$

using (4.136) for the second equality. All the approximations are again exact in the large dense graph limit. Now we notice from Fig. 4.6 that the fixed point conditions are:

$$\Theta_\mu \frac{y_\mu - w_\mu}{\Delta + \Theta_\mu} = \sum_{i}^{N} F_{\mu i} a_i - w_\mu \tag{4.189}$$

$$\Theta_\mu = \sum_{i}^{N} F_{\mu i}^2 v_i \tag{4.190}$$

$$a_i = f_{a_i}(\Sigma_i^2, R_i) \tag{4.191}$$

$$v_i = f_{c_i}(\Sigma_i^2, R_i) \tag{4.192}$$

Using (4.189), we get the final expression of the Bethe free energy in terms of the AMP quantities *at their fixed point*:

$$F \approx -\sum_{\mu}^{M} \left[\log(z_\mu) + \frac{\left(\sum_i^N F_{\mu i} a_i - w_\mu\right)^2}{2\Theta_\mu} \right] - \sum_{i}^{N} \underbrace{\left[\log\left(z_i^{AMP}\right) + \frac{(R_i - a_i)^2 + v_i}{2\Sigma_i^2} \right]}_{= -KL(P_i || P_0)} \tag{4.193}$$

$$= \frac{M}{2} \log(2\pi\Delta) + \sum_{\mu}^{M} \left[\frac{\left(y_\mu - w_\mu\right)^2}{2\left(\Delta + \Theta_\mu\right)} + \frac{1}{2} \log\left(1 + \frac{\Theta_\mu}{\Delta}\right) - \frac{\left(\sum_i^N F_{\mu i} a_i - w_\mu\right)^2}{2\Theta_\mu} \right]$$

$$+ \sum_{i}^{N} KL(P_i || P_0) \tag{4.194}$$

where we have replaced (4.165) in the last equality and:

$$P_i(x_i | \Sigma_i^2, R_i) := \frac{1}{\tilde{z}_i} P_0(x_i) \exp\left(-(R_i - x_i)^2 / (2\Sigma_i^2)\right) \tag{4.195}$$

is as usual the AMP posterior measure (4.112) of the variable x_i and we use the Kullback-Leibler divergence (3.49). Now let us simplify even further this expression by re-writing (4.189) which is true at the fixed point as:

$$w_\mu = \frac{1}{\Delta} \left(\tilde{a}_\mu (\Delta + \Theta_\mu) - \Theta_\mu y_\mu \right) \tag{4.196}$$

where we define $\tilde{a}_\mu := \sum_i^N F_{\mu i} a_i$. We want to simplify (4.194) by working out the term:

$$\frac{\left(y_\mu - w_\mu\right)^2}{2\left(\Delta + \Theta_\mu\right)} - \frac{\left(\tilde{a}_\mu - w_\mu\right)^2}{2\Theta_\mu} \tag{4.197}$$

$$= \frac{y_\mu^2}{2(\Delta+\Theta_\mu)} - \frac{\tilde{a}_\mu^2}{2\Theta_\mu} + w_\mu\left(\frac{\tilde{a}_\mu}{\Theta_\mu} - \frac{y_\mu}{\Delta+\Theta_\mu}\right) - \frac{w_\mu^2\Delta}{2\Theta_\mu(\Delta+\Theta_\mu)} \tag{4.198}$$

Now replacing w_μ by (4.196), we get after careful simplifications:

$$w_\mu\left(\frac{\tilde{a}_\mu}{\Theta_\mu} - \frac{y_\mu}{\Delta+\Theta_\mu}\right) = \frac{1}{\Delta}\left(\tilde{a}_\mu^2\frac{\Delta+\Theta_\mu}{\Theta_\mu} - 2y_\mu\tilde{a}_\mu + \frac{y_\mu^2\Theta_\mu}{\Delta+\Theta_\mu}\right) \tag{4.199}$$

$$-\frac{w_\mu^2\Delta}{2\Theta_\mu(\Delta+\Theta_\mu)} = -\frac{1}{2\Delta}\left(\tilde{a}_\mu^2\frac{\Delta+\Theta_\mu}{\Theta_\mu} - 2y_\mu\tilde{a}_\mu + \frac{y_\mu^2\Theta_\mu}{\Delta+\Theta_\mu}\right) \tag{4.200}$$

Combining everything in (4.198), we get after simplification:

$$\frac{\left(y_\mu - w_\mu\right)^2}{2\left(\Delta + \Theta_\mu\right)} - \frac{\left(\tilde{a}_\mu - w_\mu\right)^2}{2\Theta_\mu} = \frac{1}{2\Delta}\left(y_\mu - \tilde{a}_\mu\right)^2 \tag{4.201}$$

which allows to simplify the Bethe free energy (4.194) using the fixed point condition (4.190):

$$F(\{\Sigma_i^2, R_i\}_i^N) = \frac{1}{2}\sum_\mu^M\left[\frac{(y_\mu - \sum_i^N F_{\mu i}a_i)^2}{\Delta} + \log\left(1 + \frac{\sum_i^N F_{\mu i}^2 v_i}{\Delta}\right)\right] + \frac{M}{2}\log(2\pi\Delta) + \sum_i^N KL(P_i||P_0) \tag{4.202}$$

which is only true at the fixed point of the algorithm, i.e. when all the constraints (4.189), (4.190), (4.191), (4.192) are verified.

This expression of the free energy is only valid in the thermodynamic limit as we used the AMP approximations (and thus becomes exact in the infinite dense graph limit). Its expression remains the same in the vector variables case. It is the same as (24) in [78] where it has been first derived and is formally valid only at a fixed point of the algorithm, which is a minimum of this expression. It is important to understand that during the dynamic of the algorithm, if we plug at each step the AMP quantities Fig. 4.6 in this free energy, it does not necessarily decrease monotonously in time but at a fixed point, it is guaranteed that the algorithm reached a minimum of (4.202): it is a variational expression. The reasons behind why the Bethe free energy is variational when imposing the constraints (4.189), (4.190), (4.191), (4.192) are discussed in [78]. This form or any of the two previous ones (4.193), (4.194) can thus be optimized to derive learning equations, see next section. This free energy is also at the core of the methods used in [99] to solve convergence issues by using an adaptative damping in AMP.

4.3.8 Learning of the model parameters by expectation maximization

In the estimation task, the prior model parameters $\boldsymbol{\theta}$ in (3.59) or the noise variance can be unknown as well. In this case one needs to learn them in some way. This can be done in the Bayesian framework similarly to the classical expectation maximisation method. We consider that θ is the scalar parameter that we want to learn. It starts from the Bayes formula:

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})} \propto Z(\theta)P(\theta) \tag{4.203}$$

where we used that the partition function $Z(\theta) \propto P(\mathbf{y}|\theta)$ from (3.56). Thus if no prior is assumed about θ, then maximizing Z (5.2) or equivalently minimizing the Bethe free energy, considering the other parameters fixed, gives the most probable value of θ. So the method is simple: i) take your favorite form of the Bethe free energy F, ii) solve $\partial_\theta F(\theta|\mathbf{\Gamma}) = 0$ from which you extract a fixed point equation verified by θ that has been isolated: $\theta = f(\theta|\mathbf{\Gamma}^t)$ where $\mathbf{\Gamma}$ is the set of all the other quantities on which depend F and f is a given function, iii) finally add the time: $\theta^{t+1} = f(\theta^t|\mathbf{\Gamma}^t)$ to get an iterative learning equation.

As the Bethe free energy expressions given previously are variational when imposing all the fixed point conditions (4.196), (4.190), (4.191) and (4.192), we can use them in the expectation maximization procedure to fix the new values of the learned parameters. It must be understood that depending on the used form of the Bethe free energy (that are all equivalent), and even for a given fixed form, there exist usually many different ways to write fixed point equations verified by θ. Thus a learning equation must be tried empirically to assess its efficiency even if all fixed point equations are a priori equivalent.

Let us apply the method to the Gaussian noise variance to obtain its generic learning equation, that does not depend on the prior model but only on the AMP fields. As we will see, different learnings can be derived. For example starting from the Bethe free energy (4.193), the noise variance will appear only in the factor term given by (4.165). A possible fixed point equation [35] that arise naturally is in this case:

$$\Delta^{t+1} = \left[\sum_\mu^M \frac{(y_\mu - w_\mu^t)^2}{\left(1+\frac{\Theta_\mu^t}{\Delta^t}\right)^2}\right]\left[\sum_\mu^M \left(1+\frac{\Theta_\mu^t}{\Delta^t}\right)^{-1}\right]^{-1} \tag{4.204}$$

where the AMP quantities iterations are given on Fig. 4.6. But if instead we were starting from the perfectly equivalent expression (4.194), the most straightforward learning would be:

$$\Delta^{t+1} = \left[\frac{1}{M}\sum_\mu^M \left(\frac{(y_\mu - w_\mu^t)^2}{\left(\Delta^t+\Theta_\mu^t\right)^2} + \frac{\Theta_\mu^t}{\Delta^t(\Delta^t+\Theta_\mu^t)}\right)\right]^{-1} \tag{4.205}$$

Many other forms could be derived as for any learned parameter. The first learning equation will be essentially used in this thesis despite the second is valid as well. It can be easily checked that both expressions are the same at the fixed point as they should: any of the two equations

can be derived from (4.193), (4.194) or (4.202) after simplifications.

4.3.9 Dealing with non zero mean measurement matrices

Despite the AMP derivation (see sec. 4.3.3) does *not* rely on the fact that the measurement matrix has zero mean (as opposed to the state evolution, see sec. 5.3), it appears that AMP experiences strong convergence issues when the matrix has a finite mean. Recent works have shed light on solving this issue [76, 99, 100], but these advanced methods are not used in the present thesis. In this work, we used a more classical trick to deal with this problem which starts from (3.18) and noticing that:

$$\frac{1}{M}\sum_{\mu}^{M} y_\mu = \frac{1}{M}\sum_{\mu}^{M}\left(\sum_{i}^{N} F_{\mu i}\, s_i + \xi_\mu\right) \tag{4.206}$$

$$\Rightarrow < \mathbf{y} > = \sum_{i}^{N} \frac{1}{M}\left(\sum_{\mu}^{M} F_{\mu i}\right) s_i + < \boldsymbol{\xi} > \tag{4.207}$$

$$= \sum_{i}^{N} < \mathbf{F}_{\bullet,i} > s_i \tag{4.208}$$

where we have used that the noise has zero mean (if it not the case, its mean can be included in the next rescaling as well). Finally if we call $\tilde{\mathbf{y}} := \mathbf{y} - < \mathbf{y} >$ and $\tilde{\mathbf{F}}_{\bullet,i} := \mathbf{F}_{\bullet,i} - < \mathbf{F}_{\bullet,i} >$, we obtain the rescaled system where now the measurement matrix has zero mean:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{F}}\mathbf{s} + \boldsymbol{\xi} \tag{4.209}$$

that can now be solved without convergence issues. So now we always consider the measurement matrix to have zero mean, as anyway this trick can be used if it is not the case.

5 Phase transitions, asymptotic analyzes and spatial coupling

In this chapter we preset the tools allowing for the asymptotic analyzes of the linear estimation problems and of the approximate message-passing algorithm. We start by discussing how the statistical physics notion of disorder is related to noisy linear estimation problems and give a first example of phase diagram in compressed sensing. Furthermore we introduce the notion of typical complexity phase transitions in sparse linear estimation problems, which connects even more this discipline to statistical physics. It appears that three different complexity regimes are present in linear estimation. Their definitions and implications will be discussed.

Then we will present the replica method of statistical physics of disordered system and use it to compute the potential of the linear estimation problem (3.18) in the general case of a signal with vector components. This potential which is the Bethe free entropy contains the information about the typical complexity of the inference problem as a function of the external parameters, the measurement rate and the noise variance and thus allows to obtain the phase diagram of the problem.

We will derive the state evolution recursions that allow to predict the asymptotic dynamical reconstruction perfomances of the approximate message-passing algorithm, still in the vectorial case. This will be done starting directly from the approximate message-passing algorithm and then starting from the cavity equations as it is usually done. After a discussion on the Bayesian optimality under the prior matching condition, the link between the state evolution and replica analyzes will be underlined: we will show that the fixed points of the state evolution equations give back the optima of the replica potential, i.e. that the two analyzes despite being totally different in their derivations contain the same information on the static properties of the problem.

Finally, we introduce a central notion in this thesis that allows to perfom optimal inference from the information theoretical point of view: the spatial coupling technique. We will discuss its relation to physical concepts such as the nucleation theory and the notion of metastability. Then the approximate message-passing in a form more adapted to use in combination with spatial coupling and the state evolution will be derived. We end up showing how to go from

the form of the algorithm presented here to the well known Montanari's notations.

5.1 Disorder and typical complexity phase transitions in linear estimation

Statistical physics of disordered systems has been specifically created to deal with complex systems that are drawn from some stochastic process. For example, in a disordered spin model such as the fully connected Sherrington-Kirckpatrick spin glass, the interactions amplitudes between spins are random (in this case we speak about interaction disorder), or in a combinatorial optimization problem such as the independent set problem [1], the disorder is the graph instance itself (we speak about structural disorder). The specifically designed replica method [24, 67] allows to perform the necessary thermodynamic averages over these new sources of randomness that are distincts from the pure thermal agitation, i.e. the entropic contribution. It allows to compute the free entropy (i.e. minus the free energy) of the system, its potential function which should not depend on the specific disorder instance in the thermodynamic limit due to the self-averageness of the thermodynamic quantities.

The sources of randomness in the linear estimation problem (3.18) are coming from the measurement noise (that plays the role of a temperature), the random measurement matrix and signal realizations (which induce an interaction disorder). So the replica method is the method of choice and can be "straightforwardly" used in the present context to compute the potential and extract from it the phase diagram of the problem. This potential is minus the Bethe free energy (4.61), (4.194) averaged over these disorder sources as in (4.61),(4.194) all the quantities are dependent on the specific problem instance. Equivalently it is its thermodynamic limit by the self-averageness property. The Bethe free energy in its form (4.61), (4.83) or especially (4.194) is adapted for single instances of the problem, such as when deriving learning equations, see sec. 4.3.8.

Before to present in details the replica computation of the potential in linear estimation problems, we present now the typical phase diagram for sparse linear estimation and discuss the nature of the different phase transitions and typical complexity phases that exist.

5.1.1 Typical phase diagram in sparse linear estimation and the nature of the phase transitions

The replica computation of the next section allows to compute the Bethe free entropy $\Phi(E)$ of the problem, where E is the MSE (3.12) of the reconstruction. This function contains the information on the different phase transitions that happen in the problem as the control parameters, the signal density of non zero (or "large" ones in the approximates sparsity setting, see chap. 6) components ρ and the Gaussian noise variance Δ are tuned. These transitions separate three distinct phases of typical complexity: Fig. 5.1 is the typical phase diagram in the (α, ρ) plane that we will encounter in this thesis. The details of the problem to which it

corresponds will be exposed later on but it it not the point here, it is more to understand the general scenario that happens in linear estimation.

Glassy systems share common features with the present problem (3.18), such as the fact that in the out of equilibrium glass phase, local algorithms are inefficient for sampling the solution space exactly like in the hard phase of the present problem defined hereafter. But it is important to underline that the physics behind these two phases is different. The present inference problem is replica symmetric under the prior matching condition [57] and thus *does not* formally present the typical glass phenomenology like the splitting of the solution space into exponentially many disconnected clusters, i.e. the replica symmetry breaking. Let us now present in details the different phases of the problem.

The "very" easy and easy phases

The first *"very" easy phase* corresponds to the region above the Donoho-Tanner transition, the grey dashed line on Fig. 5.1. In this region ℓ_1 optimization solvers (see sec. 3.4) are efficient for reconstructing the signal, but if sparsity only is known about the signal, they are not anymore in the *easy phase* which corresponds to the region between the Donoho-Tanner and BP transitions. In this region, the potential has a unique maximum corresponding to the $MMSE$ estimate at a "low" MSE, see the blue line on Fig. 5.2. By low, we mean compared to the second maximum of the potential appearing below the BP transition explained after.

In a pure compressed sensing problem where the only knowledge about the signal is that it is sparse, the gap between the Donoho-Tanner and BP transitions is due to the fact that the minimum ℓ_1 solution of the LASSO regression (3.25) does not match the minimizer of the ℓ_0 equivalent of (3.25), i.e. the sparsest solution of the linear system, which is the true solution of the problem. If more knowledge is known about the signal, ℓ_1 optimization solvers are anyway not Bayes optimal as they do not take it into account, it only seeks for the solution with minimum ℓ_1 norm, as opposed to AMP.

In the easy phase, local algorithms such as message-passing or monte-carlo based methods are able to efficiently sample the posterior distribution which is Bayes optimal under the prior matching condition. The "easyness" of this region can be understood thanks to the Bethe free entropy. As we have shown in sec. 4.2.7 the message-passing algorithm fixed points correspond to the optima of this potential (on a single instance). In sec. 5.4 we will show that it remains the case in the thermodynamic limit. Thus we can interpret the algorithm dynamics as a sort of gradient ascent of this potential starting from a high MSE random initialization, until a fixed point is reached and the algorithm converges. Thus when the maximum is unique, AMP will find it and is thus Bayes optimal as it gives the true $MMSE$ estimate. This interpretation of gradient ascent is not rigorously correct in the sense that nothing proves that the Bethe free energy (4.194) is strictly decreased (or the Bethe free entropy strictly increased) at each time step by the message-passing but as BP (and thus AMP which is just its limit on dense graphs) is derived as fixed point equations of the potential (see sec. 4.2.7), if the AMP converges then the

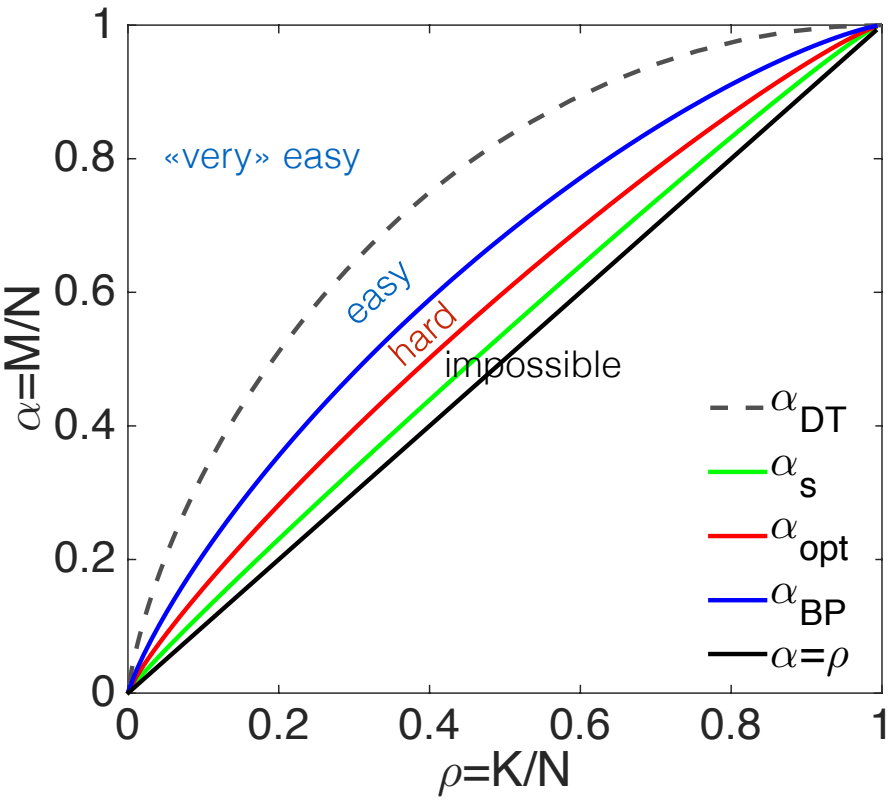

Figure 5.1 – Typical phase diagram of linear estimation problems under sparsity assumption in the plane (density of the signal ρ, measurement rate α). Are plotted the different phase transitions that appear in the problem and the typical complexity phases. The Donoho-Tanner line $\alpha_{DT}(\rho)$ separates the "very" easy phase where ℓ_1 optimization solvers are efficient from the easy phase, where the potential has a unique "low" MSE maximum and thus message-passing is Bayes optimal (under the prior matching condition) with good reconstruction results whereas convex optimization is not. The first order BP phase transition $\alpha_{BP}(\rho)$ is the largest α for which the potential function $\Phi(E)$ has two coexisting maxima. Below this line in the hard phase, message-passing is blocked in a metastable state which is not the $MMSE$ estimate and reconstruction fails, at least without spatial coupling. This hard phase is the gap we want to close thanks to spatial coupling, see sec. 5.5. The optimal transition $\alpha_{opt}(\rho)$ is the α for which the two coexisting maxima of the potential have the same height, i.e. the smallest α at fixed ρ (or highest ρ at fixed α) for which it is theoretically possible to find the signal. In the impossible phase, no algorithm can solve the estimation problem. Finally the static transition $\alpha_s(\rho)$ is defined as the smallest α for which the potential function $\Phi(E)$ has two local maxima. Below this line, all information about the signal is lost. The solid black curve correponds to the $\alpha = \rho$ line and is the fundamental limit of reconstruction (i.e. the optimal transition) in the noiseless limit.

free entropy *must* have been increased until a maxima (local or global) during the dynamics of the message-passing. Further details about the Bethe free energy and its optimization during the message-passing dynamics can be found in [78].

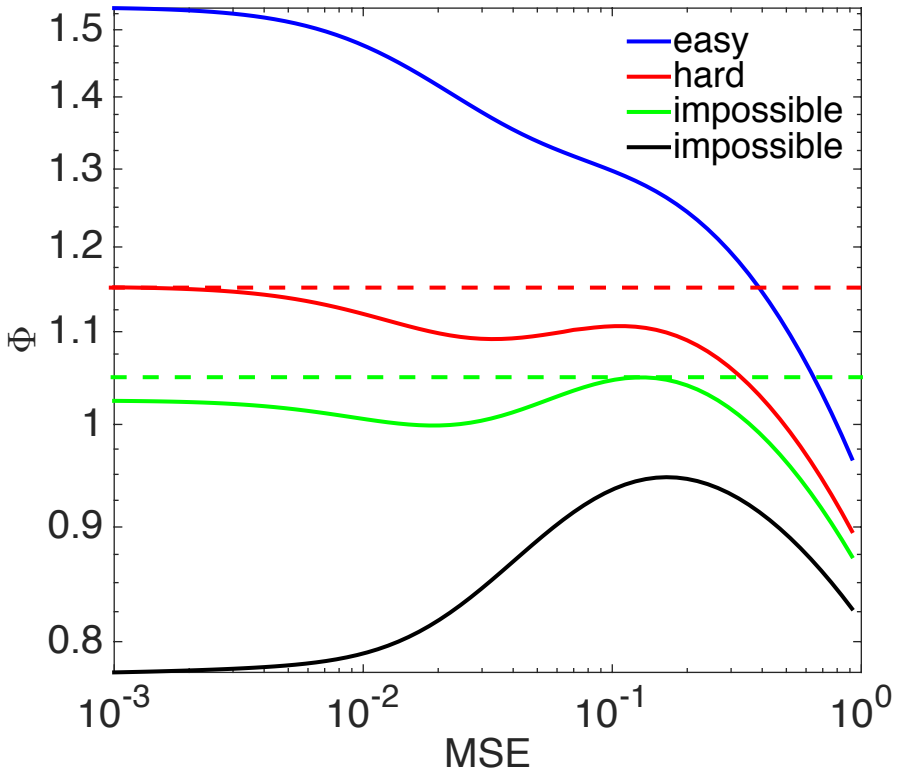

Figure 5.2 – The Bethe free entropy function $\Phi(E)$ for a noisy compressed sensing problem in the different typical complexity phases. In the easy phase, the $MMSE$ solution is unique, and the gradient ascent starting from high MSE performed by the message-passing will find it. In the hard phase, the $MMSE$ solution is still the global maximum but it coexists with a local high MSE fixed point blocking the message-passing (without spatial coupling sec. 5.5). In the impossible phase the equilibrium becomes the wrong solution and then the $MMSE$ metastable fixed point totally disappears below the static transition: no information about the solution is present anymore in the posterior distribution (3.56).

The hard phase and the BP transition

The second phase, right below the blue line and above the red one on the phase diagram Fig. 5.1 is called the *hard phase*. In this region of parameters, the problem is theoretically solvable but local algorithms cannot anymore sample the posterior. It correponds to the regime where the potential has now two distincts maxima, see the red line in Fig. 5.2: the one at low MSE corresponding to the Bayes optimal $MMSE$ estimate is present and still corresponds to the global maximum but it coexists with another spurious local maximum corresponding to a high MSE solution of the message-passing equations. This solution is refered as the *metastable state* as it not the true equilibrium given by the free entropy global maximum. So the message-passing climbing the potential starting from high MSE will reach this fixed point before being able to see the $MMSE$ one, and thus reconstruction will fail. With a monte carlo or any other local algorithm (which is of course initialized randomly, thus in the basin of attraction of the metastable wrong solution), only an exponentially long sampling of

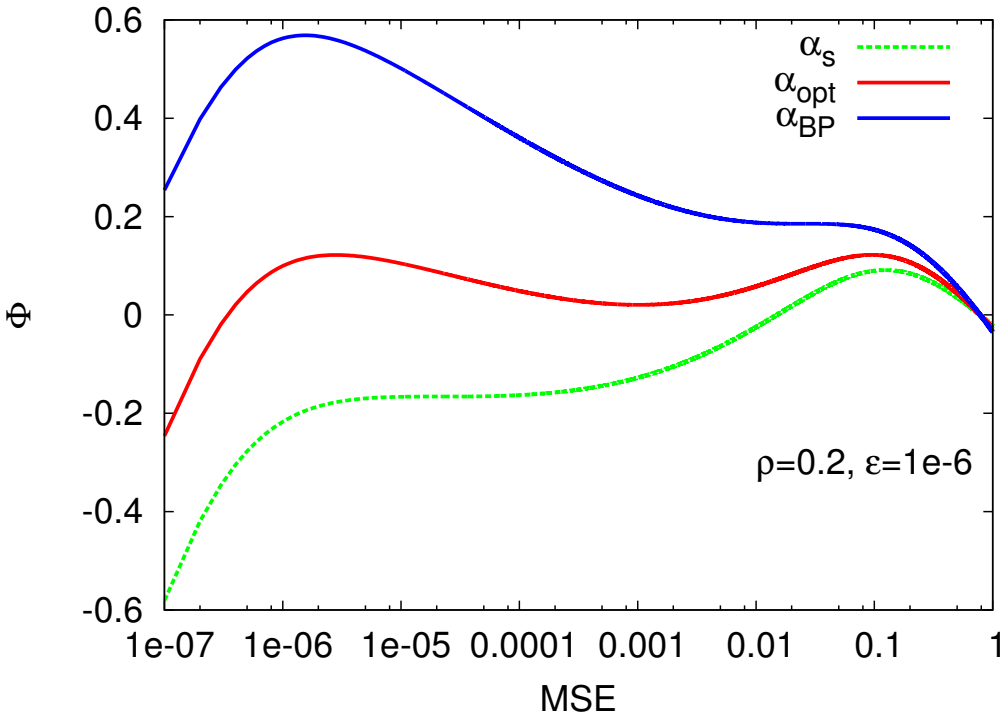

Figure 5.3 – The Bethe free entropy $\Phi(E)$ for signals of density $\rho = 0.2$, with an effective noise of variance $\Delta \in O(10^{-6})$. The three lines depict the potential for three different measurement rates corresponding to the critical values: $\alpha_{BP} = 0.3559$, $\alpha_{opt} = 0.2817$, $\alpha_s = 0.2305$. The two local maxima exists for $\alpha \in (\alpha_s, \alpha_{BP})$, at α_{opt} the low MSE maxima becomes the global one. This scenario is typical in linear estimation problems.

the solution space, passing through exponentially rare configurations with a low free entropy would allow the algorithm to ultimately reach the true equilibrium, at least without spatial coupling (sec. 5.5). This blocking of local algorithms is observed also in glassy phases of combinatorial optimization problems, but in this case the free entropy is way more rough than Fig. 5.2 due to the spontaneous replica symmetry breaking.

The transition between the easy and hard phases is called the *BP transition* or *spinodal transition* as it corresponds to the separation between a region where message-passing based algorithms are Bayes optimal from one where it fails. Exactly at the transition point, the Bethe free entropy has typically the shape given by the blue line on Fig. 5.3 with a plateau appearing at high MSE: $\alpha_{BP}(\rho)$ is defined as the largest $\alpha(\rho)$ for which the potential has two local maxima. This transition does *not* affect the Bayes optimal estimator, it is an algorithmic phase transition of the first order type: the order parameter which is the asymptotic reconstruction MSE we would get by message-passing (without spatial coupling) jumps *discontinuously* from a low (in the noisy setting) or zero (in the noiseless setting) value when situated in the easy phase to a high value in the hard phase, the jump happening exactly on the transition line. But we will see in the next that when the noise is very large, this transition can become continuous (there is no more sharp transition), and only one maximum will exist at any measurement rate

that smoothly moves to a lower MSE solution as the measurement rate increases.

This plateau of the free entropy explains the phenomenon of *critical slowing down* close to the BP transition, a phenomenon typical from first order transitions: interpreting again the message-passing as a gradient ascent of the potential, as the paramaters of the problem get closer to the BP transition this plateau with very low gradient reduces greatly the convergence rate of the algorithm until it becomes infinitely long in the large limit exactly at the transition and then it fails. The same behavior happens in many other problems such as random K-SAT [85] and is inherent to local algorithms facing first order transitions. The same phenomenon of appearance of a metastable state preventing the equilibrium to be reached occurs in the supercooled state, where a liquid is blocked in the liquid state despite the temperature is below its critical temperature of solidification.

The impossible phase and the optimal transition

The third phase is all the region below the red line on Fig. 5.1. Is it the *impossible phase*. Here, no algorithm is able to find the solution as the potential is dominated by a high MSE solution and thus even the Bayes optimal inference would fail, see the green and black curves on Fig. 5.2.

The first order transition between the hard and impossible phases, called the *optimal transition* $\alpha_{opt}(\rho)$ happens when the two coexisting free entropy maxima have exactly the same height, i.e. at the exact parameters values where the $MMSE$ solution (the global maximum) jumps from a low MSE to an high one, see the red curve on Fig. 5.3.

From an algorithmic point of view and running purely local algorithms (i.e. without considering spatial coupling), this transition will not be detected as the algorithm was failing before it in the hard phase, and continues to do so after in the impossible phase.

The static transition

Despite the fact that in all this impossible phase the algorithm behaves the same, from the potential (and thus thermodynamical) point of view, another transition occurs denoted as the *static* transition: $\alpha_s(\rho)$ is defined as the smallest α for which the potential function $\Phi(E)$ has two local maxima. It separates the impossible phase into two distinct phases, "one more impossible than the other". In the impossible phase before that the static transition happens, the potential is dominated by the high MSE solution (by definition of the impossible phase) but it remains a local maximum of the potential corresponding to the $MMSE$ estimate, which is now the metastable state. This maximum could be reached initializing the message-passing equations close enough to the solution, inside the basin of attraction of this state. But anyway, the true equilibrium correponds to a failure of the reconstruction. The static transition happens when this low error metastable state disappears, see the green curve on Fig. 5.3, and only remains the high MSE state, see the black curve on Fig. 5.2. Below the static

transition, even if we would initialize the reconstruction algorithm on the solution itself, it would converge to the high MSE state: it does not remain any information about the signal in the posterior distribution (3.56) because there are too few measurements for the signal density, there are too noisy or both at the same time.

Sum up of the different transitions and algorithmic implications

To summarize, the AMP algorithm exhibits a double phase transition. It is asymptotically equivalent to the optimal Bayes inference at large $\alpha_{BP} < \alpha < 1$, where it matches the optimal reconstruction with a small value of the MSE. At low values of $\alpha < \alpha_{opt}$ the AMP is also asymptotically equivalent to the optimal Bayes inference as the potential has only one maximum, but in this low-sampling-rate region the optimal result leads to a large MSE. In the intermediate region $\alpha_{opt} < \alpha < \alpha_{BP}$, AMP leads to large MSE, but the optimal Bayes inference leads to low MSE. This is the region where one needs to improve on AMP, using for example the spatial coupling technique discussed in sec. 5.5.

It must be understood that all these considerations and phase transitions except the Donoho-Tanner one are properties of the problem itself, not of the algorithm used to reconstruct the signal. This potential, the Bethe free entropy, is conjectured exact in the thermodynamic limit for such dense graphs. But it does not mean that the easy phase is easy for all reconstruction algorithms, it means that mean field methods are appropriate and are able to sample properly the posterior until the BP transition under the prior matching condition. But as we have seen, convex ℓ_1 optimization based solvers cannot reconstruct averywhere in the easy phase: they experience another phase transition preventing the solver to reconstruct well before the BP transition.

5.2 The replica method for linear estimation over the additive white Gaussian noise channel

The replica method leads to asymptotically exact evaluation of the logarithm of the partition function Z (5.2), which is here the Bethe free entropy. In general, if the partition function can be evaluated precisely then the marginal means with respect to the true posterior and the associated MSE of the optimal inference can be computed. More generally, it is a method for averaging logarithms of complex functions depending on some disorder. It does not always lead to a Bethe form of the potential but it appears that in cases where the factor graph corresponding to the model of interest is not a tree nor dense, the replica method is far too complex to be applied. It is thus usual to confound the potential extracted from the replica analysis and the Bethe free entropy as it can be applied only in cases where the Bethe free entropy is the true potential. This is the case in our problem (3.18), the Bethe free entropy (4.61), (4.194) is actually exact because the problem is dense and thus its average matches the potential extracted from the replica analysis.

The optimal reconstruction in compressed sensing, information that we will extract from the replica potential, was studied extensively in [101]. The replica method was used in compressed sensing similarly as in the present thesis in [102, 103] for example. Let us now derive the potential of the problem (3.18). All the computations are made considering that the signal is made of B-d sections (see Fig. 4.5) and that the prior is factorizable over them, with the same prior P_0 independently of the section. Furthermore we assume the prior matching condition which as already explained implies the replica symmetry in inference problems.

5.2.1 Replica trick and replicated partition function

We start from the definition of the free entropy potential at fixed section size B:

$$\Phi_B := \lim_{L\to\infty} \frac{1}{L}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\log(Z(\mathbf{F},\boldsymbol{\xi},\mathbf{s}))\} \tag{5.1}$$

$$Z(\mathbf{F},\boldsymbol{\xi},\mathbf{s}) = \int \left[\prod_l^L d\mathbf{x}_l P_0(\mathbf{x}_l)\right] \prod_\mu^M \sqrt{\frac{\mathrm{snr}}{2\pi}} e^{-\frac{\mathrm{snr}}{2}\left(\sum_l^L \mathbf{F}^\intercal_{\mu l}(\mathbf{s}_l-\mathbf{x}_l)+\xi_\mu\right)^2} \tag{5.2}$$

where we have used (3.18) to replace $\mathbf{y}$ and Z is the normalization constant of the full posterior distribution (3.56), i.e. the partition function, a random variable of the disorder. It can be transformed using the so called replica trick, a trivial mathematical identity:

$$\Phi_B = \lim_{L\to\infty}\lim_{n\to 0} \frac{1}{L}\frac{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{Z^n\}-1}{n} \tag{5.3}$$

where $\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}$ is the average over all the sources of disorder in (3.18). So the problem of computing the free energy is converted into computing the n^{th} moment of the partition function. Z^n is the so-called replicated partition function as it can be interpreted as the partition function of n independent systems drawn from the same distribution, refered as the replicas. The index associated with the replicas is $a \in \{1,..,n\}$:

$$Z^n = (\mathrm{snr}/2\pi)^{\frac{Mn}{2}} \int \left[\prod_{l,a}^{L,n} d\mathbf{x}_l^a P_0(\mathbf{x}_l^a)\right] \prod_\mu^M e^{-\frac{\mathrm{snr}}{2}\sum_a^n\left(\sum_l^L \mathbf{F}^\intercal_{\mu l}(\mathbf{s}_l-\mathbf{x}_l^a)+\xi_\mu\right)^2} \tag{5.4}$$

The average replicated partition function can be rearranged as:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{Z^n\} = (\mathrm{snr}/2\pi)^{\frac{Mn}{2}}\, \mathbb{E}_{\mathbf{s}}\left\{\int \left[\prod_{l,a}^{L,n} d\mathbf{x}_l^a P_0(\mathbf{x}_l^a)\right] \prod_\mu^M X_\mu\right\} \tag{5.5}$$

where we have defined:

$$X_\mu := \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\left\{ e^{-\frac{\text{snr}}{2}\sum_a^n\left(\sum_l^L \mathbf{F}_{\mu l}^\intercal(\mathbf{s}_l-\mathbf{x}_l^a)+\xi_\mu\right)^2}\right\} \tag{5.6}$$

$$= \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\left\{ e^{-\frac{\text{snr}}{2}\sum_a^n(v_\mu^a)^2}\right\} \tag{5.7}$$

$$v_\mu^a := \sum_l^L \mathbf{F}_{\mu l}^\intercal(\mathbf{s}_l-\mathbf{x}_l^a)+\xi_\mu \tag{5.8}$$

In order to compute X_μ (5.7), we can apply the central limit theorem to the quantity v_μ^a which is a sum of independent terms as the measurement matrix is i.i.d. We thus need its first two moments to define its associated Gaussian distribution. Using the fact that both the measurement matrix and the i.i.d Gaussian noise have zero mean we get:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{v_\mu^a\} = 0 \tag{5.9}$$

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{(v_\mu^a)^2\} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\left\{\sum_{l,k}^{L,L}[\mathbf{F}_{\mu l}^\intercal(\mathbf{s}_l-\mathbf{x}_l^a)]^\intercal\mathbf{F}_{\mu k}^\intercal(\mathbf{s}_k-\mathbf{x}_k^a)+2\xi_\mu\sum_l^L\mathbf{F}_{\mu l}^\intercal(\mathbf{s}_l-\mathbf{x}_l^a)+\xi_\mu^2\right\}$$

$$= \sum_{l,k}^{L,L}\left[(\mathbf{s}_l-\mathbf{x}_l^a)^\intercal\mathbb{E}_{\mathbf{F}}\left\{\mathbf{F}_{\mu l}\mathbf{F}_{\mu k}^\intercal\right\}(\mathbf{s}_k-\mathbf{x}_k^a)\right]+1/\text{snr} \tag{5.10}$$

Using the fact that each element of the matrix is i.i.d with variance $1/L$, we find that only the diagonal elements of the matrix $\mathbb{E}_{\mathbf{F}}\left\{\mathbf{F}_{\mu l}\mathbf{F}_{\mu k}^\intercal\right\}$ are non zero:

$$\mathbb{E}_{\mathbf{F}}\left\{\mathbf{F}_{\mu l}\mathbf{F}_{\mu k}^\intercal\right\} = \frac{\delta_{k,l}}{L}\boldsymbol{I}_B \tag{5.11}$$

$$\Rightarrow\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{(v_\mu^a)^2\} = 1/L\sum_l^L(\mathbf{s}_l-\mathbf{x}_l^a)^\intercal(\mathbf{s}_l-\mathbf{x}_l^a)+1/\text{snr} \tag{5.12}$$

where $\boldsymbol{I}_B$ is the identity matrix of dimension $B\times B$. Now we define new macroscopic order parameters which will be considered as the new degrees of freedom of the replicated system, instead of the individual replica states $\{\mathbf{x}_a\}$. In this way, we average out the microscopic properties of the system (the randomness instance, i.e. the replica individual states) to get access to the macroscopic ones (the observables, such as the MSE), the goal of the replica methodology. This is referred as coarse-graining in physics:

$$m_a := 1/L\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{s}_l \tag{5.13}$$

$$Q_a := 1/L\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{x}_l^a \tag{5.14}$$

$$q_{ab} := 1/L\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{x}_l^b \tag{5.15}$$

m_a is the overlap between the replica state $\hat{\mathbf{x}}^a$ and the signal $\mathbf{s}$, Q_a is the power (or self-overlap) of the replica a and q_{ab} is the overlap between replicas a and b. Rewriting the previous

moment (5.12) in terms of these new quantities, we get:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{(v_\mu^a)^2\} = <\mathbf{s}^2> -2m_a + Q_a + 1/\text{snr} \tag{5.16}$$

Exactly in the same way, we get the cross terms $\forall\ a \neq b$:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{v_\mu^a v_\mu^b\} = <\mathbf{s}^2> -(m_a + m_b) + q_{ab} + 1/\text{snr} \tag{5.17}$$

The dependence on the measurement index μ of v_μ^a (and thus X_μ as well (5.7)) is lost due to the averaging. Thus we note $v_\mu^a = v^a$, $X_\mu = X$. We now apply the replica symmetric ansatz which is valid for inference problems (and more generally planted problems) on locally tree-like or highly dense graphs under the prior matching condition such as in the present case [24, 57]. In the replica method, it is expressed by removing the replica indices in the macroscopic order parameters (thus the name of the ansatz):

$$q_{ab} = q\ \forall\ (a,b : a \neq b) \tag{5.18}$$

$$Q_a = Q\ \forall\ a \tag{5.19}$$

$$m_a = m\ \forall\ a \tag{5.20}$$

Now we have computed the necessary moments (5.16), (5.17) we can write the covariance matrix $\mathbf{G}$ of $\{v^a\}$ under this ansatz which reads $\forall\ (a,b)$:

$$G_{ab} := \mathbb{E}_{\mathbf{F},\boldsymbol{\xi}}\{v^a v^b\} \tag{5.21}$$

$$= <\mathbf{s}^2> -2m + 1/\text{snr} + q + (Q-q)\delta_{a,b} \tag{5.22}$$

$$\Rightarrow \mathbf{G} = \left(<\mathbf{s}^2> -2m + 1/\text{snr} + q\right)\mathbf{1}_n + (Q-q)\boldsymbol{I}_n \tag{5.23}$$

where $\mathbf{1}_n$ is a matrix full of ones of dimension $n \times n$. We thus have:

$$X = \mathbb{E}_{\mathbf{v}}\{e^{-\frac{\text{snr}}{2}\mathbf{v}^\intercal\mathbf{v}}\} \tag{5.24}$$

$$P(\mathbf{v}) = [(2\pi)^n \det(\mathbf{G})]^{-1/2} e^{-\frac{1}{2}\mathbf{v}^\intercal\mathbf{G}^{-1}\mathbf{v}} \tag{5.25}$$

The explicit computation of X by Gaussian integral gives:

$$X = [(2\pi)^n \det(\mathbf{G})]^{-1/2} \int d\mathbf{v} e^{-\frac{1}{2}\mathbf{v}^\intercal(\mathbf{G}^{-1}+\text{snr}\boldsymbol{I}_n)\mathbf{v}} \tag{5.26}$$

$$= [\det(\boldsymbol{I}_n + \text{snr}\mathbf{G})]^{-1/2} \tag{5.27}$$

The eigenvectors of $\mathbf{G}$ are one eigenvector $[1]_a^n = [1,1,\ldots,1]$ with associated eigenvalue $Q-q+n\left(1-2m+1/\text{snr}+q\right)$ and $n-1$ eigenvectors of the type $[0,..,0,-1,1,0,..,0]$ with the couples $[-1,1]$ shifting by one component from one eigenvector to the next. Their degenerated

eigenvalue is $Q-q$. Therefore:

$$\det\left(\boldsymbol{I}_n+\text{snr}\mathbf{G}\right)=\frac{1+\text{snr}\left[Q-q+n\left(<\mathbf{s}^2>-2m+1/\text{snr}+q\right)\right]}{\left[1+\text{snr}(Q-q)\right]^{1-n}} \tag{5.28}$$

from which we get:

$$\lim_{n\to 0} X=e^{-\frac{n}{2}\left[\frac{q-2m+<\mathbf{s}^2>+1/\text{snr}}{Q-q+1/\text{snr}}+\log(1/\text{snr}+Q-q)-\log(1/\text{snr})\right]} \tag{5.29}$$

When computing (5.5), we need to enforce the constraints that the new order parameters satisfy their definitions (5.15). This is done by the usual trick of rewriting 1 as the inverse Fourier transform of its Fourier transform and plugging this expression in the definition of the averaged replicated partition function that we are computing:

$$\begin{aligned}1=&\int\left[\prod_a^n dQ_a d\hat{Q}_a dm_a d\hat{m}_a\right]\left[\prod_{b,a<b}^{n,(n-1)/2} dq_{ab}d\hat{q}_{ab}\right]\\&\exp\left[-\sum_a^n \hat{m}_a(m_aL-\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{s}_l)+\sum_a^n\hat{Q}_a(Q_aL/2-1/2\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{x}_l^a)\right.\\&\left.-\sum_{b,a<b}^{n,(n-1)/2}\hat{q}_{ab}(q_{ab}L-\sum_l^L(\mathbf{x}_l^a)^\intercal\mathbf{x}_l^b)\right]\end{aligned} \tag{5.30}$$

Plugging this into the average replicated partition function (5.5) expression we get:

$$\begin{aligned}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{Z^n\}=&(\text{snr}/2\pi)^{\frac{Mn}{2}}\int\left[\prod_a^n dQ_a d\hat{Q}_a dm_a d\hat{m}_a\right]\left[\prod_{b,a<b}^{n,(n-1)/2} dq_{ab}d\hat{q}_{ab}\right]\\&\exp\left[L\left(\frac{1}{2}\sum_a^n\hat{Q}_aQ_a-\frac{1}{2}\sum_{b,a\neq b}^{n,(n-1)}\hat{q}_{ab}q_{ab}-\sum_a^n\hat{m}_am_a\right)\right]\left[\prod_\mu^M X\right]\\&\underbrace{\left(\int_{\mathbb{R}^B}d\mathbf{s}P_0(\mathbf{s})\int_{\mathbb{R}^{Bn}}\left[\prod_a^n d\mathbf{x}^aP_0(\mathbf{x}^a)\right]\exp\left[-\frac{1}{2}\sum_a^n\hat{Q}_a(\mathbf{x}^a)^\intercal\mathbf{x}^a+\frac{1}{2}\sum_{b,b\neq a}^{n,(n-1)}\hat{q}_{ab}(\mathbf{x}^a)^\intercal\mathbf{x}^b+\sum_a^n\hat{m}_a(\mathbf{x}^a)^\intercal\mathbf{s}\right]\right)}_{:=\Gamma}{}^L\end{aligned} \tag{5.31}$$

At this stage, it is worth noticing that thanks to the average over the disorder $(\mathbf{F},\boldsymbol{\xi},\mathbf{s})$ and the introduction of the replica macroscopic orders parameters (5.13), (5.15), (5.14) we converted a sytem which partition function was (5.4) i.e. made of n i.i.d strongly disordered replicas each with their own interacting variables into an asymptotically (we averaged over the disorder and used the central limit theorem to get there) equivalent system made of n interacting replicas but independent of the original sources of disorder. Furthermore, their variables are effectively independent one of the other inside the same replica but interact with their equivalent in the other replicas: for example the $\mathbf{x}_l^a$ variable is effectively independent of $\{\mathbf{x}_k^a\}_{k\neq l}^L$ but interacts with $\{\mathbf{x}_l^b\}_{b\neq a}^n\ \forall b$ with effective interaction $\hat{q}_{ab}$, the dual variable of (5.15), it interacts with itself through the self field $\hat{Q}_a$ (5.14) and with the signal through $\hat{m}_a$ (5.13). The interactions between variables belonging to the same replica are "hidden" in the replica

order parameters that are now variables which control the new coupling interactions between replicas ($\{\hat{m}_a, \hat{q}_{ab}, \hat{Q}_a\}$).

These new effective interactions are way more easy to deal with as the dependence on the disorder (noise and measurement matrix realizations) has been averaged out when applying the central limit theorem to v^a, which naturally gave rise to the replica order parameters. Looking at Γ, we realize that there is a Gaussian coupling between the replicas. As usual in this situation (like is the AMP derivation sec. 4.3), we use the Stratanovitch transform to decouple them by linearizing the exponent, the payoff being an additional Gaussian integral to perform at the end. In B dimensions the transform is given by:

$$e^{\frac{\hat{q}}{2}\sum_{b,a\neq b}^{n,(n-1)}\mathbf{x}_a^\intercal\mathbf{x}_b} = \prod_i^B e^{\frac{\hat{q}}{2}\sum_{b,a\neq b}^{n,(n-1)}x_{a,i}x_{b,i}} \tag{5.32}$$

$$= \prod_i^B \int \mathscr{D}z_i\, e^{\sqrt{\hat{q}}\,z_i\sum_a^n x_{a,i} - \frac{\hat{q}}{2}\sum_a^n {x_{a,i}}^2} \tag{5.33}$$

$$= \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z}\, e^{\sqrt{\hat{q}}\,\mathbf{z}^\intercal\sum_a^n\mathbf{x}_a} e^{-\frac{\hat{q}}{2}\sum_a^n\mathbf{x}_a^\intercal\mathbf{x}_a} \tag{5.34}$$

where we remind that $\mathscr{D}\mathbf{z} := \prod_i^B \mathscr{D}z_i = \prod_i^B \mathcal{N}(z_i|0,1)dz_i$ is a unit centered B-d Gaussian measure and sums of the form $\sum_a^n\mathbf{x}_a = [\sum_a^n x_{a,i}]_i^B$ are vectors. Using the replica symmetric ansatz we obtain that:

$$\Gamma = \int_{\mathbb{R}^B} d\mathbf{s}\mathscr{D}\mathbf{z}\,P_0(\mathbf{s})\underbrace{\left[\int_{\mathbb{R}^B} d\mathbf{x}P_0(\mathbf{x})e^{-\frac{1}{2}(\hat{Q}+\hat{q})\mathbf{x}^\intercal\mathbf{x}+\hat{m}\mathbf{x}^\intercal\mathbf{s}+\mathbf{z}^\intercal\mathbf{x}\sqrt{\hat{q}}}\right]^n}_{:=f(\mathbf{z},\mathbf{s})} \tag{5.35}$$

In addition we have:

$$\Gamma \underset{n\to 0}{\approx} \exp\left(n\int_{\mathbb{R}^B} d\mathbf{s}\mathscr{D}\mathbf{z}P_0(\mathbf{s})\log\big(f(\mathbf{z},\mathbf{s})\big)\right) \tag{5.36}$$

Combining (5.36) and (5.31) under the replica symmetric ansatz, we get the expression of the averaged replicated partition function as $n\to 0$:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{x}}\{Z^n\} \underset{n\to 0}{\approx} \int dQd\hat{Q}dmd\hat{m}dqd\hat{q}e^{nL\tilde{\Phi}_B(m,\hat{m},q,\hat{q},Q,\hat{Q})} \tag{5.37}$$

where the replica potential, up to irrelevant constants independent on the order parameters is:

$$\begin{aligned}\tilde{\Phi}_B(m,\hat{m},q,\hat{q},Q,\hat{Q}) = &\frac{1}{2}\left(\hat{Q}Q+\hat{q}q-2\hat{m}m\right)\\ &-\frac{\alpha B}{2}\left(\frac{q-2m+<\mathbf{s}^2>+1/\mathrm{snr}}{Q-q+1/\mathrm{snr}}+\log(1/\mathrm{snr}+Q-q)\right)\\ &+\int_{\mathbb{R}^B} d\mathbf{s}\mathscr{D}\mathbf{z}P_0(\mathbf{s})\log\left(\int_{\mathbb{R}^B} d\mathbf{x}P_0(\mathbf{x})e^{\hat{m}\mathbf{x}^\intercal\mathbf{s}+\sqrt{\hat{q}}\mathbf{z}^\intercal\mathbf{x}-\frac{1}{2}(\hat{q}+\hat{Q})\mathbf{x}^\intercal\mathbf{x}}\right)\end{aligned} \tag{5.38}$$

where it must be kept in mind that the vectors in the last integral are one B-d section, not the overall vectors.

5.2.2 Saddle point estimation

For the replica trick (5.3) to be formally valid, the limit $n \to 0$ should be taken before the limit over L. But we need to estimate the integral (5.37) by its saddle point as it is intractable otherwise. This can be justified only if the limit $L \to \infty$ is performed before the limit over n, but in the same time the expression (5.37) has been obtained already considering n very small.

We thus assume that n is small enough for (5.37) to be accurate but yet fixed. Then we assume that the limits commute and perform the saddle point estimation of the integral before to really let $n \to 0$. This is not rigorous, but heuristically verified in many different models, including inference problems. The saddle point estimate is performed by taking the optimum of $\tilde{\Phi}_B$, given by its fixed point value with respect to the different order parameters. The resulting potential actually corresponds to the desired Bethe free entropy as seen from (5.3):

$$\Phi_B := \tilde{\Phi}_B(m^*, \hat{m}^*, q^*, \hat{q}^*, Q^*, \hat{Q}^*) \tag{5.39}$$

The optimization gives these fixed point values, denoted with stars:

$$\frac{\partial \tilde{\Phi}_B}{\partial m} = 0 \Rightarrow \hat{m}^* = \frac{\alpha B}{Q^* - q^* + 1/\mathrm{snr}} \tag{5.40}$$

$$\frac{\partial \tilde{\Phi}_B}{\partial q} = 0 \Rightarrow \hat{q}^* = \alpha B \frac{1/\mathrm{snr} + B < \mathbf{s}^2 > -2m^* + q^*}{(Q^* - q^* + 1/\mathrm{snr})^2} \tag{5.41}$$

$$\frac{\partial \tilde{\Phi}_B}{\partial Q} = 0 \Rightarrow \hat{Q}^* = \alpha B \frac{2m^* - B < \mathbf{s}^2 > -2q^* + Q^*}{(Q^* - q^* + 1/\mathrm{snr})^2} \tag{5.42}$$

to be plugged into the previous potential (9.9) to get its most general form. But as we assume the prior matching condition, further simplifications are possible.

5.2.3 The prior matching condition

As shown after in sec. 5.3.3 together with sec. 5.4, the matching prior condition implies:

$$q^* = m^* \tag{5.43}$$

$$Q^* = B < \mathbf{s}^2 > \tag{5.44}$$

$$E = < \mathbf{s}^2 > - \frac{m^*}{B} \tag{5.45}$$

which implies for their conjugate parameters (5.40), (5.41), (5.42):

$$\hat{q}^* = \hat{m}^* = \frac{\alpha B}{BE + 1/\mathrm{snr}} \tag{5.46}$$

$$\hat{Q}^* = 0 \tag{5.47}$$

where $E := <(\mathbf{s}-\mathbf{x})^2>$ is the MSE, i.e. the observable of the system. We get the final expression of the Bethe free entropy:

$$\begin{aligned}\Phi_B(E) = &-\frac{\alpha B}{2}\left(\log(1/\mathrm{snr}+BE)+\frac{B<\mathbf{s}^2>-BE}{1/\mathrm{snr}+BE}\right)\\ &+\int_{\mathbb{R}^B} d\mathbf{s} P_0(\mathbf{s}) \mathscr{D}\mathbf{z} \log\left(\int_{\mathbb{R}^B} d\mathbf{x} P_0(\mathbf{x}) \exp\left(\frac{\mathbf{s}^\intercal\mathbf{x}}{\Sigma(E)^2}+\frac{\mathbf{z}^\intercal\mathbf{x}}{\Sigma(E)}-\frac{\mathbf{x}^\intercal\mathbf{x}}{2\Sigma(E)^2}\right)\right)\end{aligned} \tag{5.48}$$

where:

$$\Sigma(E)^2 := 1/\hat{m}^* = \frac{1/(B\mathrm{snr})+E}{\alpha} \tag{5.49}$$

This expression is general for linear estimation of fixed section size B signals. It will be used in most of the theoretical studies in this thesis. The asymptotic variance (5.49) of the maximum likelihood estimate increases linearly both with the noise variance and the MSE. Furthermore, their common effect is enhanced as the measurement rate decreases.

5.3 The cavity method for linear estimation: state evolution analysis

The state evolution analysis, referred as the cavity method in physics [24, 67] is a statistical analysis that allows to monitor the approximate message-passing algorithm dynamics and performance in the limit of reconstructing infinitely large signals, i.e. in the thermodynamic limit. We consider the case of i.i.d Gaussian matrices $\mathbf{F}$ for which state evolution has been originally derived [97] and then proved to be rigorous in large generality [104] (see also [35, 89, 105]). As in the replica computation of the previous section, this assumption is essential in order to decouple the signal components in the analysis. Extension to more general ensembles such as row orthogonal matrices could be considered [106] but it is out of the scope of the present thesis. In addition, as we will show in chap. 7 in great details, the state evolution analysis derived in the i.i.d Gaussian matrices case is a good predictive tool of the reconstruction performances of the AMP algorithm even with structured operators, despite not predicting well the dynamics before convergence nor being rigorous.

As before, we consider the general case of signals with B-d sections and assume the proper scaling to get a codeword with power $P = 1$: $F_{ij} \sim \mathcal{N}\left(F_{ij}|0, 1/L\right) \forall\ (i,j) \Rightarrow F_{ij} \in O(1/\sqrt{L})\ \forall\ (i,j)$. As in the replica analysis, we consider the case of a factorizable prior over the sections with the same prior for every sections. This drastically simplifies the analysis due to the induced

symmetry between all the sections. We will look at the section index dependent prior case in the chapter about superposition codes in the sec. 9.5.1. Furthermore, we consider having perfect knowledge of the channel noise statistical properties as the prior which generated the signals as well so that we place ourselves under the prior matching case.

5.3.1 Derivation starting from the approximate message-passing algorithm

One can perform the analysis starting from the cavity quantities defined in the AMP derivation starting from BP, see sec. 4.3.3 as it is more classicaly done [35]. This will be done in the next section, but here we will follow another path starting from the AMP algorithm itself. We refer to Fig. 4.6 for the definitions of all the quantities that we will use in the derivation.

The aim is to evaluate the asymptotic AMP posterior estimate of a section at each time in order to compute the asymptotic mean square error $E^{t+1}(E^t)$ as a function of its value at time t. The posterior estimate is given by applying the denoising function $\mathbf{f}_{a_l}$ to the variables $((\mathbf{\Sigma}_l^{t+1})^2, \mathbf{R}_l^{t+1})$, see Fig. 4.6. We thus need to get an asymptotic estimate of $\mathbf{R}_l^{t+1}$, the average of variable $\mathbf{x}_l$ at time $t+1$ with respect to the likelihood. $(\mathbf{\Sigma}_l^{t+1})^2$ is its associated variance. Injecting in $\mathbf{R}_l^{t+1}$ the expression of w_μ^{t+1} as a function of the previous time quantities and using (3.18) to replace the measurement by its expression in terms of the signal, measurement matrix and noise (i.e. the disorder sources), we get the following expression for $\mathbf{R}_l^{t+1}$:

$$\mathbf{R}_l^{t+1} = \mathbf{a}_l^t + (\mathbf{\Sigma}_l^{t+1})^2 \sum_\mu^M \frac{\mathbf{F}_{\mu l}}{1/\mathrm{snr} + \Theta_\mu^{t+1}} \left[\sum_k^L \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \underbrace{\Theta_\mu^{t+1} \frac{y_\mu - w_\mu^t}{1/\mathrm{snr} + \Theta_\mu^t}}_{:=\Lambda_\mu^t} \right] \tag{5.50}$$

$$= \mathbf{a}_l^t + (\mathbf{\Sigma}_l^{t+1})^2 \sum_\mu^M \frac{\mathbf{F}_{\mu l}}{1/\mathrm{snr} + \Theta_\mu^{t+1}} \left[\mathbf{F}_{\mu l}^\intercal (\mathbf{s}_l - \mathbf{a}_l^t) + \sum_{k \neq l}^L \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] \tag{5.51}$$

Now we use the fact that Θ_μ^t is asymptotically independent of μ as we can replace the $F_{\mu i}^2$ elements by the matrix variance $1/L$:

$$\Theta_\mu^t \approx \Theta^t := 1/L \sum_i^N v_i^t \tag{5.52}$$

$$\Rightarrow (\mathbf{\Sigma}_l^{t+1})^2 \approx \frac{1/\mathrm{snr} + \Theta^{t+1}}{B\alpha} \mathbf{1}_B \tag{5.53}$$

where we have used $M = LB\alpha$ and $\mathbf{1}_B$ is a vector of ones of size B. This simplifies the expression to:

$$\mathbf{R}_l^{t+1} = \mathbf{a}_l^t + \frac{1}{B\alpha} \sum_\mu^M \mathbf{F}_{\mu l} \left[\sum_{k \neq l}^{L-1} \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] + \frac{1}{B\alpha} \sum_\mu^M \mathbf{F}_{\mu l} \left[\mathbf{F}_{\mu l}^\intercal (\mathbf{s}_l - \mathbf{a}_l^t) \right] \tag{5.54}$$

Now we notice that we can simplify the last term in the previous equality:

$$\sum_{\mu}^{M}\mathbf{F}_{\mu l}\left[\mathbf{F}_{\mu l}^{\intercal}(\mathbf{s}_l-\mathbf{a}_l^t)\right]=\underbrace{\left[\sum_{\mu}^{M}F_{\mu i}^2(s_i-a_i^t)\right]_{i\in l}}_{=B\alpha(\mathbf{s}_l-\mathbf{a}_l^t)}+\underbrace{\left[\sum_{\mu}^{M}\sum_{j\in l:j\neq i}^{B-1}F_{\mu i}F_{\mu j}(s_j-a_j^t)\right]_{i\in l}}_{\in O(1/\sqrt{L})} \tag{5.55}$$

We can neglect the second term (B remains finite as L diverges) as we will keep only $O(1)$ terms in the computation of the moments of the Gaussian fluctations of $\mathbf{R}_l^{t+1}$ around $\mathbf{s}_l$. This leads to the expression:

$$\mathbf{R}_l^{t+1}\approx\mathbf{s}_l+\frac{1}{B\alpha}\underbrace{\sum_{\mu}^{M}\mathbf{F}_{\mu l}\left[\sum_{k\neq l}^{L-1}\mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k-\mathbf{a}_k^t)+\xi_\mu+\Lambda_\mu^t\right]}_{:=\mathbf{r}_l^{t+1}} \tag{5.56}$$

Now we notice from (4.137) and the definition of Θ_μ in the AMP algorithm Fig. 4.6 that:

$$\Lambda_\mu^t\approx\sum_{k}^{L}\mathbf{F}_{\mu k}^{\intercal}\boldsymbol{\epsilon}_{a_{k\mu}} \tag{5.57}$$

where $\boldsymbol{\epsilon}_{a_{k\mu}}\in O(1/\sqrt{L})$ is given by (4.132), (4.137). Using the independence assumption of the operator $\mathbf{F}$ elements, we can apply the central limit theorem to $\mathbf{r}_l^{t+1}$ which is thus Gaussian distributed with moments that we compute now. We remind the reader that the noise has zero mean and we note that the $MSE(\mathbf{a},\mathbf{s})$ (3.12) tends to its average over the disorder in the large size signals limit:

$$E^{t+1}=<\left(\mathbf{s}-\mathbf{a}^{t+1}\right)^2>\underset{L\to\infty}{\to}\frac{1}{B}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\left(\mathbf{s}_l-\mathbf{a}_l^{t+1}\right)^{\intercal}\left(\mathbf{s}_l-\mathbf{a}_l^{t+1}\right)\right\} \tag{5.58}$$

where $\mathbf{a}^{t+1}=\mathbb{E}_{\mathbf{x}|\mathbf{y}}^{t+1}\{\mathbf{x}\}=\left[\mathbf{f}_{a_l}\left((\boldsymbol{\Sigma}_l^{t+1})^2,\mathbf{R}_l^{t+1}\right)\right]_l^L$ is the AMP posterior estimate of the signal at time $t+1$ and we put the index l to underline that we speak about a section, not the overall signal (which is the same for the MSE in the thermodynamic limit). The matrix $\mathbf{F}$ elements having 0 mean, only the terms with even power of the matrix elements in the various sums that appear remain because of the average over the disorder that we will use so that, using (5.57), (4.137) we obtain for its first moment:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_l^{t+1}\}=\underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu}^{M}\mathbf{F}_{\mu l}\left[\sum_{k\neq l}^{L-1}\mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k-\mathbf{a}_k^t)+\xi_\mu\right]\right\}}_{=\mathbf{0}_B}+\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu}^{M}\mathbf{F}_{\mu l}\sum_{k}^{L}\mathbf{F}_{\mu k}^{\intercal}\boldsymbol{\epsilon}_{a_{k\mu}}\right\} \tag{5.59}$$

$$\approx\underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu}^{M}(\mathbf{F}_{\mu l}^3)^{\intercal}\mathbf{v}_l\frac{y_\mu-w_\mu^t}{1/\mathrm{snr}+\Theta^t}\right\}}_{\in O(1/L)} \tag{5.60}$$

$$=\mathbf{0}_B \tag{5.61}$$

Now the cross terms, with $l' \neq l$:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_l^{t+1}\mathbf{r}_{l'}^{t+1}\} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu,\nu}^{M,M} \mathbf{F}_{\mu l}\mathbf{F}_{\nu l'} \left[\sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] \right.$$
$$\left. \left[\sum_{k'\neq l'}^{L-1} \mathbf{F}_{\nu k'}^{\intercal}(\mathbf{s}_{k'} - \mathbf{a}_{k'}^t) + \xi_\nu + \Lambda_\nu^t \right] \right\} \tag{5.62}$$
$$= \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}\mathbf{F}_{\mu l'} \left[\mathbf{F}_{\mu l'}^{\intercal}(\mathbf{s}_{l'} - \mathbf{a}_{l'}^t) + \xi_\mu + \Lambda_\mu^t \right] \right. \tag{5.63}$$
$$\left. \left[\mathbf{F}_{\mu l}^{\intercal}(\mathbf{s}_l - \mathbf{a}_l^t) + \xi_\mu + \Lambda_\mu^t \right] \right\} \tag{5.64}$$
$$= \underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 \mathbf{F}_{\mu l'}^2 (\mathbf{s}_{l'} - \mathbf{a}_{l'}^t)(\mathbf{s}_l - \mathbf{a}_l^t) \right\}}_{\in O(1/L)}$$
$$\approx \mathbf{0}_B \tag{5.65}$$

Here if the matrix elements were not i.i.d we would obtain non trivial crossed terms that would greatly complexify the analysis, as all the sections would become correlated. Finally its diagonal second moment:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{(\mathbf{r}_l^{t+1})^2\} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu,\nu}^{M,M} \mathbf{F}_{\mu l}\mathbf{F}_{\nu l} \left[\sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] \right.$$
$$\left. \left[\sum_{k'\neq l}^{L-1} \mathbf{F}_{\nu k'}^{\intercal}(\mathbf{s}_{k'} - \mathbf{a}_{k'}^t) + \xi_\nu + \Lambda_\nu^t \right] \right\} \tag{5.66}$$
$$= \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 \left[\sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] \right.$$
$$\left. \left[\sum_{k'\neq l}^{L-1} \mathbf{F}_{\mu k'}^{\intercal}(\mathbf{s}_{k'} - \mathbf{a}_{k'}^t) + \xi_\mu + \Lambda_\mu^t \right] \right\} \tag{5.67}$$
$$= \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 \left[\sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k - \mathbf{a}_k^t) \right] \left[\sum_{k'\neq l}^{L-1} \mathbf{F}_{\mu k'}^{\intercal}(\mathbf{s}_{k'} - \mathbf{a}_{k'}^t) \right] \right\}$$
$$+ \frac{\alpha B}{\mathrm{snr}} + \underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 (\Lambda_\mu^t)^2 \right\}}_{\in O(L^{-3/2})}$$
$$+ 2\underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 \Lambda_\mu^t \left[\sum_{k\neq l}^{L-1} \mathbf{F}_{\mu k}^{\intercal}(\mathbf{s}_k - \mathbf{a}_k^t) \right] \right\}}_{=\mathbf{0}_B} \tag{5.68}$$

$$\approx \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}} \left\{ \sum_{\mu}^{M} \mathbf{F}_{\mu l}^2 \left[\sum_{k \neq l}^{L-1} (\mathbf{F}_{\mu k}^2)^\intercal (\mathbf{s}_k - \mathbf{a}_k^t)^2 \right] \right\} + \frac{\alpha B}{\text{snr}} \tag{5.69}$$

$$\approx \frac{M}{L^2} \sum_{k}^{L} \mathbb{E}_{\mathbf{s}} \{ (\mathbf{s}_k - \mathbf{a}_k^t)^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) \} + \frac{\alpha B}{\text{snr}}$$

$$= B\alpha \left(1/\text{snr} + BE^t \right) \tag{5.70}$$

where $\mathbf{0}_B$ is a vector of zeros of size B and we have used (5.58) with the MSE definition (3.12) and $N = LB$. Thanks to the independence assumptions and the performed averages, we get a variance that is independent of the component and that depends only on the global MSE. From all these computations, we can now write $\mathbf{R}_l^t$ using (5.56) as a Gaussian variable, dependent on the other components only through E:

$$\mathbf{r}_l^{t+1} \sim \mathcal{N} \left(\mathbf{r}_l^{t+1} \middle| \mathbf{0}_B, B\alpha \left(1/\text{snr} + BE^t \right) \boldsymbol{I}_B \right) \tag{5.71}$$

$$\Rightarrow \mathbf{R}_l^{t+1} \sim \mathcal{N} \left(\mathbf{R}_l^{t+1} \middle| \mathbf{s}_l, \frac{1/(\text{snr}B) + E^t}{\alpha} \boldsymbol{I}_B \right) \tag{5.72}$$

where $\boldsymbol{I}_B$ is the identity of size B. It remains to perform the average over the signal $\mathbf{s}$ distribution. The equivalence between all the sections due to the prior which is the same for all of them implies that we can consider only the MSE evolution in a single section instead of the overall one. Thus the state evolution in the matching prior case, with knowledge of the channel noise is:

$$E^{t+1} = \frac{1}{B} \sum_{i \in l}^{B} \int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} \left[f_{a_i} \left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l) \right) - s_i \right]^2 \tag{5.73}$$

$$\Sigma^{t+1}(E^t) := \sqrt{\frac{1/(\text{snr}B) + E^t}{\alpha}} \tag{5.74}$$

$$\mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l) := \mathbf{s}_l + \mathbf{z}\, \Sigma^{t+1}(E^t) \tag{5.75}$$

where $\mathbf{z}$ is a i.i.d unit centered Gaussian B-d vector. In sec. 5.3.3 we will show that this recursion can be written in other equivalent ways under the prior matching condition. The definition (5.74) at its fixed point matches what we obtained in the replica computation of the Bethe free entropy (5.49).

5.3.2 Alternative derivation starting from the cavity quantities

We now re-derive the state evolution using the cavity quantities used to derive AMP, instead of starting from the final TAP equations, but the derivation sis very close. We start from the MSE (again, focusing on a unique section as they are all equivalent), which is the observable we

want to predict:

$$E^t = < (\mathbf{f}_{a_l}((\mathbf{\Sigma}_l^t)^2, \mathbf{R}_l^t) - \mathbf{s}_l)^\intercal (\mathbf{f}_{a_l}((\mathbf{\Sigma}_l^t)^2, \mathbf{R}_l^t) - \mathbf{s}_l) > \tag{5.76}$$

$$= \frac{1}{B} \mathbb{E}_{\mathbf{R}_l^t} \left\{ (\mathbf{f}_{a_l}((\mathbf{\Sigma}_l^t)^2, \mathbf{R}_l^t) - \mathbf{s}_l)^\intercal (\mathbf{f}_{a_l}((\mathbf{\Sigma}_l^t)^2, \mathbf{R}_l^t) - \mathbf{s}_l) \right\} \tag{5.77}$$

We thus need to derive the distribution of $\mathbf{R}_l^t$, defined by (4.126), $(\mathbf{\Sigma}_l^t)^2$ being its variance. Plugging the definitions (4.108) and (4.109) into (4.126) and using the fact that:

$$\sum_{k \neq l} \mathbf{v}_{k\mu} = \sum_{k \neq l} \mathbf{v}_k + O(1) \tag{5.78}$$

$$\Rightarrow \sum_{k \neq l} (\mathbf{F}_{\mu k}^2)^\intercal \mathbf{v}_{k\mu} = 1/L \sum_{k \neq l} \mathbf{v}_k + O(1/L) \tag{5.79}$$

coming from (4.133) for the first equality, and the fact that the $\mathbf{F}^2$ elements can be safely replaced by the matrix variance in the thermodynamic limit, we obtain after simple algebraic simplifications similar to the ones at the beginning of the previous section:

$$\mathbf{R}_l^t = \frac{\sum_\mu^M \mathbf{B}_{\mu l}}{\sum_\mu^M \mathbf{A}_{\mu l}} \tag{5.80}$$

$$= \mathbf{s}_l + \frac{1}{B\alpha} \underbrace{\sum_\mu^M \mathbf{F}_{\mu l} \left(\xi_\mu + \sum_{j \neq l} \mathbf{F}_{\mu j}^\intercal (\mathbf{s}_j - \mathbf{a}_{j\mu}^t) \right)}_{:= \mathbf{p}_l^t} \tag{5.81}$$

where we used the definition of the measurement rate $\alpha := M/BL$ and the relation between the measurement and the signal (3.18) to replace y_μ. From the central limit theorem and the independence assumption of the $\mathbf{F}$ elements, $\mathbf{p}_l^t$ is a Gaussian random variable. Actually by identification with (5.56), we recognize that $\mathbf{p}_l^t = \mathbf{r}_l^t$ which moments have already been computed in the previous section and the final state evolution recursion (5.73) is thus found back.

5.3.3 The prior matching condition and Bayesian optimality

Let us us now discuss some implications of the prior matching condition, i.e. of the knowledge of the true generating model of the signal. We will demonstrate relations that are true when this assumption is verified thanks to the state evolution analysis and then discuss what it implies in terms of replicas.

The prior matching conditions from the state evolution

We will now show that the prior matching condition asymptotically implies that at each time step during the reconstruction, the MSE of the posterior estimate by AMP equals the posterior variance of this estimate. We start from the state evolution results for the asymptotic MSE

through time (5.73), (5.74), (5.75). The aim is to show that it is equal to the average posterior variance defined as:

$$V^{t+1} := \frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l\, P_0(\mathbf{s}_l)\int_{\mathbb{R}^B}\mathscr{D}\mathbf{z}\, f_{c_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z},\mathbf{s}_l)\right) \tag{5.82}$$

$$= \frac{1}{B}\left[\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l^\intercal\mathbf{x}_l\}\} - \mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\}\right] \tag{5.83}$$

together with the pevious definitions (5.74), (5.75), where $\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\}$ is the asymptotic average with respect to the posterior estimate by AMP. Its dependence on the time dependent AMP fields $((\Sigma^{t+1})^2, \mathbf{R}^{t+1})$ is implicit. Furthermore, we denote by $\mathbb{E}_{\mathbf{y}}\{\} := \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\}$ the disorder average, which is the integral over $P_0(\mathbf{s})$ and $\mathscr{D}\mathbf{z}$ in (5.82). Expanding the MSE (5.73) we get:

$$E^{t+1} = \frac{1}{B}\sum_{i\in l}^{B}\Bigg[\int_{\mathbb{R}^B} d\mathbf{s}_l\, P_0(\mathbf{s}_l)\int_{\mathbb{R}^B}\mathscr{D}\mathbf{z}\, f_{a_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z},\mathbf{s}_l)\right)^2$$
$$-2\int_{\mathbb{R}^B} d\mathbf{s}_l\, P_0(\mathbf{s}_l)\, s_i\int_{\mathbb{R}^B}\mathscr{D}\mathbf{z} f_{a_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z},\mathbf{s}_l)\right) + \int_{\mathbb{R}^B} d\mathbf{s}_l\, P_0(\mathbf{s}_l)\, s_i^2\Bigg] \tag{5.84}$$

$$= \frac{1}{B}\left[\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} - 2\mathbb{E}_{\mathbf{y}}\{\mathbf{s}_l^\intercal\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} + \mathbb{E}_{\mathbf{s}}\{\mathbf{s}_l^\intercal\mathbf{s}_l\}\right] \tag{5.85}$$

We start by proving the equality between the first and second term (up to the 2). From now on, we skeep the time index for sake of readibility. Using the definition of the denoiser f_{a_i} (4.115), we get:

$$\int d\mathbf{s}_l\, P_0(\mathbf{s}_l)\int\mathscr{D}\mathbf{z}\, f_{a_i}\left((\Sigma)^2, \mathbf{R}(\mathbf{z},\mathbf{s}_l)\right)^2$$
$$= \int d\mathbf{s}_l\, P_0(\mathbf{s}_l)\int\mathscr{D}\mathbf{z}\,\frac{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal(\mathbf{s}_l+\mathbf{z}\Sigma)}{\Sigma^2}\right)x_i\right]^2}{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal(\mathbf{s}_l+\mathbf{z}\Sigma)}{\Sigma^2}\right)\right]^2} \tag{5.86}$$

where we have expanded the square in the exponent and simplified the $\mathbf{x}_l$ independent term with its normalization one in the denoiser expression. Now we use the following change of variable: $\mathbf{z}' := (\mathbf{s}_l + \mathbf{z}\Sigma)$:

$$= \int\mathscr{D}\mathbf{z}'\int d\mathbf{s}_l\, P_0(\mathbf{s}_l)\exp\left(\frac{\mathbf{s}_l^\intercal\mathbf{z}'}{\Sigma} - \frac{\mathbf{s}_l^\intercal\mathbf{s}_l}{2\Sigma^2}\right)\frac{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal\mathbf{z}'}{\Sigma}\right)x_i\right]^2}{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal\mathbf{z}'}{\Sigma}\right)\right]^2} \tag{5.87}$$

$$= \int\mathscr{D}\mathbf{z}'\,\frac{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal\mathbf{z}'}{\Sigma}\right)x_i\right]^2}{\left[\int d\mathbf{x}_l\, P_0(\mathbf{x}_l)\exp\left(-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal\mathbf{z}'}{\Sigma}\right)\right]} \tag{5.88}$$

Now the second term of (5.84) using the same change of variable:

$$\int \mathscr{D}\mathbf{z}' \int d\mathbf{s}_l \, P_0(\mathbf{s}_l) \, s_i \exp\left(\frac{\mathbf{s}_l^\intercal \mathbf{z}'}{\Sigma} - \frac{\mathbf{s}_l^\intercal \mathbf{s}_l}{2\Sigma^2}\right) \frac{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right) x_i\right]}{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right)\right]} \tag{5.89}$$

$$= \int \mathscr{D}\mathbf{z}' \frac{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right) x_i\right]^2}{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right)\right]} \tag{5.90}$$

So the two terms are equal:

$$\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} = \mathbb{E}_{\mathbf{y}}\{\mathbf{s}_l^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} \tag{5.91}$$

This allows to re-write the MSE as:

$$E^{t+1} = \frac{1}{B}\sum_{i\in l}^{B}\left[\int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) s_i^2 - \int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} \, f_{a_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l)\right)^2\right] \tag{5.92}$$

$$= \mathbb{E}_{\mathbf{s}}\{\mathbf{s}_l^\intercal \mathbf{s}_l\} - \mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} \tag{5.93}$$

Let us now show that under the prior matching condition, $\mathbb{E}_{\mathbf{s}}\{s_i^2\} = \mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{x_i^2\}\}$ which will complete the proof. We do so by using again the same change of variable:

$$\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{x_i^2\}\} = \int d\mathbf{s}_l \, P_0(\mathbf{s}_l) \int \mathscr{D}\mathbf{z} \frac{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal (\mathbf{s}_l + \mathbf{z}\Sigma)}{\Sigma^2}\right) x_i^2\right]}{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal (\mathbf{s}_l + \mathbf{z}\Sigma)}{\Sigma^2}\right)\right]} \tag{5.94}$$

$$= \int \mathscr{D}\mathbf{z}' \int d\mathbf{s}_l \, P_0(\mathbf{s}_l) \exp\left(\frac{\mathbf{s}_l^\intercal \mathbf{z}'}{\Sigma} - \frac{\mathbf{s}_l^\intercal \mathbf{s}_l}{2\Sigma^2}\right) \frac{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right) x_i^2\right]}{\left[\int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right)\right]} \tag{5.95}$$

$$= \int \mathscr{D}\mathbf{z}' \int d\mathbf{x}_l \, P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2} + \frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma}\right) x_i^2 \tag{5.96}$$

$$= \int d\mathbf{x}_l P_0(\mathbf{x}_l) \exp\left(-\frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma^2}\right) x_i^2 \int d\mathbf{z}' \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mathbf{x}_l^\intercal \mathbf{z}'}{\Sigma} - \frac{(\mathbf{z}')^\intercal \mathbf{z}'}{2}\right) \tag{5.97}$$

$$= \int d\mathbf{x}_l P_0(\mathbf{x}_l) x_i^2 \tag{5.98}$$

$$= \mathbb{E}_{\mathbf{s}}\{s_i^2\} \tag{5.99}$$

where the last equality is true due to the matching prior condition. This with (5.83) and (5.93) implies that $E^t = V^t \; \forall \; t$. Of course this is true only if the two quantities are initialized with same value, but as long as they are, they remain equal at any time step: we say the the algorithm remains on the Nishimori line. We summarize the diverse relations that are true

under the prior matching condition:

$$E^t = V^t \ \forall \ t \tag{5.100}$$

$$\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} = \mathbb{E}_{\mathbf{y}}\{\mathbf{s}_l^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} \tag{5.101}$$

$$\mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l^\intercal \mathbf{x}_l\}\} = \mathbb{E}_{\mathbf{s}}\{\mathbf{s}_l^\intercal \mathbf{s}_l\} \tag{5.102}$$

To summarize, this implies that the three following forms of the state evolution are perfectly equivalent under the prior matching condition:

$$E^{t+1} = V^{t+1} = \frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} \left[f_{a_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l)\right) - s_i \right]^2 \tag{5.103}$$

$$= \frac{1}{B}\left[\mathbb{E}_{\mathbf{s}}\{\mathbf{s}_l^\intercal \mathbf{s}_l\} - \sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) s_i \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} \, f_{a_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l)\right) \right] \tag{5.104}$$

$$= \frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l \, P_0(\mathbf{s}_l) \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} \, f_{c_i}\left((\Sigma^{t+1})^2, \mathbf{R}^{t+1}(\mathbf{z}, \mathbf{s}_l)\right) \tag{5.105}$$

together with (5.74) and (5.75). One form can be more practical to use depending on the denoising functions. The second form (5.104) can be computationally easier and faster to deal with but can be more dangerous to use than the two other forms of the state evolution because it can become negative if the difference is really small due to finite numerical precision.

The prior matching conditions in terms of replicas

Let us discuss the physical meaning of the macroscopic order parameters introduced in the replica computation and make the connection between these and the equalities obtained thanks to the Nishimori condition (5.100), (5.101), (5.102).

The replicas are interpreted as different possible solutions (microstates) to the noisy problem (3.18), the fluctuations coming from the disorder. Using the replica symmetric ansatz is assuming that all the replicas belong to the same pure state, i.e. the space of solutions does not split into many disconnected components and thus each replicas have same statistical properties, given by the replica macroscopic order parameters.

• The meaning of m (5.13) in the replica symmetric ansatz is the overlap between the signal and the signal estimate given by the average over all the replica states. We can thus identify m with $m = \mathbb{E}_{\mathbf{y}}\{\mathbf{s}_l^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\}$.

• The order parameter Q (5.14) is the average selfoverlap of the replicas and is thus naturally identified with $Q = \mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l^\intercal \mathbf{x}_l\}\} = \mathbb{E}_{\mathbf{s}}\{\mathbf{s}_l^\intercal \mathbf{s}_l\}$.

• Finally, q (5.15) is the overlap between different replicas. In the phase where reconstruction is possible as enough information about the signal is contained in the measurement, the differences between the different replica states should be dominated by the noise, and the

features present in all the replica states is the true information on the signal. So the overlap q between infinitely large replicas cancels out the noise-induced fluctuations in average, and remains only the average over the replicas to the square, i.e. the squared signal estimate. Thus $q = \mathbb{E}_{\mathbf{y}}\{\mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}^\intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} = \mathbb{E}_{\mathbf{y}}\{\mathbf{s}_l \intercal \mathbb{E}_{\mathbf{x}|\mathbf{y}}\{\mathbf{x}_l\}\} = m$.

The second equalities have been obtained thanks to the Nishimori conditions (5.100), (5.101) and (5.102).

5.4 The link between replica and state evolution analyzes: derivation of the state evolution from the average Bethe free entropy

We now show that the fixed point conditions of the Bethe free entropy computed by the replica method are giving back the state evolution recursion of the MSE at its fixed point. We restrict ourselves to the section independent prior case but the derivation for generic prior is done in a similar manner.

As we place ourselves under the matching prior condition, the previous section sec. 5.3.3 have shown that the conditions assumed in sec. 5.2.3 are true. Thus starting from the replica potential expression (5.38) at its fixed point with respect to all its parameters, we should be able to derive the state evolution, for example the form (5.104). By identification of (5.104) with (5.45), we should find that the fixed point of m gives the integral part of (5.104). Let us prove it. The fixed point condition for m around the optimal values of all the parameters of the Bethe free energy gives:

$$\left.\frac{\partial\tilde{\Phi}_B}{\partial\hat{m}}\right|_{(\hat{m}^*,\hat{q}^*,\hat{Q}^*,m^*,q^*,Q^*)} = 0 \tag{5.106}$$

$$\begin{aligned}\Rightarrow m^*(E) &= \int d\mathbf{s}_l \mathscr{D}\mathbf{z} P_0(\mathbf{s}_l) \int d\mathbf{x}_l P_0(\mathbf{x}_l) \frac{\mathbf{s}_l^\intercal \mathbf{x}_l}{Z(\mathbf{s}_l,\mathbf{z},E)} \exp\left\{\frac{\mathbf{x}_l^\intercal \mathbf{s}_l}{\Sigma(E)^2} + \frac{\mathbf{z}^\intercal \mathbf{x}_l}{\Sigma(E)} - \frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma(E)^2}\right\} \\ &= \sum_{i\in l}^{B} \int d\mathbf{s}_l \mathscr{D}\mathbf{z} P_0(\mathbf{s}_l) s_i \int d\mathbf{x}_l P_0(\mathbf{x}_l) x_i \frac{1}{Z(\mathbf{s}_l,\mathbf{z},E)} \exp\left\{\frac{\mathbf{x}_l^\intercal \mathbf{s}_l}{\Sigma(E)^2} + \frac{\mathbf{z}^\intercal \mathbf{x}_l}{\Sigma(E)} - \frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma(E)^2}\right\}\end{aligned} \tag{5.107}$$

using (5.48), (5.49) and where $Z(\mathbf{s}_l,\mathbf{z},E)$ is a partition function:

$$Z(\mathbf{s}_l,\mathbf{z},E) := \int d\mathbf{x}_l P_0(\mathbf{x}_l) \exp\left\{\frac{\mathbf{x}_l^\intercal \mathbf{s}_l}{\Sigma(E)^2} + \frac{\mathbf{z}^\intercal \mathbf{x}_l}{\Sigma(E)} - \frac{\mathbf{x}_l^\intercal \mathbf{x}_l}{2\Sigma(E)^2}\right\} \tag{5.108}$$

We recognize the expression of the denoiser f_{a_i} (4.115) when we use the definitions (5.74) and

(5.75):

$$f_{a_i}(\Sigma^2,\mathbf{R}) = \frac{1}{\tilde{Z}(\Sigma^2,\mathbf{R})}\int d\mathbf{x}_l x_i P_0(\mathbf{x}_l)\exp\left\{-\frac{(\mathbf{x}_l-\mathbf{R})^\intercal(\mathbf{x}_l-\mathbf{R})}{2\Sigma^2}\right\} \tag{5.109}$$

$$= \frac{1}{\tilde{Z}(\mathbf{s}_l,\mathbf{z},E)}\int d\mathbf{x}_l x_i P_0(\mathbf{x}_l)\exp\left\{-\frac{(\mathbf{x}_l-\mathbf{s}_l-\mathbf{z}\Sigma)^\intercal(\mathbf{x}_l-\mathbf{s}_l-\mathbf{z}\Sigma)}{2\Sigma^2}\right\} \tag{5.110}$$

$$= \frac{1}{Z(\mathbf{s}_l,\mathbf{z},E)}\int d\mathbf{x}_l x_i P_0(\mathbf{x}_l)\exp\left\{\frac{\mathbf{x}_l^\intercal\mathbf{s}_l}{\Sigma(E)^2}+\frac{\mathbf{z}^\intercal\mathbf{x}_l}{\Sigma(E)}-\frac{\mathbf{x}_l^\intercal\mathbf{x}_l}{2\Sigma(E)^2}\right\} \tag{5.111}$$

From this together with (5.107) we get that at its fixed point, the replica order parameter m verifies:

$$m^* = \sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l P_0(\mathbf{s}_l) s_i \int_{\mathbb{R}^B}\mathscr{D}\mathbf{z} f_{a_i}(\Sigma(E)^2,\mathbf{R}(\mathbf{z},\mathbf{s}_l)) \tag{5.112}$$

which is exactly the integral of (5.104). Thus using (5.45) we get:

$$E = <\mathbf{s}^2> -\frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l P_0(\mathbf{s}_l) s_i \int_{\mathbb{R}^B}\mathscr{D}\mathbf{z} f_{a_i}(\Sigma(E)^2,\mathbf{R}(\mathbf{z},\mathbf{s}_l)) \tag{5.113}$$

which is exactly the expression of the state evolution (5.104) as the empirical and true averages are equal in the thermodynamic limit. Thus the fixed point conditions verified at the optimum of the Bethe free entropy gives back the state evolution at its fixed point (when the time index is droped).

From this analysis, we can now assert that using the state evolution analysis to find fixed points of the algorithm or extracting this information from the potential (5.48) are totally equivalent. In addition, this validates further the replica analysis as an exact procedure in the present context. The state evolution thus finds the optima of the potential $\Phi_B(E)$ and thus when it has two local maxima the fixed point of the state evolution depends on the initial condition of the recursion.

5.4.1 Different forms of the Bethe free entropy for different fixed points algorithms

It is worth noticing that the previous derivation is the asymptotic equivalent of the derivation in sec. 4.2.7 of the belief propagation algorithm as fixed point equations extracted from the Bethe free energy on a single instance (4.83). Actually depending on the limit taken and the graph topology, we get different expressions of the Bethe free energy, and thus different fixed point equations and message-passings.

• When using the Bethe free entropy/energy parametrized by the cavity messages (4.83), justified on locally tree-like graphs, the fixed points equations directly give the canonical belief propagation for a single graph instance, see sec. 4.2.7.

• Now in the case of large densely connected graphs, the Bethe free entropy becomes (4.194)

and deriving its fixed point equations would lead to the approximate message-passing algorithm, that has been instead directly derived from belief propagation in sec. 4.3. This free entropy and algorithm are dependent on the disorder instance.

• Finally, when the dense graph is really taken to be infinitely large (or equivalently averaged over the disorder by the self averageness property), the Bethe free entropy is the one extracted from the replica analysis (5.48) and the associated fixed point algorithm is the state evolution, the asymptotic AMP.

One could think also about a Bethe free entropy for tree-like graphs, but averaged over the disorder. In this case the fixed point equations predicting the BP algorithm behavior on infinite graphs is referred as density evolution, the equivalent of the state evolution for AMP.

5.5 Spatial coupling for optimal inference in linear estimation

As already discussed, mean-field systems (systems for which mean-field techniques such as message-passing algorithms are appropriate) are of two types: the problems defined on sparse random graphs due to their tree-like property, such as the independent set [1] or many other combinatorial optimization problems and problems defined on densely connected graphs such as in the present case (3.18). As discussed in sec. 4.3.2 these systems are equivalent to infinite dimensional homogeneous systems. Here nucleation (a local change of thermodynamical phase) cannot occur: an infinite number of dimensions implies that there is no notion of locality in these systems and thus no nucleus can spread as any apparition of a different phase nucleus inside another one has an infinite energy cost, as the "surface" of the nucleus is itself infinite.

As discussed in sec. 5.1.1, in inference the phases we are dealing with are computational: easy/hard/impossible inference phases. In order to have nucleation of an easy phase inside an hard one and to allow this nucleus to propagate inside the full system, one has to introduce a dimensionality, or structure in the problem. This is done by spatial coupling using a properly designed coding operator, see Fig. 5.4. It mimics the strategy employed by the nature. We use again the example of supercooled water which is blocked in the metastable liquid state by a first order transition despite it is below its critical temperature. If a nucleus of crystal appears somewhere in the fluid, the surface between the two phases has an energetic cost $C_1 2\pi R^2$ where R is the radius of the nucleus that we consider spheric. But as the system is 3-d, this cost remains always finite. Now if the nucleus is big enough, the reduction in energy by $C_2 4/3\pi R^3$ due to the bulk of the small crystal nucleus (the true equilibrium state at this temperature) counterbalances the surface energy term and the entropy loss $\Delta S < 0$ due to the crystal: $\Delta F = F_{with\ nucleus} - F_{no\ nucleus} = C_1 2\pi R^2 - C_2 4/3\pi R^3 - T\Delta S < 0$ and the nucleus spreads in the overall system which finally reaches its true equilibrium macrostate. $C_1, C_2 > 0$ are constants that depend on the microscopic physical interactions between atoms.

Spatial coupling thanks to which the theoretical optimal thresholds can be saturated was de-

veloped in error-correcting codes [107–109] and has been extensively used in the compressed sensing setting as well [34, 35, 90, 110]. It rigorously allows to reach the information theoretical bound in LDPC codes [108] and in compressed sensing in the random i.i.d measurement matrix case [108, 110] as we will see. Rigorous proofs of this was worked out in [110]. The robustness to measurement noise was also discussed in [35, 110]. In addition, spatial coupling is used as a proof technique for understanding properties of uncoupled ensembles [111]. Furthermore, it is applicable in a very wide range of graphical models and allows to asymptotically solve constraint satisfaction problems until their satisfability threshold [112, 113], the last point in the phase diagram until which it is theoretically possible to find a solution to the problem, see sec. 5.1.1. The spatial coupling strategy is conjectured efficient to solve a problem only if a first order phase transition is present, which is the case in a very large class of interesting problems.

5.5.1 The spatially-coupled operator and the reconstruction wave propagation

Let us now describe how the spatial coupling is implemented in the context of sparse linear estimation through the measurement operator construction. The spatial structure decribed above is induced in the signal by the block structure of the measurement (or coding) matrix, see Fig. 5.4. Other designs are possible [35, 114] but in the present thesis, spatially-coupled operators will always be of the form Fig. 5.4 as they are empirically very efficient. Looking at Fig. 5.4, we clearly see that if we consider only the diagonal blocks on the matrix, from the measurements point of view the signal becomes the concatenation of independent subsystems $\mathbf{s} = [\mathbf{s}_c]_c^{L_c}$ where L_c is the number of such blocks. But the matrix has also blocks on the the upper and some lower diagonals as well: these couple the different sub-systems $\{\mathbf{s}_c\}_c^{L_c}$. The blocks on the lower diagonals couple the signal block $\mathbf{s}_c$ with the w previous ones $\{\mathbf{s}_{c-i},\}_i^w$, where w is called the coupling window. They have elements with statistically the same amplitude as the diagonal blocks. The upper diagonal blocks weakly couple the block $\mathbf{s}_c$ with the next one $\mathbf{s}_{c+1}$. This forward coupling strength is tuned by the J parameter, the variance of the elements of the random values inside the block. The matrix structure is thus fully encoded through a $L_r \times L_c$ matrix $\mathbf{J}$ with element $J_{r,c}$ giving the (non rescaled by $1/L$) variance $\in O(1)$ of the block at the l^{th} block-row, c^{th} block-column of the matrix.

Let us assume that we want to solve a compressed sensing instance generated in the hard phase, where the equilibrium is given by a low MSE estimate and reconstruction should succeed from a thermodynamic point of view. Unfortunately, the algorithm is stuck in the metastable high MSE state by the first order BP transition. The nucleus of "crystal", i.e. of easy phase called the *seed*, is induced by the matrix first block-row: as seen on Fig. 5.4, these blocks are closer to square than the next block-rows, i.e. they have a higher effective measurement rate $\alpha_{seed} > \alpha_{rest}$, where α_{rest} is the effective measurement rate of the next block-rows which can be asymptotically as small as the Bayes optimal measurement rate α_{opt}. As the forward coupling is relatively weak, the first measurements represented by the darker grey part of the measurement vector in Fig. 5.4 contain essentially information about

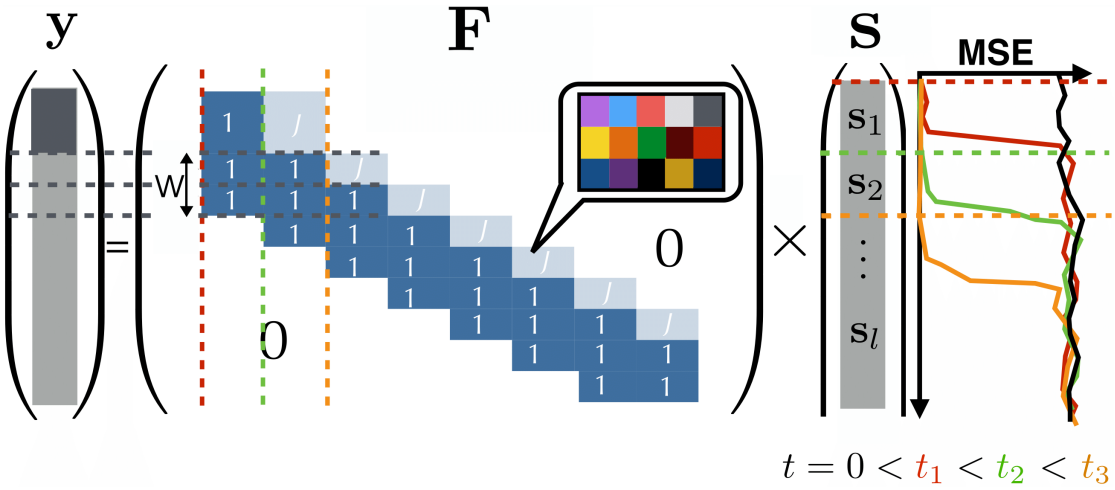

Figure 5.4 – Representation of the spatially-coupled sensing matrix used in this thesis. The operator is decomposed in $L_r \times L_c$ blocks, each being made of N/L_c columns and $\alpha_{seed} N/L_c$ lines for the blocks of the first block-row, $\alpha_{rest} N/L_c$ lines for the following block-rows (these follow from the definition of $\alpha := M/N$ combined with (5.114)), with $\alpha_{seed} > \alpha_{BP} \geq \alpha_{rest} > \alpha_{opt}$. The figure shows that each block is itself a small random operator with i.i.d elements. There is a number w (the coupling window) of lower diagonal blocks with elements of variance 1 as the diagonal blocks, the upper diagonal blocks have elements of variance $J < 1$ where $\sqrt{J}$ is the coupling strength, all the other blocks contain only zeros. The matrix structure is thus encoded through a $L_r \times L_c$ matrix $\mathbf{J}$ with element $J_{r,c}$ giving the variance $\in O(1)$ of the block at the l^{th} block-row and c^{th} block-column. The colored dotted lines help to visualize the block decomposition of the signal $\mathbf{s} = [\mathbf{s}_c]_c^{L_c}$ induced by the operator structure. Each block of the signal will be reconstructed at different times in the algorithm, as depicted by the right part of the plot which shows how the wave of reconstruction propagates though time, see the color code behind: at $t = 0$, the MSE is homogeneously high (black curve) and as time goes on, blocks are reconstructed from the seed until the end of the signal. The parameters that define the spatially-coupled operator ensemble are $(L_c, L_r, w, \sqrt{J}, \alpha_{seed}, \alpha_{rest})$. In the thesis, the matrix elements are rescaled by $1/\sqrt{L}$ such that the variances are rescaled by $1/L$ to enforce the measurements to be $\in O(1)$.

the components of the signal inside its first block $\mathbf{s}_1$. As we enforce $\alpha_{seed} > \alpha_{BP}$, the "signal seed" $\mathbf{s}_1$ is easily reconstructed by message-passing and then this new information which is coupled to the next signal block helps the reconstruction of this neighboring $\mathbf{s}_2$ by increasing its effective measurement rate. This triggers a wave of reconstruction from the top of the signal where the seed is until its end. This is represented on Fig. 5.4: initially the MSE is homogeneously high in all the signal (black curve on the MSE plot on the right of the signal), and as the time steps increase, more and more blocks are reconstructed until the signal is fully reconstructed. [115, 116] present detailed studies of the spatial coupling in compressed sensing and the dynamical properties of the reconstruction wave front.

One has to be careful to ensure that these sub-systems $\{\mathbf{s}_c\}_c^{L_c}$ remain large enough for the assumptions behind the approximate message-passing to be valid inside each of them: each

$\mathbf{s}_c = [\mathbf{s}_l]_{l\in c}$ must itself be a mean-field system, which is true if $L \gg L_c$. Concurrently, however, the larger L_c, the better it is to get closer to the optimal treshold α_{opt}. This is due to the fact that the overall measurement rate α is a weighted average of the effective measurement rate of the seed block α_{seed} and that of the remaining ones α_{rest}:

$$\alpha_{rest} = \frac{\alpha L_c - \alpha_{seed}}{L_r - 1} := \alpha\left(\frac{L_c - \beta_{seed}}{L_r - 1}\right) \tag{5.114}$$

In practice, α is fixed and $\alpha_{seed} := \alpha\beta_{seed}$ as well as by fixing β_{seed}. α_{rest} is then deduced from (5.114). In the remaining of this thesis, we will define the spatially-coupled ensemble by $(L_c, L_r, w, \sqrt{J}, \alpha, \beta_{seed})$ or $(L_c, L_r, w, \sqrt{J}, \alpha_{seed}, \alpha_{rest})$ equivalently. This relation (5.114) is equivalent to:

$$\alpha = \frac{(L_r - 1)\alpha_{rest} + \alpha_{seed}}{L_c} \underset{L_c, L_r \to \infty}{\to} \alpha_{rest} > \alpha_{opt} \tag{5.115}$$

where α_{rest} can asymptotically be as small as α_{opt}. Thus spatial coupling allows to *asymptotically* reach the optimal transition, the information theoretic limit of reconstruction. In this way, the gap between α_{opt} and α_{BP}, the hard phase discussed in sec. 5.1.1 is filled and AMP becomes asymptotically Bayes optimal for any $\alpha > \alpha_{opt}$.

In the case of vectorial components, one may be careful to induce the blocks such that the sections are not "cut" by the induced structure, but it appears empirically that it does not change anything, so we forget about this detail in the algorithm implementation.

5.6 The spatially-coupled approximate message-passing algorithm

We define $\mathbf{e}_c$ with $c \in \{1, \dots, L_c\}$, a vector of size N/L_c, as the c^{th} block of $\mathbf{e}$ (of size N) and $\mathbf{f}_r$ with $r \in \{1, \dots, L_r\}$, a vector of size $\alpha_r N/L_c$ as the r^{th} block of $\mathbf{f}$ (of size M). For example, in Fig. 5.4, the signal $\mathbf{s}$ is decomposed as $\mathbf{s} = [\mathbf{s}_c]_c^{L_c}$. The notation $i \in c$ (resp. $\mu \in r$) means all the components of $\mathbf{e}$ that are in $\mathbf{e}_c$ (resp. all the components of $\mathbf{f}$ that are in $\mathbf{f}_r$). The algorithm (for scalar signals and matrices, see Fig. 7.2 for the complex case) requires four different operators performing the following operations:

$$\tilde{O}_\mu(\mathbf{e}_c) := \sum_{i\in c}^{N/L_c} F_{\mu i}^2 e_i \tag{5.116}$$

$$O_\mu(\mathbf{e}_c) := \sum_{i\in c}^{N/L_c} F_{\mu i} e_i \tag{5.117}$$

$$\tilde{O}_i(\mathbf{f}_r) := \sum_{\mu\in r}^{\alpha_r N/L_c} F_{\mu i}^2 f_\mu \tag{5.118}$$

$$O_i(\mathbf{f}_r) := \sum_{\mu\in r}^{\alpha_r N/L_c} F_{\mu i} f_\mu \tag{5.119}$$

$t \leftarrow 0$
$\delta \leftarrow \epsilon + 1$
while $t < t_{max}$ **and** $\delta > \epsilon$ **do**
  $\Theta_\mu^{t+1} \leftarrow \sum_c^{L_c} \tilde{O}_\mu(\mathbf{v}_c^t)$
  $w_\mu^{t+1} \leftarrow \sum_c^{L_c} O_\mu(\mathbf{a}_c^t) - \Theta_\mu^{t+1} \frac{y_\mu - w_\mu^t}{1/\mathrm{snr} + \Theta_\mu^t}$
  $\Sigma_i^{t+1} \leftarrow \left[\sum_r^{L_r} \tilde{O}_i \left([1/\mathrm{snr} + \mathbf{\Theta}_r^{t+1}]^{-1} \right) \right]^{-1/2}$
  $R_i^{t+1} \leftarrow a_i^t + (\Sigma_i^{t+1})^2 \sum_r^{L_r} O_i \left(\frac{\mathbf{y}_r - \mathbf{w}_r^{t+1}}{1/\mathrm{snr} + \mathbf{\Theta}_r^{t+1}} \right)$
  $v_i^{t+1} \leftarrow f_{c_i} \left((\mathbf{\Sigma}_{l_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1} \right)$
  $a_i^{t+1} \leftarrow f_{a_i} \left((\mathbf{\Sigma}_{l_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1} \right)$
  $t \leftarrow t + 1$
  $\delta \leftarrow ||\mathbf{a}^{t+1} - \mathbf{a}^t||_2^2$
end while
return $\mathbf{a}^t$

Figure 5.5 – The AMP algorithm written with operators. Here for example $\mathbf{w}_r^{t+1} := [w_\mu^{t+1}]_{\mu \in r}$ and $\mathbf{a}_c^t := [\mathbf{a}_l^t]_{l \in c}$. This form underlines how AMP is operating when spatially-coupled operators are used instead of homogeneous matrices and takes advantage from this structure. l_i is the index of the section to which the i^{th} 1-d variable belongs to. ϵ is the accuracy for convergence and t_{max} the maximum number of iterations. A suitable initialization for the quantities is ($a_i^{t=0} = \mathbb{E}_{P_0}(x)$, $v_i^{t=0} = \mathrm{Var}_{P_0}(x)$, $w_\mu^{t=0} = y_\mu$). Once the algorithm has converged, i.e. the quantities do not change anymore from iteration to iteration, the estimate of the l^{th} signal section is $\mathbf{a}_l^t$. If needed, the damping scheme of Fig. 4.6 can be used.

α_r is the measurement rate of all the blocks at the r^{th} block-row, for example on Fig. 5.4, $\alpha_1 = \alpha_{seed}$ and $\alpha_j = \alpha_{rest}\ \forall\ j > 1$.

This version of AMP is perfectly equivalent to Fig. 4.6 but underlines the spatially-coupled structure of the matrix. Furthermore, it will be useful when defining the structured operators making use of fast transforms such as Hadamard and Fourier operators, see chap. 7. Actually, even in the case where no fast transforms are used, this form of AMP can be advantageous as the matrix is sparse, and thus it can avoid the useless time consuming products with the many zeros in the matrix.

5.6.1 Further simplifications for random matrices with zero mean and equivalence with Montanari's notations

We can go further in the approximation of some quantities computed by the algorithm Fig. 5.5 by considering that the large signal limit allows to replace the elements $F_{\mu i}^2$ by the matrix variance $F_{\mu i}^2 \approx J_{r_\mu, c_i}/L$. This allows to derive the so called full-TAP equations for AMP. This is justified by the fact that the average with respect to the matrix realization of all the quantities appearing in the algorithm depending on such squared elements (such as Θ_μ) are $\in O(1)$ whilst their variance $\in O(1/N)$, see [35]. Thus in the large signal limit, we can neglect their instance-

$t \leftarrow 0$
$\delta \leftarrow \epsilon + 1$
while $t < t_{max}$ **and** $\delta > \epsilon$ **do**
  $\Theta_r^{t+1} \leftarrow \sum_c^{L_c} \tilde{O}_r(\mathbf{v}_c^t)$
  $w_\mu^{t+1} \leftarrow \sum_c^{L_c} O_\mu(\mathbf{a}_c^t) - \Theta_{r_\mu}^{t+1} \frac{y_\mu - w_\mu^t}{1/\mathrm{snr} + \Theta_{r_\mu}^t}$
  $\Sigma_c^{t+1} \leftarrow \left[\sum_r^{L_r} \tilde{O}_c\left([1/\mathrm{snr} + \Theta_r^{t+1}]^{-1}\right)\right]^{-1/2}$
  $R_i^{t+1} \leftarrow a_i^t + (\Sigma_{c_i}^{t+1})^2 \sum_r^{L_r} O_i\left(\frac{\mathbf{y}_r - \mathbf{w}_r^{t+1}}{1/\mathrm{snr} + \Theta_r^{t+1}}\right)$
  $v_i^{t+1} \leftarrow f_{c_i}\left((\Sigma_{c_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1}\right)$
  $a_i^{t+1} \leftarrow f_{a_i}\left((\Sigma_{c_i}^{t+1})^2, \mathbf{R}_{l_i}^{t+1}\right)$
  $t \leftarrow t + 1$
  $\delta \leftarrow ||\mathbf{a}^{t+1} - \mathbf{a}^t||_2^2$
end while
return $\mathbf{a}^t$

Figure 5.6 – The simplified (with respect to Fig. 5.5) full-TAP AMP algorithm, where we have approximated the squared elements of the matrix by the its variance. For example in Fig. 5.5, $\mathbf{\Theta}_r^{t+1} := [\Theta_\mu^{t+1}]_{\mu \in r}$ was a vector of measure-index dependent components whereas now Θ_r^{t+1} is a scalar with same value $\forall \mu \in r$. Notations must not be confused: l_i is the index of the unique section to which the 1-d component i belongs to, whereas c_i is the index of the block to which the section l_i belongs to, etc. If needed, the damping scheme of Fig. 4.6 can be used.

dependent fluctuations using this simplification. Considering the most general version of AMP Fig. 5.5 where it is written in terms of the operators, the dependency in squared matrix elements is just in the $\tilde{O}_\mu$ (5.116) and $\tilde{O}_i$ (5.118) operators which can thus be approximated as:

$$\tilde{O}_\mu(\mathbf{e}_c) \approx \tilde{O}_{r_\mu}(\mathbf{e}_c) := \frac{J_{r_\mu,c}}{L} \sum_{i \in c}^{N/L_c} e_i \tag{5.120}$$

$$\tilde{O}_i(\mathbf{f}_r) \approx \tilde{O}_{c_i}(\mathbf{f}_r) := \frac{J_{r,c_i}}{L} \sum_{\mu \in r}^{\alpha_r N/L_c} f_\mu \tag{5.121}$$

They now depend only on the block indices and thus also Θ_r and Σ_c that are derived from them. From this, we can write a simplified form for the AMP algorithm Fig. 5.6.

Equivalence with Montanari's notations

Now we show how to go from this simplified algorithm to the equivalent notation of Montanari [117] in the case of an homogeneous matrix. Starting from Fig. 5.6, we plug the AMP field R_i^{t+1}

expression into the denoiser:

$$a_i^{t+1} = f_{a_i}\left((\Sigma_{c_i}^{t+1})^2, \underbrace{\left[a_j^t + (\Sigma_{c_i}^{t+1})^2 \sum_r^{L_r} O_i\left(\frac{\boldsymbol{\tau}_r^t}{1/\text{snr} + \Theta_r^{t+1}}\right)\right]_{j\in l_i}}_{:=\mathbf{R}_{l_i}^{t+1}}\right) \tag{5.122}$$

where we define the residual $\boldsymbol{\tau}_r^t := \mathbf{y}_r - \mathbf{w}_r^{t+1} = \left[\tau_\mu^t = y_\mu - w_\mu^{t+1}\right]_{\mu\in r}$ and we have defined the blocks such that all the 1-d components inside the same section are in the same block for the derivation. Now using the iteration of $\mathbf{w}_r^{t+1}$ in Fig. 5.5 we get:

$$\boldsymbol{\tau}_r^t = \mathbf{y}_r - \left[\sum_c^{L_c} O_\mu(\mathbf{a}_c^t)\right]_{\mu\in r} + \Theta_r^{t+1}\frac{\mathbf{y}_r - \mathbf{w}_r^t}{1/\text{snr} + \Theta_r^t} \tag{5.123}$$

$$= \mathbf{y}_r - \left[\sum_c^{L_c} O_\mu(\mathbf{a}_c^t)\right]_{\mu\in r} + \frac{\Theta_r^{t+1}\boldsymbol{\tau}_r^{t-1}}{1/\text{snr} + \Theta_r^t} \tag{5.124}$$

Using the definition of Θ_r^{t+1} from the algorithm Fig. 5.6 together with (5.120) we obtain:

$$\Theta_r^{t+1} = \frac{B}{L_c}\sum_c^{L_c} J_{r,c} < f_c^t >_c \tag{5.125}$$

$$= \frac{B}{L_c}\sum_c^{L_c} J_{r,c}(\Sigma_c^t)^2 < (f_a^t)' >_c \tag{5.126}$$

where we have used the property (4.142) of the denoising function for the last equality and we define the shorthand notations $<>_c$ of the empirical average restricted to one block c :

$$< f_c^t >_c := \frac{L_c}{N}\sum_{i\in c}^{N/L_c} f_{c_i}\left((\Sigma_c^t)^2, \mathbf{R}_{l_i}^t\right) \tag{5.127}$$

$$< (f_a^t)' >_c := \frac{L_c}{N}\sum_{l\in c}^{L/L_c}\sum_j^B \left.\frac{\partial f_{a_{l(j)}}(x,\mathbf{y})}{\partial y_j}\right|_{(\Sigma_c^t)^2, \mathbf{R}_l^t} \tag{5.128}$$

where $l(j) \in \{1,\dots,N\}$ is the index of the j^{th} 1-d variable belonging to the section l, where $j \in \{1,\dots,B\}$. From here we can show that this form gives back the Montanari's one in the homogeneous operator case ($L_c = L_r = J_{r,c} = 1$). In this case, the quantities in the algorithm become:

$$\Theta^{t+1} = B(\Sigma^t)^2 < (f_a^t)' > \tag{5.129}$$

$$(\Sigma^{t+1})^2 = \left[\frac{1}{L}\sum_\mu^{\alpha LB}\frac{1}{1/\text{snr} + \Theta^{t+1}}\right]^{-1} \tag{5.130}$$

$$= \frac{\Theta^{t+1} + 1/\text{snr}}{B\alpha} \tag{5.131}$$

$$= \frac{B(\Sigma^t)^2 < (f_a^t)' > + 1/\text{snr}}{B\alpha} \tag{5.132}$$

$$a_i^{t+1} = f_{a_i}\left((\Sigma^{t+1})^2, \left[a_j^t + (\Sigma^{t+1})^2 \sum_\mu^M \frac{F_{\mu j}\tau_\mu^t}{\Theta^{t+1} + 1/\text{snr}}\right]_{j\in l_i}\right) \tag{5.133}$$

$$= f_{a_i}\left((\Sigma^{t+1})^2, \left[a_j^t + \frac{1}{B\alpha}\sum_\mu^M F_{\mu j}\tau_\mu^t\right]_{j\in l_i}\right) \tag{5.134}$$

$$\tau_\mu^t = y_\mu - \sum_i^N F_{\mu i} a_i^t + \tau_\mu^{t-1}\frac{\Theta^{t+1}}{1/\text{snr} + \Theta^t} \tag{5.135}$$

$$= y_\mu - \sum_i^N F_{\mu i} a_i^t + \tau_\mu^{t-1}\frac{B(\Sigma^t)^2 < (f_a^t)' >}{B\alpha(\Sigma^t)^2} \tag{5.136}$$

$$= y_\mu - \sum_i^N F_{\mu i} a_i^t + \tau_\mu^{t-1}\frac{< (f_a^t)' >}{\alpha} \tag{5.137}$$

where we used the definition of Σ^{t+1} from Fig. 5.6 and (5.129), (5.131) to simplify (5.133) and obtain (5.136). The very last step is to rescale the coding matrix by dividing its elements by $B\alpha : \tilde{\mathbf{F}} := \mathbf{F}/(B\alpha)$. The measure $\tilde{\mathbf{y}}$ is thus rescaled in the same way. We finally obtain the Montanari's form of AMP for homogeneous matrices and B-d vectorial components signals:

$$\tilde{\tau}_\mu^t = \tilde{y}_\mu - \sum_i^N \tilde{F}_{\mu i} a_i^t + \frac{\tilde{\tau}_\mu^{t-1} < (f_a^t)' >}{\alpha} \tag{5.138}$$

$$(\Sigma^{t+1})^2 = \frac{(\Sigma^t)^2 < (f_a^t)' > +1/(B\text{snr})}{\alpha} \tag{5.139}$$

$$a_i^{t+1} = f_{a_i}\left((\Sigma^{t+1})^2, \left[a_j^t + \sum_\mu^M \tilde{F}_{\mu j}\tilde{\tau}_\mu^t\right]_{j\in l_i}\right) \tag{5.140}$$

where $\tilde{\boldsymbol{\tau}}^t$ is the rescaled residual, and:

$$< (f_a^t)' > := \frac{1}{N}\sum_l^L\sum_j^B \left.\frac{\partial f_{a_{l(j)}}(x,\mathbf{y})}{\partial y_j}\right|_{(\Sigma^t)^2, \left[a_j^{t-1} + \sum_\mu^M \tilde{F}_{\mu j}\tilde{\tau}_\mu^{t-1}\right]_{j\in l_i}} \tag{5.141}$$

$$= \frac{1}{N(\Sigma^t)^2}\sum_i^N f_{c_i}\left((\Sigma^t)^2, \left[a_j^{t-1} + \sum_\mu^M \tilde{F}_{\mu j}\tilde{\tau}_\mu^{t-1}\right]_{j\in l_i}\right) \tag{5.142}$$

5.7 State evolution analysis in the spatially-coupled measurement operator case

The derivation of the state evolution with spatial coupling is very similar to the full operator case, see sec. 5.3. The difference is that now each block of the matrix Fig. 5.4 can have a different variance, and thus one must be vigilant when performing the derivation. We give here the main steps, the details being similar to the full case. All the computations are done keeping in mind the limit $L \gg L_c, L_r$ for which AMP is valid with spatial coupling. We start from the algorithm Fig. 5.5 and the operators definitions (5.116), (5.117), (5.118), (5.119). As in sec. 5.3, we need to study the fluctuations of the AMP field, the random variable of the disorder

that takes as input the denoisers:

$$\mathbf{R}_l^{t+1} = \mathbf{a}_l^t + (\mathbf{\Sigma}_l^{t+1})^2 \sum_r^{L_r} \sum_{\mu\in r}^{\alpha_r N/L_c} \frac{\mathbf{F}_{\mu l}}{1/\mathrm{snr} + \Theta_\mu^{t+1}} \left[\sum_c^{L_c} \sum_{k\in c}^{L/L_c} \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right] \tag{5.143}$$

where Λ_μ^t is defined in terms of the AMP quantities in (5.50). The variance of the matrix elements depend only on the block to which they belong, thus in the thermodynamic limit when we replace the square matrix elements by their variance, it simplifies the variance of the measure estimate which end up depending only on the block line index:

$$\begin{aligned}
\Theta_\mu &= \sum_c^{L_c} \sum_{l\in c}^{L/L_c} (\mathbf{F}_{\mu l}^2)^\intercal \mathbf{v}_l && (5.144)\\
&\approx \frac{1}{L} \sum_c^{L_c} J_{r_\mu, c} \sum_{l\in c}^{L/L_c} \sum_{i\in l}^{B} v_i && (5.145)\\
&=: \Theta_{r_\mu} && (5.146)\\
\Rightarrow \Lambda_\mu^t &= \Theta_{r_\mu}^{t+1} \frac{y_\mu - w_\mu^t}{1/\mathrm{snr} + \Theta_{r_\mu}^t} && (5.147)
\end{aligned}$$

where $J_{r,c} \in O(1)$ is the not yet rescaled by $1/L$ variance of the elements of the block of the spatially-coupled operator that is at the r^{th} block-line, c^{th} block-column, see Fig. 5.4. The notation r_μ (c_l) means the block index $r \in \{1,\dots,L_r\}$ ($c \in \{1,\dots,L_c\}$) to which the factor index μ (section index l) belongs to. The previous simplification implies from the definition of $(\mathbf{\Sigma}_l^{t+1})^2$ in Fig. 5.5 that:

$$\begin{aligned}
(\mathbf{\Sigma}_l^{t+1})^2 &= \frac{L_c}{B} \left(\sum_r^{L_r} \frac{J_{r,c_l}\alpha_r}{1/\mathrm{snr} + \Theta_r^{t+1}} \right)^{-1} \mathbf{1}_B && (5.148)\\
&= (\Sigma_{c_l}^{t+1})^2 \mathbf{1}_B && (5.149)
\end{aligned}$$

which thus just depend on the block-column index c_l to which the section l belongs to. We deduce from (5.148) the simplified expression of $\mathbf{R}_l^{t+1}$:

$$\begin{aligned}
\mathbf{R}_l^{t+1} &\approx \mathbf{a}_l^t + (\Sigma_{c_l}^{t+1})^2 \sum_r^{L_r} \frac{1}{1/\mathrm{snr} + \Theta_r^{t+1}} \sum_{\mu\in r}^{\alpha_r N/L_c} \mathbf{F}_{\mu l} \left[\sum_c^{L_c} \sum_{k\in c: k\neq l}^{L/L_c} \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right]\\
&\quad + (\Sigma_{c_l}^{t+1})^2 \underbrace{\sum_r^{L_r} \frac{1}{1/\mathrm{snr} + \Theta_r^{t+1}} \sum_{\mu\in r}^{\alpha_r N/L_c} \mathbf{F}_{\mu l} \left[\mathbf{F}_{\mu l}^\intercal (\mathbf{s}_l - \mathbf{a}_l^t) \right]}_{:=U} && (5.150)\\
&\approx \mathbf{s}_l + (\Sigma_{c_l}^{t+1})^2 \sum_r^{L_r} \frac{1}{1/\mathrm{snr} + \Theta_r^{t+1}} \underbrace{\sum_{\mu\in r}^{\alpha_r N/L_c} \mathbf{F}_{\mu l} \left[\sum_c^{L_c} \sum_{k\in c: k\neq l}^{L/L_c} \mathbf{F}_{\mu k}^\intercal (\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t \right]}_{:=\mathbf{r}_{rl}^{t+1}} && (5.151)
\end{aligned}$$

where we have used the same approximation as in (5.55) which implies:

$$U = (\Sigma_{c_l}^{t+1})^{-2}(\mathbf{x}_l - \mathbf{a}_l^t) + O(1/\sqrt{L}) \tag{5.152}$$

Now we define:

$$\mathbf{r}_l^{t+1} := (\Sigma_{c_l}^{t+1})^2 \sum_r^{L_r} \frac{\mathbf{r}_{rl}^{t+1}}{1/\mathrm{snr} + \Theta_r^{t+1}} \tag{5.153}$$

Using the independence assumption about the matrix elements we can now compute by central limit theorem the moments of the Gaussian distributed variables $\mathbf{r}_{rl}^{t+1}$ in order to deduce the moments of $\mathbf{r}_l^{t+1}$. As in sec. 5.3, we only keep the $O(1)$ terms. We can actually identify $\mathbf{r}_{rl}^{t+1}$ to $\mathbf{r}^{t+1}$ of (5.56) and thus, the computations are exactly the same as in the full operator case up to the values of the variances that are different. Using again (4.137), (5.57) remains valid, so that we get a similar result to (5.59) and (5.61):

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_{rl}^{t+1}\} \approx \mathbf{0}_B \tag{5.154}$$

$$\Rightarrow \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_l^{t+1}\} \approx \mathbf{0}_B \tag{5.155}$$

Identifying in (5.62) $\mathbf{r}_{rl}^{t+1}$ with $\mathbf{r}_l^{t+1}$ which is defined by (5.56), the result (5.65) implies that if $l' \neq l$ then $\forall\ r'$:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_{rl}^{t+1}\mathbf{r}_{r'l'}^{t+1}\} \approx \mathbf{0}_B \tag{5.156}$$

Furthermore, in the case where $r' \neq r$, we have:

$$\begin{aligned}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_{rl}^{t+1}\mathbf{r}_{r'l}^{t+1}\} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\Bigg\{ &\sum_{\mu\in r}\sum_{\nu\in r'} \mathbf{F}_{\mu l}\mathbf{F}_{\nu l}\left[\sum_c^{L_c}\sum_{k\in c:k\neq l}^{L/L_c} \mathbf{F}_{\mu k}^{\mathsf{T}}(\mathbf{s}_k - \mathbf{a}_k^t) + \xi_\mu + \Lambda_\mu^t\right] \\ &\left[\sum_{c'}^{L_c}\sum_{k'\in c':k'\neq l}^{L/L_c} \mathbf{F}_{\nu k'}^{\mathsf{T}}(\mathbf{s}_{k'} - \mathbf{a}_{k'}^t) + \xi_\nu + \Lambda_\nu^t\right]\Bigg\} \end{aligned} \tag{5.157}$$

$$= \mathbf{0}_B \tag{5.158}$$

so $\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_{rl}^{t+1}\mathbf{r}_{r'l'}^{t+1}\}$ can be different of zero only if $l = l'$ and $r = r'$. It implies that the cross terms cancel as well:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{\mathbf{r}_l^{t+1}\mathbf{r}_{l'}^{t+1}\} \approx \mathbf{0}_B \tag{5.159}$$

The only moment that changes with respect to the full matrix case is the variance term. Skipping some steps similar to (5.66), (5.67), we get:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{(\mathbf{r}_{rl}^{t+1})^2\} = \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu\in r}^{\alpha_r N/L_c}\mathbf{F}_{\mu l}^2\left[\sum_{c}^{L_c}\sum_{k\in c:k\neq l}^{L/L_c}\mathbf{F}_{\mu k}^{\mathsf{T}}(\mathbf{s}_k-\mathbf{a}_k^t)\right]\left[\sum_{c'}^{L_c}\sum_{k'\in c':k'\neq l}^{L/L_c}\mathbf{F}_{\mu k'}^{\mathsf{T}}(\mathbf{s}_{k'}-\mathbf{a}_{k'}^t)\right]\right\}$$
$$+\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu\in r}^{\alpha_r N/L_c}\mathbf{F}_{\mu l}^2\xi_\mu^2\right\}+\underbrace{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu\in r}^{\alpha_r N/L_c}\mathbf{F}_{\mu l}^2(\Lambda_\mu^t)^2\right\}}_{\in O(L^{-3/2})}$$
$$+\underbrace{2\,\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu\in r}^{\alpha_r N/L_c}\mathbf{F}_{\mu l}^2\Lambda_\mu^t\left[\sum_{c}^{L_c}\sum_{k\in c:k\neq l}^{L/L_c}\mathbf{F}_{\mu k}^{\mathsf{T}}(\mathbf{s}_k-\mathbf{a}_k^t)\right]\right\}}_{=\mathbf{0}_B} \tag{5.160}$$
$$\approx \mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\sum_{\mu\in r}^{\alpha_r N/L_c}\mathbf{F}_{\mu l}^2\left[\sum_{c}^{L_c}\sum_{k\in c:k\neq l}^{L/L_c}(\mathbf{F}_{\mu k}^2)^{\mathsf{T}}(\mathbf{s}_k-\mathbf{a}_k^t)^2\right]\right\}+\frac{\alpha_r B J_{r,c_l}}{\mathrm{snr}L_c} \tag{5.161}$$
$$\approx \frac{\alpha_r B J_{r,c_l}}{\mathrm{snr}L_c}+\sum_{\mu\in r}^{\alpha_r N/L_c}\frac{J_{r,c_l}}{L}\left[\sum_{c}^{L_c}\frac{J_{r,c_c}}{L}\underbrace{\sum_{k\in c:k\neq l}^{L/L_c}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{(\mathbf{s}_k-\mathbf{a}_k^t)^{\mathsf{T}}(\mathbf{s}_k-\mathbf{a}_k^t)\}}_{\approx E_c^t N/L_c}\right] \tag{5.162}$$
$$= \frac{\alpha_r B J_{r,c_l}}{L_c}\left(\frac{1}{\mathrm{snr}}+\frac{B}{L_c}\sum_{c}^{L_c}J_{r,c_c}E_c^t\right) \tag{5.163}$$

where E_c^t is the MSE at time t of the block c of the signal, see Fig. 5.4:

$$E_c^t := \frac{L_c}{N}\sum_{k\in c}^{L/L_c}\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\{(\mathbf{s}_k-\mathbf{a}_k^t)^{\mathsf{T}}(\mathbf{s}_k-\mathbf{a}_k^t)\} \tag{5.164}$$

The variance of $\mathbf{r}_r^{t+1}$ is deduced from (5.153) using (5.158):

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{(\mathbf{r}_l^{t+1})^2\right\} = (\Sigma_{c_l}^{t+1})^4\sum_{r,r'}^{L_r,L_r}\frac{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{\mathbf{r}_{rl}^{t+1}\mathbf{r}_{r'l}^{t+1}\right\}}{(1/\mathrm{snr}+\Theta_r^{t+1})(1/\mathrm{snr}+\Theta_{r'}^{t+1})} \tag{5.165}$$
$$= (\Sigma_{c_l}^{t+1})^4\sum_{r}^{L_r}\frac{\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{(\mathbf{r}_{rl}^{t+1})^2\right\}}{(1/\mathrm{snr}+\Theta_r^{t+1})^2} \tag{5.166}$$

We define the average variance of the posterior estimates inside the block c as:

$$V_c^t := \frac{L_c}{N}\sum_{l\in c}^{L/L_c}\sum_{i\in l}^{B}v_i^t \tag{5.167}$$

The matching prior conditions (5.100), (5.101) and (5.102) remain true "per block" as the derivation performed in sec. 5.3.3 just assumed the definition of the denoisers and the prior matching condition: we would obtain the same replacing $\left((\Sigma^{t+1})^2,\mathbf{R}^{t+1}(\mathbf{z},\mathbf{s}_l)\right)$ by $\left((\Sigma_c^{t+1})^2,\mathbf{R}_c^{t+1}(\mathbf{z},\mathbf{s}_l)\right)$, the only block index dependent quantities. These conditions allow to greatly simplify the analysis. It implies an equality between the mean variance per block and the MSE per block. It becomes $V_c^t = E_c^t\ \forall\ c\in\{1,\dots,L_c\}$ at each time step if they are initialized with same value.

From this and (5.146), we can re-write Θ_r as:

$$\Theta_r^t = \frac{B}{L_c}\sum_c^{L_c} J_{r,c} V_c^t \tag{5.168}$$

$$= \frac{B}{L_c}\sum_c^{L_c} J_{r,c} E_c^t \tag{5.169}$$

We plug this expression into (5.163) and using (5.166), (5.148) we get:

$$\mathbb{E}_{\mathbf{F},\boldsymbol{\xi},\mathbf{s}}\left\{(\mathbf{r}_l^{t+1})^2\right\} = (\Sigma_{c_l}^{t+1})^4 \frac{B}{L_c}\sum_r^{L_r} \frac{\alpha_r J_{r,c_l}(1/\mathrm{snr}+\Theta_r^{t+1})}{(1/\mathrm{snr}+\Theta_r^{t+1})^2} \tag{5.170}$$

$$= (\Sigma_{c_l}^{t+1})^2 \mathbf{1}_B \tag{5.171}$$

So now we know the distribution of $\mathbf{R}_l^{t+1}$ from (5.151):

$$\mathbf{r}_l^{t+1} \sim \mathcal{N}\left(\mathbf{r}_l^{t+1} \middle| \mathbf{0}_B, (\Sigma_{c_l}^{t+1})^2 \boldsymbol{I}_B\right) \tag{5.172}$$

$$\Rightarrow \mathbf{R}_l^{t+1} \sim \mathcal{N}\left(\mathbf{R}_l^{t+1} \middle| \mathbf{s}_l, (\Sigma_{c_l}^{t+1})^2 \boldsymbol{I}_B\right) \tag{5.173}$$

From the same arguments as in the full case derivation sec. 5.3, we finally obtain the following state evolution over the block mean square error E_c^{t+1} inside the block c in the $L \to \infty$ limit:

$$E_c^{t+1} = \frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l P_0(\mathbf{s}_l) \int_{\mathbb{R}^B} \mathscr{D}\mathbf{z}\left[f_{a_i}\left((\Sigma_c^{t+1})^2, \mathbf{R}_c^{t+1}(\mathbf{z},\mathbf{s}_l)\right) - s_i\right]^2 \tag{5.174}$$

$$\Sigma_c^{t+1}\left(\{E_{c'}^t\}_{c'}^{L_c}\right) = \left[B\sum_r^{L_r} \frac{\alpha_r J_{rc}}{L_c/\mathrm{snr} + B\sum_{c'}^{L_c} J_{rc'} E_{c'}^t}\right]^{-1/2} \tag{5.175}$$

$$\mathbf{R}_c^{t+1}(\mathbf{z},\mathbf{s}_l) := \mathbf{s}_l + \mathbf{z}\Sigma_c^{t+1} \tag{5.176}$$

where f_{a_i} is the denoiser (4.115). As explained previously, the prior matching conditions of sec. 5.3.3 imply the same equalities *per block* as in the full case, so that like in sec. 5.3 the two following forms of state evolution for spatially-coupled matrices are equivalent to (5.174):

$$E_c^{t+1} = \frac{1}{B}\left[\mathbb{E}_{\mathbf{s}_c}\{\mathbf{s}_l^\intercal \mathbf{s}_l\} - \sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l P_0(\mathbf{s}_l)\int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} f_{a_i}\left((\Sigma_c^{t+1})^2, \mathbf{R}_c^{t+1}(\mathbf{z},\mathbf{s}_l)\right) s_i\right] \tag{5.177}$$

$$= \frac{1}{B}\sum_{i\in l}^{B}\int_{\mathbb{R}^B} d\mathbf{s}_l P_0(\mathbf{s}_l)\int_{\mathbb{R}^B} \mathscr{D}\mathbf{z} f_{c_i}\left((\Sigma_c^{t+1})^2, \mathbf{R}_c^{t+1}(\mathbf{z},\mathbf{s}_l)\right) \tag{5.178}$$

together with (5.175) and (5.176), where $\mathbb{E}_{\mathbf{s}_c}\{\mathbf{s}_l^\intercal \mathbf{s}_l\}/B$ is the average of the squared signal's components of the block c, see Fig. 5.4 (in all this thesis, the statistical properties of the signal are homogeneous). f_{c_i} outputs the posterior variance and is given by (4.116).

Signal processing Part III

6 Compressed sensing of approximately sparse signals

Compressed sensing is designed to measure sparse signals directly in a compressed form. However, most signals of interest are only approximately sparse, i.e. even though the signal contains only a small fraction of relevant large components, the other ones are not strictly equal to zero but are only close to it. In this chapter we model the approximately sparse i.i.d signal with a sum of two Gaussian distributions, one with large variance correponding to the informative support of the signal, the second for the small components (which act as an effective noise) with smaller variance, and we study its compressed sensing with dense random matrices. We use replica calculations to determine the mean square error of the Bayes optimal reconstruction for such signals as a function of the variance of the small components, the density of large components and the measurement rate. We then study the approximate message-passing algorithm for approximate sparsity and we quantify the region of parameters for which it achieves optimality (for large systems). Finally, we show that in the region where the AMP algorithm with the homogeneous measurement matrices is not optimal, the spatial coupling allows to restore optimality. Even though we limit ourselves to this special class of signals and assume the prior matching condition, many qualitative features of the results stay true for other signal models and when the distribution of the signal-components is not known as well [118].

The ℓ_1-minimization based algorithms [30, 32] are widely used for compressed sensing of approximately sparse signals. They are very general as discussed in sec. 3.4 and provide good performances in many situations. They, however, do not achieve optimal reconstruction even when the statistical properties of the signal are known, see sec. 5.1.1.

As we shall see the AMP algorithm for homogeneous measurement matrices matches asymptotically the performance of the optimal reconstruction in a large part of the parameter space. However in some region of parameters that define the hard phase, it is suboptimal as the BP transition blocks the algorithm before the optimal one, see sec. 5.1.1. In compressed sensing the spatial coupling was first tested in [119] who did not observe any improvement for the present bi-Gaussian model for reasons that we will clarify later in this chapter. Basically, the spatial coupling provides improvements only if a first order phase transition is present, but

for the variance of small components that was tested in [119] there is no such transition: it appears only for slightly smaller values.

6.1 The bi-Gaussian prior for approximate sparsity

We study compressed sensing for approximately sparse signals: the N-d scalar components signals (i.e. $L = N, B = 1$ in the general equations for vector components signals presented in the previous theoretical chapters chap. 4 and chap. 5) that we consider have i.i.d components, K of these being drawn from a distribution $\phi(s)$ and the density of such components is $\rho = K/N$. The remaining $N-K$ components are Gaussian with zero mean and small variance ϵ :

$$P_0(\mathbf{s}) = \prod_{i=1}^{N} P_0(s_i) \tag{6.1}$$

$$= \prod_{i=1}^{N} [\rho\phi(s_i) + (1-\rho)\mathcal{N}(s_i|0,\epsilon)] \tag{6.2}$$

Of course no real signal of interest is truly i.i.d. However, our analysis also applies to non i.i.d signals which empirical distribution of components is converging to $P_0(s_i)$, this condition being sufficient [110]. We focus on the special case of a Gaussian $\phi(s_i) = \mathcal{N}(s_i|0,\sigma^2 = 1)$ of zero mean and unit variance. Although the numerical results depend on the form of $\phi(s_i)$, the overall picture is robust with respect to this choice. We further assume the prior matching condition: the parameters of $P(\mathbf{s})$ are known and used in the algorithm. The bi-Gaussian model for approximately sparse signals (6.2) was previously used in compressed sensing, for example in [91, 119].

For simplicity we assume the measurements to be noiseless, the case of noisy measurements can be treated similarly by the AMP algorithm and learned by expectation maximization (see sec. 4.3.8) if unknown. We consider the measurement matrix $\mathbf{F}$ having i.i.d components of zero mean and variance $1/N$. The measurements are obtained through (3.18). In the numerical experiments, the components of the matrix are Gaussian distributed, but the asymptotic analysis does not depend on the details of the components distribution, as long as they are i.i.d.

The Bayes optimal estimation is intractable in the general case, but as discussed in sec. 5.1.1 the AMP estimation is Bayes optimal under the matching prior condition before its spinodal transition, which blocks its convergence. We will use an asymptotic replica analysis of the Bayes optimal reconstruction, which allows to compute the asymptotic MSE as a function of the parameters of the signal distribution (ρ,ϵ) and of the measurement rate α. This allows to obtain the phase diagram of the problem.

6.1.1 Learning of the prior model parameters

If the prior parameters are unknown, they can be learned efficiently through expectation maximization described in sec. 4.3.8. Here we could start from the Bethe free energy (4.194) and derive fixed point equations but as the parameters have simple interpretations, we can derive learnings more easily. Actually, the easiest way is exactly as we will do later on in sec. 8.2.5 to which we refer: it just requires to compute the posterior probability estimates $\{P(x_i^t \in \mathcal{N})\}_i^N$ at time t that the signal components have been generated by the Gaussian part of the prior:

$$P(x_i^t \in \mathcal{N}) = (1-\rho)\frac{\int dx_i \mathcal{N}(x_i|0,\epsilon)\mathcal{N}(x_i|R_i^t,(\Sigma_i^t)^2)}{\int dx_i P_0(x_i)\mathcal{N}(x_i|R_i^t,(\Sigma_i^t)^2)} \tag{6.3}$$

where $P_0(x_i)$ is given by (6.2) and $(R_i^t,(\Sigma_i^t)^2)$ are the AMP fields at time t. Then from this, the parameters are easily derived, see sec. 8.2.5. For example, the density of informative components (that have been generated by ϕ is (6.2)) is just (8.30). The other parameters are learned similarly.

6.2 Reconstruction of approximately sparse signals with the approximate message-passing algorithm

The generic AMP algorithm in its scalar form Fig. 4.6 is now studied. Only the denoisers f_{a_i} and f_{c_i} depend explicitely on the signal model $P_0(\mathbf{s})$ (6.2). Referring to the sec. 4.3.6 and using the table Tab. 4.1 for the prior construction, we get directly the denoisers for bi-Gaussian approximate sparsity:

$$f_a(\Sigma^2,R) = \frac{\sum_{a=1}^2 w_a e^{-\frac{R^2}{2(\Sigma^2+\sigma_a^2)}}\frac{R\sigma_a^2}{(\Sigma^2+\sigma_a^2)^{\frac{3}{2}}}}{\sum_{a=1}^2 w_a \frac{1}{\sqrt{\Sigma^2+\sigma_a^2}} e^{-\frac{R^2}{2(\Sigma^2+\sigma_a^2)}}} \tag{6.4}$$

$$f_c(\Sigma^2,R) = \frac{\sum_{a=1}^2 w_a e^{-\frac{R^2}{2(\Sigma^2+\sigma_a^2)}}\frac{\sigma_a^2\Sigma^2(\Sigma^2+\sigma_a^2)+R^2\sigma_a^4}{(\Sigma^2+\sigma_a^2)^{\frac{5}{2}}}}{\sum_{a=1}^2 w_a \frac{1}{\sqrt{\Sigma^2+\sigma_a^2}} e^{-\frac{R^2}{2(\Sigma^2+\sigma_a^2)}}} - f_a(\Sigma^2,R)^2 \tag{6.5}$$

where we use the notation f_a/f_c instead of f_{a_i}/f_{c_i} of Fig. 4.6 in the scalar components signal case. For the approximately sparse signals that we consider here we have:

$$w_1 = \rho \tag{6.6}$$

$$\sigma_1^2 = \sigma^2 = 1 \tag{6.7}$$

$$w_2 = 1-\rho \tag{6.8}$$

$$\sigma_2^2 = \epsilon \tag{6.9}$$

6.2.1 State evolution of the algorithm with homogeneous measurement matrices

In the limit of large system sizes, i.e. when the parameters (ρ, ϵ, α) are fixed whereas $N \to \infty$, the evolution of the AMP algorithm is described exactly using the state evolution [104] given by (5.105) (or any of the other two equivalent forms) under the prior matching condition, as it is the case here. When $B = 1, L = N$ it becomes in the noiseless case snr $\to \infty$:

$$E^{t+1} = \int ds P_0(s) \int \mathscr{D}z f_c\left(\frac{E^t}{\alpha}, s + z\sqrt{\frac{E^t}{\alpha}}\right) \tag{6.10}$$

Now plugging the bi-Gaussian prior $P_0(s) = \sum_a^2 w_a \mathcal{N}(s|0,\sigma_a^2)$ in (6.10), using the fact that the sum of two independent Gaussian random variables is a new Gaussian random variable with mean and variance given by the sum of the means and variances of the original random variables plus the fact that:

$$\int du \mathcal{N}(u|m,v) f(u) = \int \mathscr{D}z f(\sqrt{v}z + m) \tag{6.11}$$

we directly obtain the final simplified state evolution recursion:

$$E^{t+1} = \sum_{a=1}^{2} w_a \int \mathscr{D}z f_c\left(\frac{E^t}{\alpha}, z\sqrt{\sigma_a^2 + \frac{E^t}{\alpha}}\right) \tag{6.12}$$

where again $\mathscr{D}z = e^{-z^2/2}/\sqrt{2\pi}dz$ is a unit centered Gaussian measure. The initialization corresponding to the one for the algorithm is $E^{t=0} = \mathrm{Var}_{P_0}(x) = (1-\rho)\epsilon + \rho\sigma^2$.

In Fig. 6.1 we plot the analytical prediction for the time evolution of the MSE computed from the state evolution (6.12), and we compare it to the one measured in one run of the AMP algorithm for a system size $N = 3 \cdot 10^4$. The agreement for such system size is already excellent. As we see, when the measurement rate is too low, the algorithm converges to an high MSE. Furthermore, we observe that when reconstruction succeeds, the MSE falls to a value comparable to the small components variance, here $\epsilon = 10^{-6}$.

6.2.2 Study of the optimal reconstruction limit by the replica method

As discussed in sec. 5.1.1, for measurement rates below the BP transition and above the static transition $\alpha_s(\rho) < \alpha < \alpha_{BP}(\rho)$, the state evolution equation (6.12) has two different stable fixed points. In particular, if the iterations are initialized with $E \to 0$, one will reach a fixed point with much lower MSE than initializing with large $E = 1$. In fact, if $\alpha_s(\rho) \leq \alpha_{opt}(\rho) < \alpha < \alpha_{BP}(\rho)$ the low error fixed point determines the MSE that would be achieved by the exact Bayes optimal inference. Let us now compute the phase diagram of compressed sensing for bi-Gaussian approximately sparse signals from the Bethe free entropy using the replica method. We start from the general potential valid under the prior matching condition (5.48), which becomes in

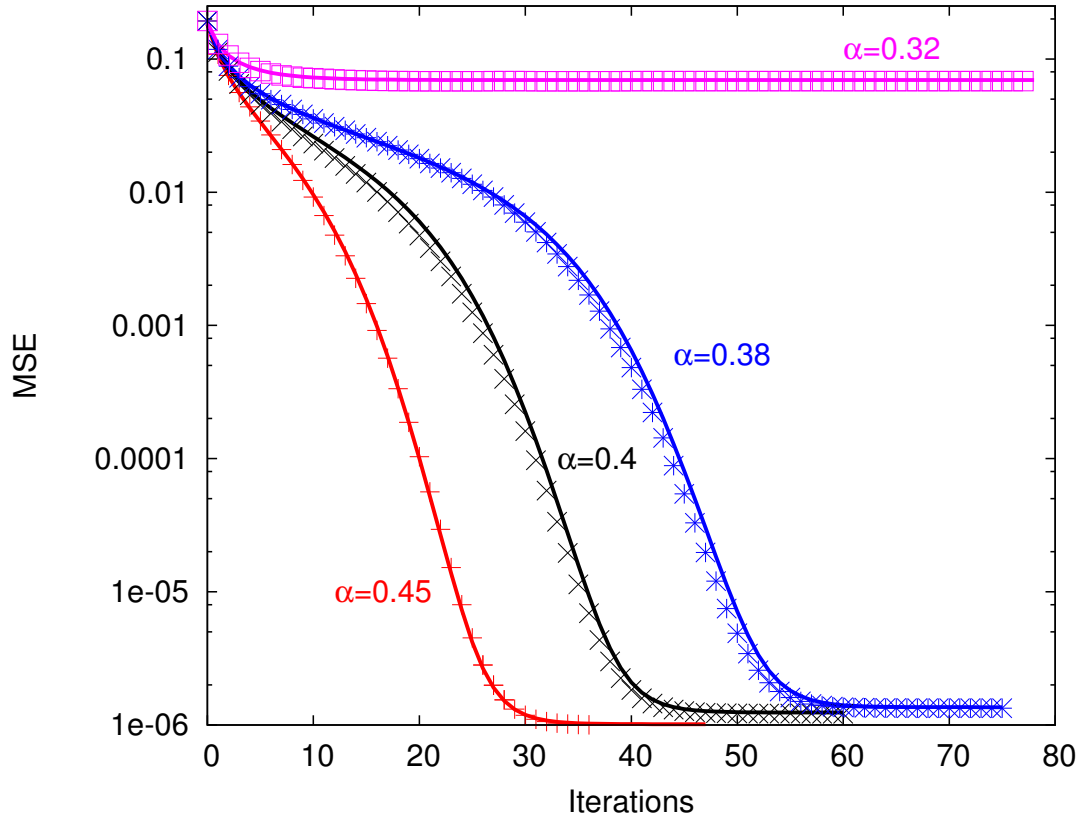

Figure 6.1 – Time evolution of the MSE the AMP algorithm achieves (crosses) compared to the asymptotic $N \to \infty$ evolution obtained from the state evolution (6.12) (full lines) for different measurement rates. Data are obtained for a signal with density of large component $\rho = 0.2$ and variance of the small components $\epsilon = 10^{-6}$. The algorithm was used for a signal of $N = 3 \cdot 10^4$ components.

the scalar components $B = 1$ and noiseless snr $\to \infty$ case:

$$\begin{aligned}\Phi(E) &= \Phi_{B=1}(E|\text{snr} \to \infty) \\ &= -\frac{\alpha}{2}\left(\log(E) + \frac{<\mathbf{s}^2>}{E}\right) + \int P_0(s)\mathscr{D}z \log\left(\int P_0(x) e^{\frac{sx}{\Sigma(E)^2} + \frac{zx}{\Sigma(E)} - \frac{x^2}{2\Sigma(E)^2}}\right)\end{aligned} \tag{6.13}$$

up to irrelevant constants that do not depend on the MSE. $\Sigma(E) = \sqrt{E/\alpha}$ is defined by (5.49). We define I as the integral appearing in the previous expression. Now using the bi-Gaussian prior $P_0(s) = \sum_i^2 w_i \mathcal{N}(s|0,\sigma_i^2)$ we can compute I by Gaussian integral:

$$\int dx e^{-ax^2+bx} = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}} \tag{6.14}$$

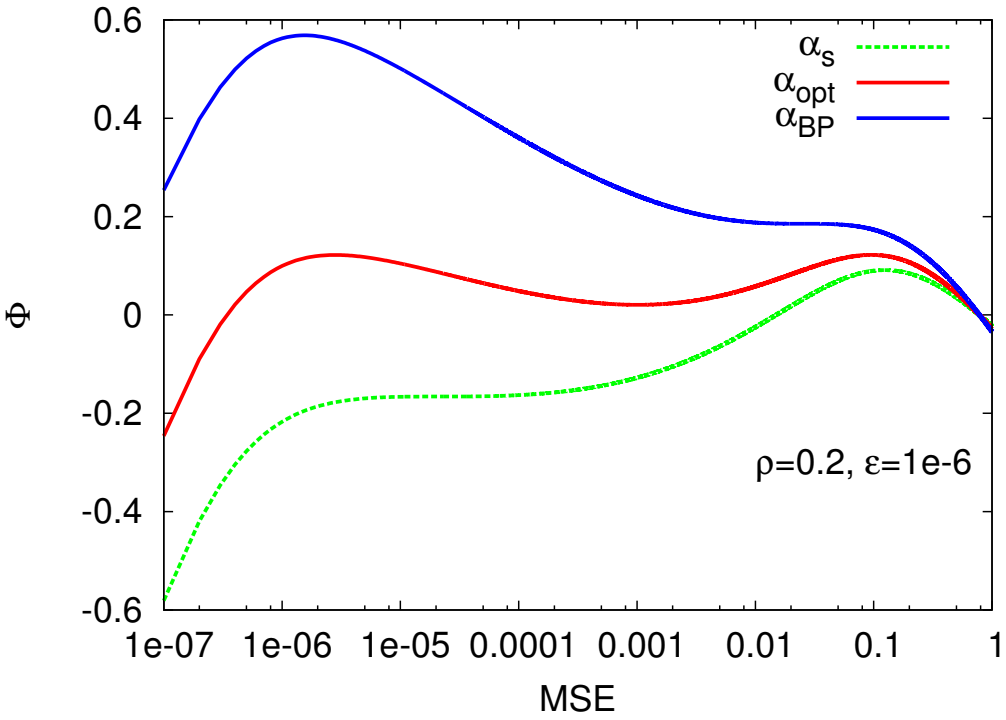

Figure 6.2 – The Bethe free entropy $\Phi(E)$ for compressed sensing of approximately sparse signals of density $\rho = 0.2$, with variance of the small components $\epsilon = 10^{-6}$. The three lines depict the potential for three different measurement rates corresponding to the critical values: $\alpha_{BP} = 0.3559$ below which AMP is not Bayes optimal anymore without spatial coupling, $\alpha_{opt} = 0.2817$, $\alpha_s = 0.2305$. The two local maxima exists for $\alpha \in [\alpha_s, \alpha_{BP}]$, and at $\alpha > \alpha_{opt}$ the low MSE maxima is the global one, i.e. the $MMSE$ estimate. Below the static transition $\alpha < \alpha_s$, all information about the signal is lost and only remain the spurious solution at high MSE.

to get:

$$
\begin{aligned}
I &= \sum_i^2 w_i \int ds \mathcal{D}z \mathcal{N}(s|0,\sigma_i^2) \log\left(\sum_j^2 \frac{w_j}{\sqrt{2\pi\sigma_j^2}} \int dx e^{-\frac{x^2}{2}(1/\Sigma(E)^2+1/\sigma_j^2)+\frac{x}{\Sigma(E)}(s/\Sigma(E)+z)} \right) \\
&= \sum_i^2 w_i \int ds \mathcal{D}z \mathcal{N}(s|0,\sigma_i^2) \log\left(\sum_j^2 \frac{w_j}{\sqrt{\sigma_j^2/\Sigma(E)^2+1}} e^{\frac{(s/\Sigma(E)+z)^2}{2(1+\Sigma(E)^2/\sigma_j^2)}} \right)
\end{aligned}
\tag{6.15}
$$

Now we use again the simple form of the moments of the sum of two independent Gaussian random variables together with (6.11) and defining $u := s/\Sigma(E) + z$ we get:

$$I = \sum_i^2 w_i \int du \mathcal{N}(u|0, \sigma_i^2/\Sigma(E)^2 + 1) \log\left(\sum_j^2 \frac{w_j}{\sqrt{\sigma_j^2/\Sigma(E)^2 + 1}} e^{\frac{u^2}{2(1+\Sigma(E)^2/\sigma_j^2)}} \right) \tag{6.16}$$

$$= \sum_i^2 w_i \int \mathscr{D}z \log\left(\sum_j^2 \frac{w_j}{\sqrt{\sigma_j^2/\Sigma(E)^2 + 1}} e^{\frac{z^2(1+\sigma_i^2/\Sigma(E)^2)}{2(1+\Sigma(E)^2/\sigma_j^2)}} \right) \tag{6.17}$$

The final Bethe free entropy expression is thus:

$$\Phi(E) = -\frac{\alpha}{2}\left(\log(E) + \frac{w_1\sigma_1^2 + w_2\sigma_2^2}{E} \right) + \sum_i^2 w_i \int \mathscr{D}z \log\left(\sum_j^2 \frac{w_j}{\sqrt{\sigma_j^2/\Sigma(E)^2 + 1}} e^{\frac{z^2(1+\sigma_i^2/\Sigma(E)^2)}{2(1+\Sigma(E)^2/\sigma_j^2)}} \right) \tag{6.18}$$

In Fig. 6.2 we plot the function $\Phi(E)$ for a signal of density $\rho = 0.2$, variance of small components $\epsilon = 10^{-6}$ and three different values of the measurement rate α corresponding to the critical values at which happen the different phase transitions. We will show next that at a fixed signal density ρ, for a variance of the small components lower than a critical value $\epsilon < \epsilon_c(\rho)$, the optimal Bayes reconstruction has a transition at a critical value $\alpha = \alpha_{opt}(\rho)$ separating a regime with a small value (comparable to ϵ) of the MSE obtained at $\alpha > \alpha_{opt}(\rho)$ from a phase with a large value of the MSE obtained at $\alpha < \alpha_{opt}(\rho)$. As discussed in sec. 5.1.1, this is a first order phase transition (as the BP transition) in the sense that the Bayes optimal MSE jumps discontinuously at $\alpha = \alpha_{opt}(\rho)$.

In this intermediate hard region $\alpha_{opt}(\rho) < \alpha < \alpha_{BP}(\rho)$ the AMP performance can be improved with the use of spatially-coupled measurement matrices and with a proper choice of the parameters defining these matrices, one can approach the performance of the optimal Bayes inference in the large system size limit for any measurement rate.

Finally for higher variance of the small components $\epsilon > \epsilon_c(\rho)$ there is no more phase transition for any $0 < \alpha < 1$. In this regime, AMP always achieves optimal Bayes inference and the MSE that it obtains varies continuously from 0 at $\alpha = 1$ to $O(1)$ values at low measurement rate α.

6.3 Phase diagrams for compressed sensing of approximately sparse signals

In Fig. 6.3 we plot the MSE to which the state evolution converges if initialized at large value of MSE - such initialization corresponds to the iterations of AMP when the actual signal is not known. For $\epsilon = 0.01$ we also compare explicitly to a run of AMP for a system size of

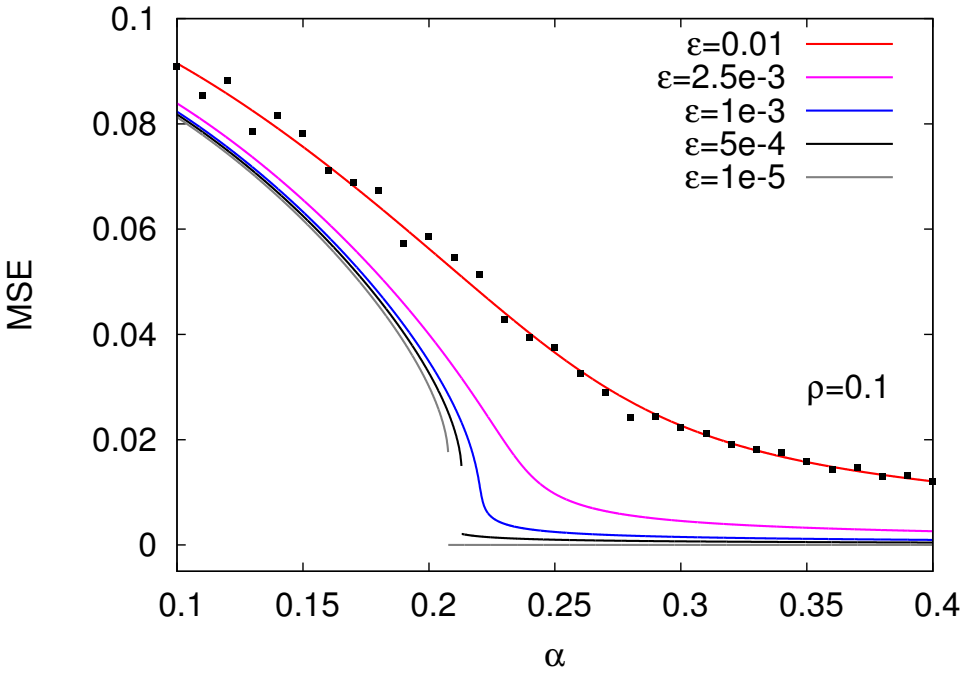

Figure 6.3 – The MSE achieved by the AMP algorithm. The lines correspond to the evaluation of the MSE from the state evolution (6.12), the data points to the MSE achieved by the AMP algorithm on single instances with $N = 3 \cdot 10^4$. The data are for signals with density $\rho = 0.1$ and several values of variance of the small components ϵ as a function of the measurement rate α. The MSE grows continuously as α decreases for $\epsilon > \epsilon_c(\rho = 0.1) = 0.00075$. For smaller values of the small components variance, a first order phase transition is present and the MSE jumps discontinuously at $\alpha_{BP}(\rho = 0.1, \epsilon)$.

$N = 3 \cdot 10^4$. Depending on the value of the density ρ and variance ϵ, two situations are possible: for relatively large ϵ, as the measurement rate α decreases the final MSE grows continuously from $E = 0$ at $\alpha = 1$ to $E = E^{t=0}$ at $\alpha = 0$. For lower values of ϵ the MSE achieved by AMP has a discontinuity at $\alpha_{BP}(\rho, \epsilon)$ at which the second maxima of $\Phi(E)$ appears. Note that the case of $\epsilon = 0.01$ was tested in [91], the case of $\epsilon = 0.0025$ in [119]. This why the authors of [119] did not observe any improvement by spatial coupling: for these parameters (ρ, ϵ), there is no first order transition and thus spatial coupling is useless as AMP is anyway Bayes optimal at any measurement rate α.

In Fig. 6.4 we plot in solid blue line the MSE to which the AMP asymptotically converges and compare it to the MSE achieved by the optimal Bayes inference (in dashed red line), i.e. the MSE corresponding to the global maximum of $\Phi(E)$. We see that, when the discontinuous transition point $\alpha_{BP}(\rho, \epsilon)$ exists, then in the region $\alpha_{opt}(\rho, \epsilon) < \alpha < \alpha_{BP}(\rho, \epsilon)$ AMP is suboptimal. We remind that in the limit $\epsilon \to 0$, exact reconstruction is possible for any $\alpha > \alpha_{opt}(\rho) = \rho$. We see that for $\alpha < \alpha_{opt}(\rho, \epsilon)$ and for $\alpha > \alpha_{BP}(\rho, \epsilon)$ the performance of AMP matches asymptotically the performance of the Bayes optimal inference. The two regions are,

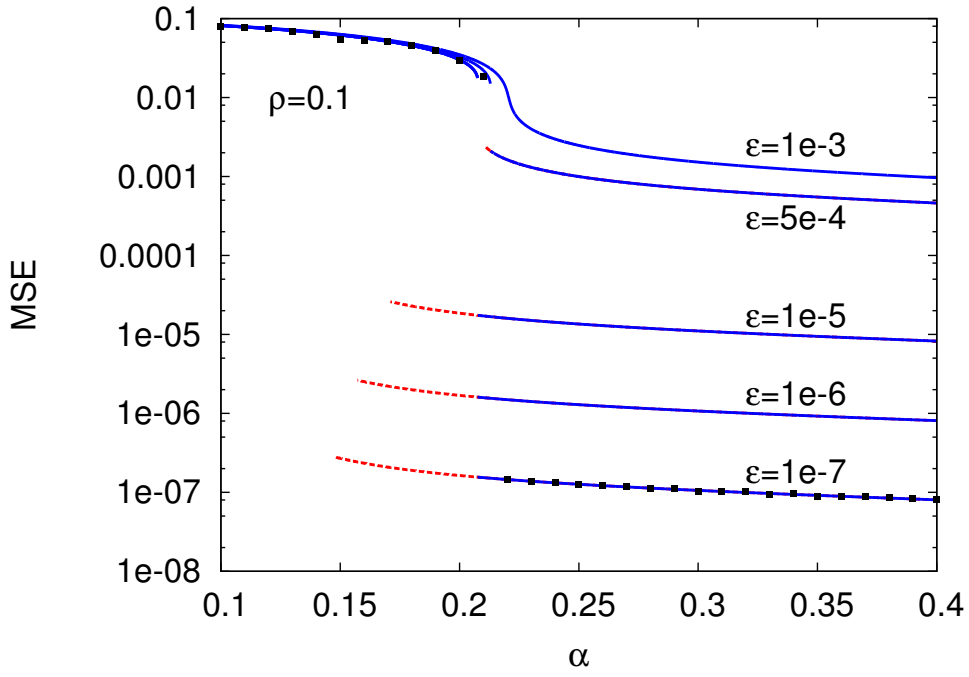

Figure 6.4 – MSE achieved asymptotically by the AMP (blue solid lines) compared to the MSE achieved by the Bayes optimal inference (red dashed lines) as evaluated using the state evolution, initializing it from $E \to 0$ to get the Bayes optimal MSE or $E = 1$ for the AMP asymptotic MSE. The data points correspond to the MSE achieved by the AMP algorithm for $N = 3 \cdot 10^4$. The optimal MSE jumps at $\alpha_{opt}(\rho,\epsilon)$. Hence for $\epsilon < \epsilon_c(\rho = 0.1) = 0.00075$ there is a range of measurement rates $[\alpha_{opt}(\rho = 0.1,\epsilon), \alpha_{BP}(\rho = 0.1,\epsilon)]$ for which the AMP algorithm is asymptotically suboptimal. In this gap, spatial coupling can be used to restore the optimality of AMP.

however, quite different as discussed in sec. 5.1.1. For $\alpha < \alpha_{opt}(\rho,\epsilon)$ the final MSE is relatively large, whereas for $\alpha > \alpha_{BP}$ the final MSE is of order ϵ and hence is this region the problem shows a very good stability towards approximate sparsity.

In Fig. 6.5 we summarize the critical values of $\alpha_{BP}(\epsilon,\alpha)$ and $\alpha_{opt}(\epsilon,\alpha)$ for a signal of density $\rho = 0.1$ as a function of the variance of the small components and the measurement rate. Note that for $\epsilon > \epsilon_c(\rho = 0.1) = 0.00075$ there are no phase transitions anymore, hence for this large value of ϵ, the AMP algorithm matches asymptotically the optimal Bayes inference at any α. Note that in the limit of exactly sparse signal $\epsilon \to 0$, the values $\alpha_{opt}(\rho) \to \rho$ and $\alpha_s(\rho) \to \rho$ whereas $\alpha_{BP}(\rho) \to 0.2076$, hence for $\alpha > 0.2076$ the AMP algorithm is very robust with respect to appearance of approximate sparsity since the transition $\alpha_{BP}(\rho)$ has a very weak ϵ-dependence, as seen in Fig. 6.5.

In Fig. 6.6 we plot the phase diagram at fixed variance ϵ in the density ρ, measurement rate α plane. The only space for improvement is in the region $\alpha_{opt}(\rho,\alpha) < \alpha < \alpha_{BP}(\rho,\alpha)$, which shrinks as ϵ increases. In this region, AMP is not optimal because the potential $\Phi(E)$ has two

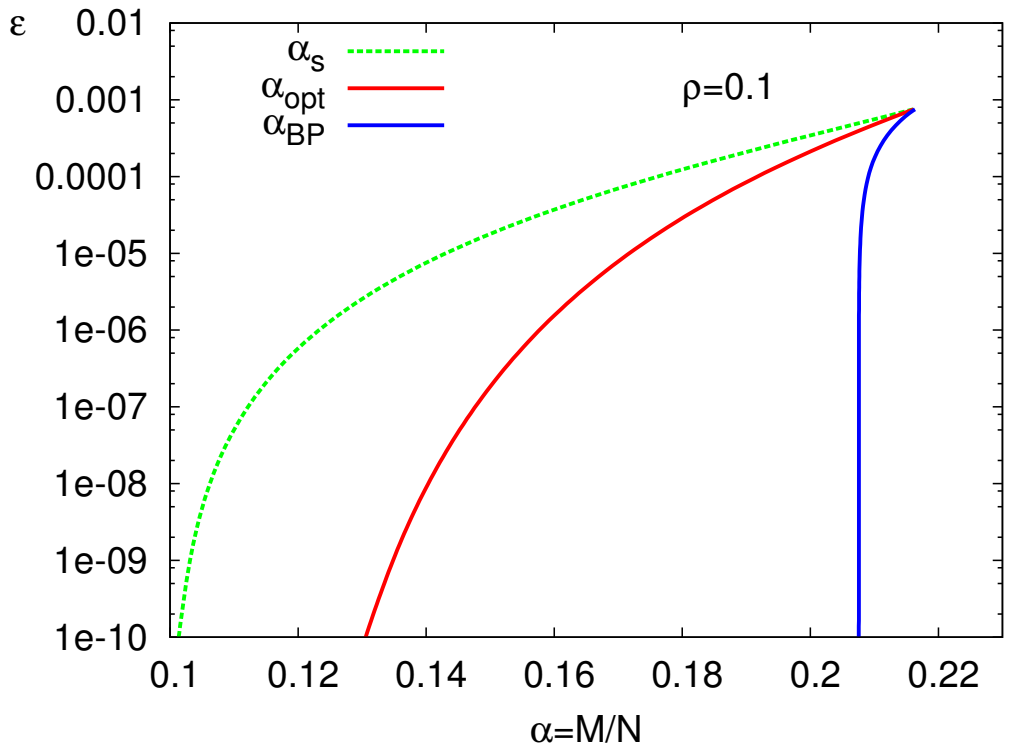

Figure 6.5 – Phase diagram for compressed sensing of approximately sparse signals. The density of the large signal components is $\rho = 0.1$, we are changing the measurement rate α and the variance of the small components ϵ. The critical values of measurement rates $\alpha_{opt}(\epsilon,\alpha)$, $\alpha_{BP}(\epsilon,\alpha)$ and $\alpha_s(\epsilon,\alpha)$ are plotted. For homogeneous measurement matrices, AMP does not achieve optimal reconstruction in the area between $\alpha_{opt}(\epsilon,\alpha)$ (red curve) and $\alpha_{BP}(\epsilon,\alpha)$ (blue curve). For any measurement rate above the tri-critical point where the three transitions curves meet, there is no more phase transitions and AMP is always Bayes optimal for any ϵ and the MSE becomes a continuous function of α.

maxima, and the iterations are blocked in the "wrong" metastable local maximum of the potential $\Phi(E)$ with the largest E.

6.4 Reconstruction of approximately sparse signals with optimality achieving matrices

A first order phase transition that is causing a failure (sub-optimality) of the AMP algorithm appears also in the case of truly sparse signals [34], see sec. 5.1.1. In that case [34] showed that with the so-called seeding (i.e. spatially-coupled) measurement matrices, the AMP algorithm is able to restore asymptotically optimal performance as discussed in sec. 5.5. This was proven rigorously in [110]. Using arguments from the theory of crystal nucleation, it was argued heuristically in [34] that spatial coupling provides improvement whenever, but only if, a

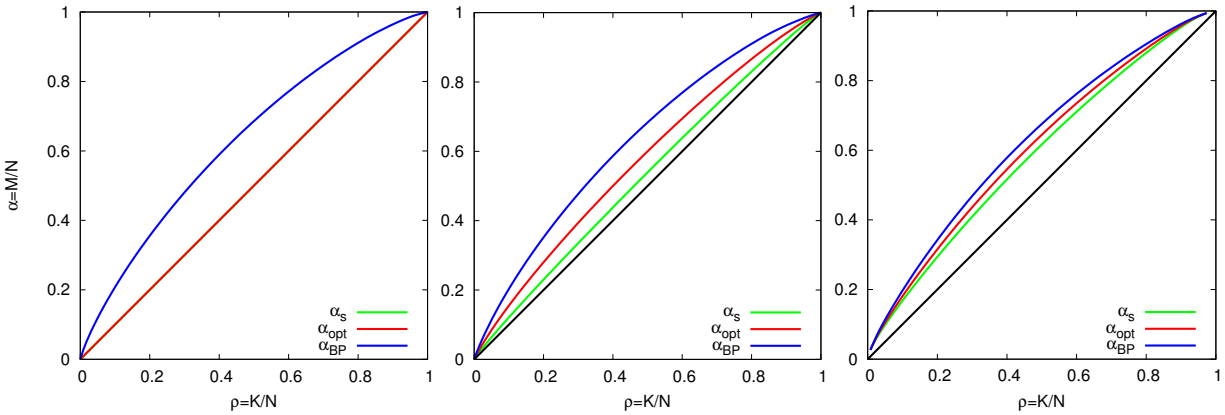

Figure 6.6 – Phase diagrams for compressed sensing of approximately sparse signals in the (α,ρ) plane with variance of small components $\epsilon = 0$ (left), $\epsilon = 10^{-6}$ (center) and $\epsilon = 10^{-4}$ (right). As ϵ increases, the space for improvement of the AMP results by spatial coupling, situated between the optimal and BP transitions, shrinks until it will totally disappear.

first order phase transition is present. Spatial coupling was first suggested for compressed sensing in [119] where the authors tested cases without a first order phase transition (see Fig. 6.3), hence no improvement was observed. Here we show that for measurement rates $\alpha_{opt}(\rho,\epsilon) < \alpha < \alpha_{BP}(\rho,\epsilon)$, seeding matrices allow to restore optimality also for the inference of approximately sparse signals.

6.4.1 Restoring optimality thanks to spatial coupling

In order to restore the asymptotic optimality of AMP with approximately sparse signals, we use spatially-coupled measurement matrices of the form Fig. 5.4. The state evolution for such block matrices have been derived in sec. 5.7. The general recursion is given by (5.178). As the derivation in the present setting is exactly the same as in the full measurement matrix case of sec. 6.2.1 up to the block index, we give here directly the spatially-coupled state evolution recursion for approximate sparsity:

$$E_c^{t+1} = \sum_{a=1}^{2} w_a \int \mathscr{D}z f_c\left((\Sigma_c^{t+1})^2, z\sqrt{\sigma_a^2 + (\Sigma_c^{t+1})^2}\right) \tag{6.19}$$

$$\Sigma_c^{t+1}\left(\{E_{c'}^t\}_{c'}^{L_c}\right) = \left[\sum_r^{L_r} \frac{\alpha_r J_{rc}}{\sum_{c'}^{L_c} J_{rc'} E_{c'}^t}\right]^{-1/2} \tag{6.20}$$

where we have used (5.175) in the noiseless case with $B = 1$ and α_r is the measurement rate of all the blocks at the r^{th} block-row, J_{rc} the $O(1)$ variance of the (r,c)-block elements see Fig. 5.4. This kind of evolution belongs to the class for which threshold saturation (asymptotic achievement of performance matching the optimal Bayes inference solver) was proven in [120] (when $L_c \to \infty$, $W \to \infty$ and $L_c/W \gg 1$). This asymptotic guarantee is reassuring, but one must check if finite N corrections are gentle enough to be able to perform compressed sensing close

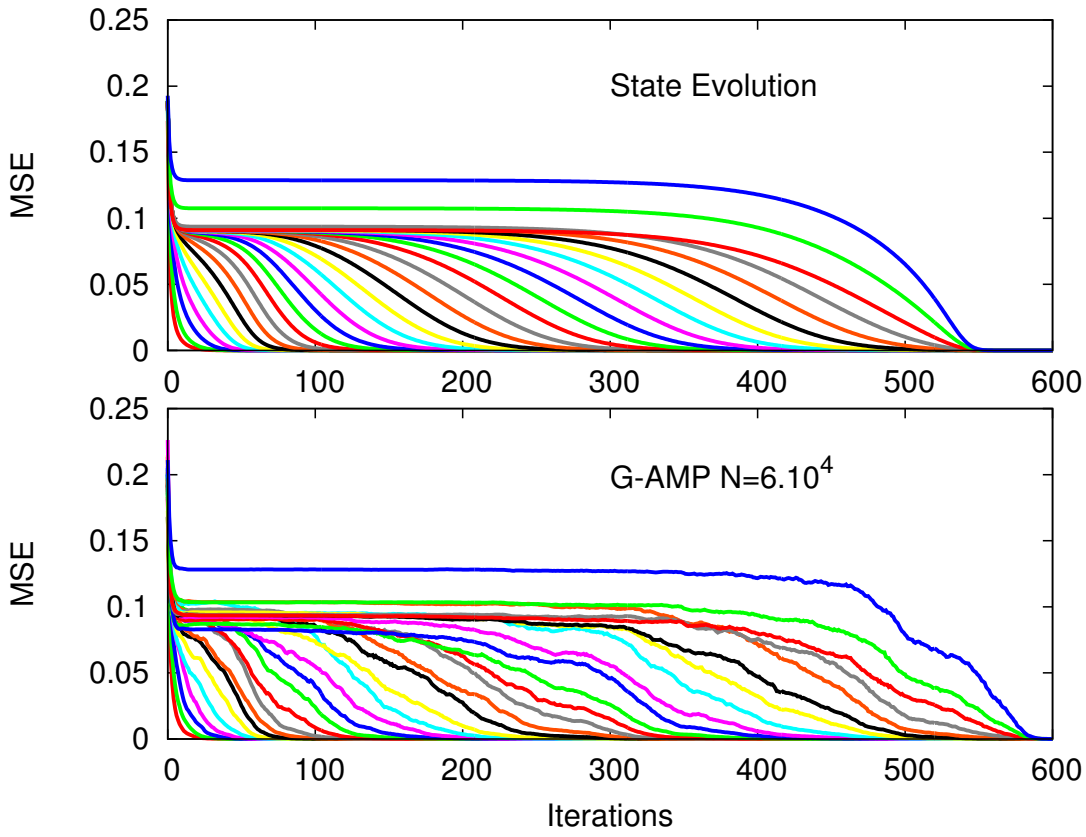

Figure 6.7 – Evolution of the MSE in reconstruction of an approximately sparse signal with density $\rho = 0.2$, variance of small components $\epsilon = 10^{-6}$ at measurement rate $\alpha = 0.303$. The state evolution on the top is compared to the evolution of the algorithm for a signal size $N = 6 \cdot 10^4$ on the bottom. The measurement is performed using a seeding matrix with the following parameters: ($\alpha_{seed} = 0.4$, $\alpha_{rest} = 0.29$, $W = 3$, $J = 0.2$, $L_c = 30$, $L_r = 31$). Each colored curve correspond to a different signal block (see Fig. 5.4), and we clearly see the reconstruction wave propagating.

to $\alpha_{opt}(\rho, \epsilon)$ even for practical system sizes, see sec. 3.1.7. We hence devote the next section to numerical experiments showing that the AMP algorithm is indeed able to reconstruct close to optimality with spatially-coupled matrices.

In Fig. 6.7 we show the spatially-coupled state evolution compared to the evolution of the AMP algorithm for system size $N = 6 \cdot 10^4$. The signal was of density $\rho = 0.2$ and $\epsilon = 10^{-6}$, the parameters of the measurement matrix are in the second line of Tab. 6.1, $L_c = 30$ giving measurement rate $\alpha = 0.303$ which is deep in the region where AMP for homogeneous measurement matrices is not Bayes optimal and gives large MSE (for any $\alpha < \alpha_{BP}(\rho = 0.2, \epsilon = 10^{-6}) = 0.356$). We see finite size fluctuations, but the overall evolution corresponds well to the asymptotic curve, and we see the reconstruction wave propagation happening.

In Fig. 6.8 we plot the asymptotic convergence time needed to achieve reconstruction with $E \approx$

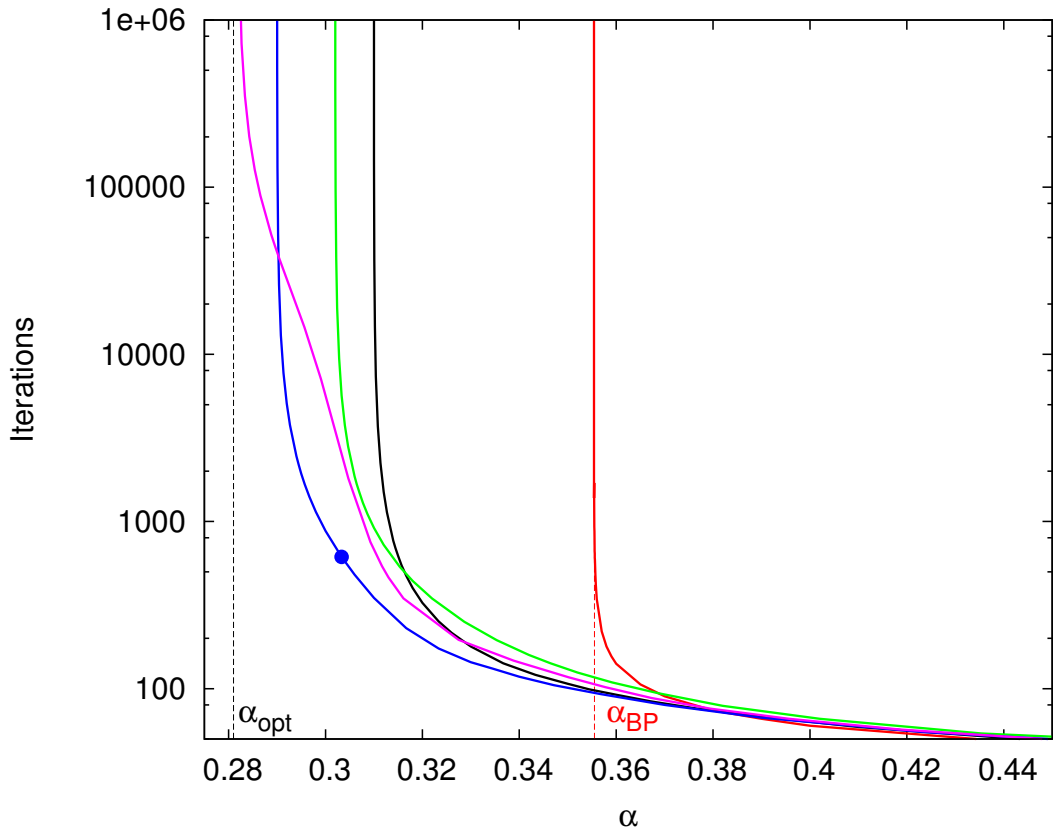

Figure 6.8 – The convergence time of AMP for large system sizes estimated by the state evolution as a function of the measurement rate α. Data are for signals with density $\rho = 0.2$, variance of small components $\epsilon = 10^{-6}$. The red line is obtained using an homogeneous measurement matrix, the vertical dashed line corresponds to the limit this approach can achieve $\alpha_{BP}(\epsilon = 10^{-6}, \rho = 0.2) = 0.3554$. All the other lines are obtained using a spatially-coupled matrix with parameters specified in Table 6.1 and varying α_{rest} by changing L_c which are related by (5.114), the resulting measurement rate α is computed from (5.115). With these seeding matrices and using large L_c, reconstruction is possible at least down to $\alpha_{rest} = 0.282$ which is very close to the measurement rate $\alpha_{opt} = 0.2817$. The blue point corresponds to the evolution illustrated in Fig. 6.7. The divergence of the convergence time of AMP approaching the phase transitions is the critical slowing down discussed in sec. 5.1.1, a typical behavior of local algorithms near first order phase transitions, like the present BP and optimal ones.

ϵ for several sets of parameters of the seeding matrices, see Tab. 6.1. Each color corresponds to a different L_c, which changes the α_{rest} thanks to (5.114). With a proper choice of the parameters, we see that we can reach an optimal reconstruction for values of α extremely close to $\alpha_{opt}(\epsilon, \rho)$. Note, however, that the number of iterations needed to converge diverges as $\alpha \to \alpha_{opt}(\epsilon, \rho)$. This critical slowing down typical of first order phase transitions is discussed in sec. 5.1.1. This is very similar to what has been obtained in the case of purely sparse signals in [34, 110].

color	α_{seed}	α_{rest}	J	W	L_r
purple	0.4	0.282	0.3	3	L_c+2
blue	0.4	0.290	0.2	3	L_c+1
green	0.4	0.302	0.001	2	L_c+1
black	0.4	0.310	0.4	3	L_c+1

Table 6.1 – Parameters of the seeding matrices used in Fig. 6.8. The α_{rest} is modified by changing L_c, the link between them being (5.114).

6.4.2 Finite size effects influence on spatial coupling performances

It is important to point out that these theoretical analyzes are valid for $N \to \infty$ only. Since we eventually work with finite size signals, in practice, finite size effects slightly degrade this asymptotic threshold saturation, see sec. 3.1.7. This is a well known effect in coding theory where a major question is how to optimise finite-length codes (see for instance [23, 121]).

In Fig. 6.9 we plot the fraction of cases in which the algorithm reached successful reconstruction for different system sizes as a function of the number of blocks L_c. We see that for a given size as the number of blocks is growing, i.e. as the size of one block decreases, the performance deteriorates. As expected the situation improves when the size increases. Analysies of the data presented in Fig. 6.9 suggest that the size of one block that is needed for good performance grows roughly linearly with the number of blocks L_c. This suggests that the probability of failure to transmit the information to every new block is roughly inversely proportional to the block size. The algorithm nevertheless reconstructs signals at rates close to the optimal one for system sizes of practical interest. This figure emphasizes the tradeoff between measurement rate decrease and probability of success in the reconstruction because as L_c increases, we can theoretically decode closer to the optimal threshold as seen from (5.115). But in the same time, it increases the finite size effects influence and thus lowers the probability of success.

Fig. 6.10 is the phase diagram for a variance of the small components $\epsilon = 10^{-6}$ in the (α, ρ) plane and shows how with spatial coupling, we can reconstruct instances generated in the hard phase between the BP transition and the optimal one. We notice that the pink crosses corresponding to these finite size instances reconstructed by spatial coupling for fixed size $N = 2^{14}$ are approximately at a constant distance of the optimal transition, and thus as the region allowing for improvement $[\alpha_{opt}(\epsilon = 10^{-6}, \rho), \alpha_{BP}(\epsilon = 10^{-6}, \rho)]$ gets smaller when ρ decreases, the gain with respect to the BP transition decreases as well. But even for this small size, a non negligible gain is possible when ρ is not too small.

6.5 Some results on real images

We now present some results on a potential application of the approximate sparsity prior: image reconstructions. The appropriate sparsifying basis for natural images is the wavelet

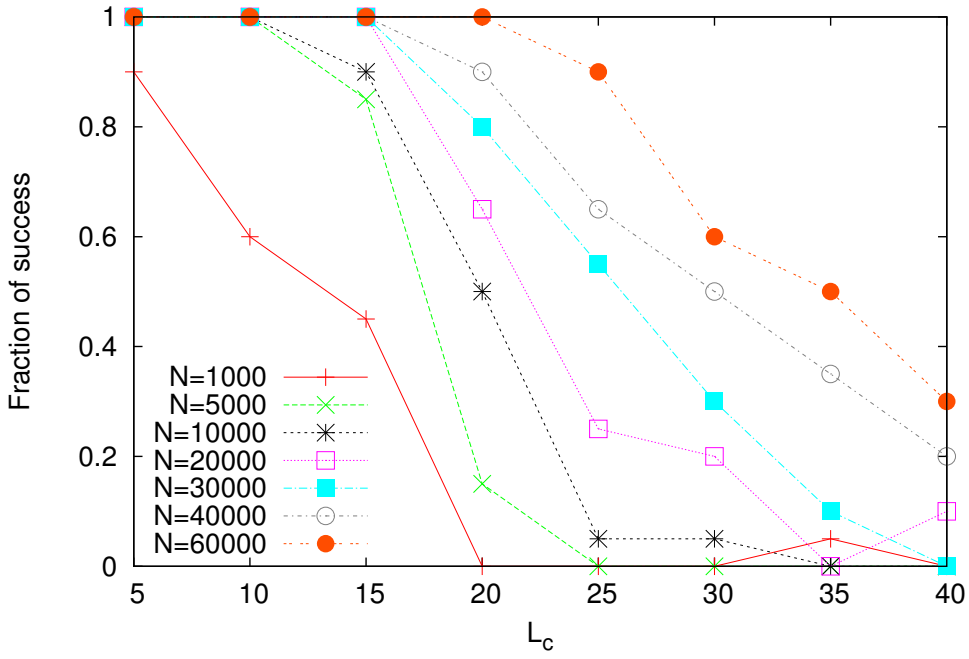

Figure 6.9 – Fraction of instances (over 20 attempts) that were solved by the algorithm in less than twice the number of iterations predicted by the density evolution for different system sizes, as a function of the number of blocks L_c. We used the parameters that lead to the blue curve in Fig. 6.8 (i.e. second line of Table 6.1). As $N \to \infty$, reconstruction is reached in all the instances, as predicted by the state evolution. For finite N, however, reconstruction is not reached when L_c is too large. But in the same time, as L_c increases we can theoretically decode closer to the optimal threshold as seen from (5.115). Thus there is a tradeoff between measurement rate decrease and probability of success in the reconstruction.

basis, but the resulting signal is not truly sparse but compressible (3.22). This can be seen from Fig. 6.11 where we show the sorted coefficients of the 4-step Haar transformed Lena and peppers images, see Fig. 6.12 and Fig. 6.13. It appears that the energy is concentrated on few coefficients but the smaller ones follow a "power law-like" distribution. Compressed sensing is thus appropriate but perfect reconstruction is impossible in the compressed regime $\alpha < 1$. We perform an experiment where the Lena and peppers images are 4-step Haar transformed, and the coefficients Fig. 6.11 are reconstructed for different measurement rates, with a purely sparsifying prior $P_0(\mathbf{x}) = \prod_i^N [(1-\rho)\delta(x_i) + \rho\mathcal{N}(x_i|m,\sigma^2)]$ or with the approximate sparsity prior (6.2). All the prior parameters (and the noise variance in the purely sparse case) are learned through expectation maximisation, see sec. 4.3.8. The results Fig. 6.12 and Fig. 6.13, both in terms of the MSE of the reconstructed wavelets coefficients and in terms of the "by-eyes" quality of reconstruction are homogeneously better with the approximate sparsity prior at any measurement rate. In the case of Lena, it is even stronger as the by-eyes quality of reconstruction of Lena at $\alpha = 0.415$ with the approximate sparsity prior is better than the image reconstructed at $\alpha = 0.65$ with the sparsity inducing prior and both the by-eyes quality

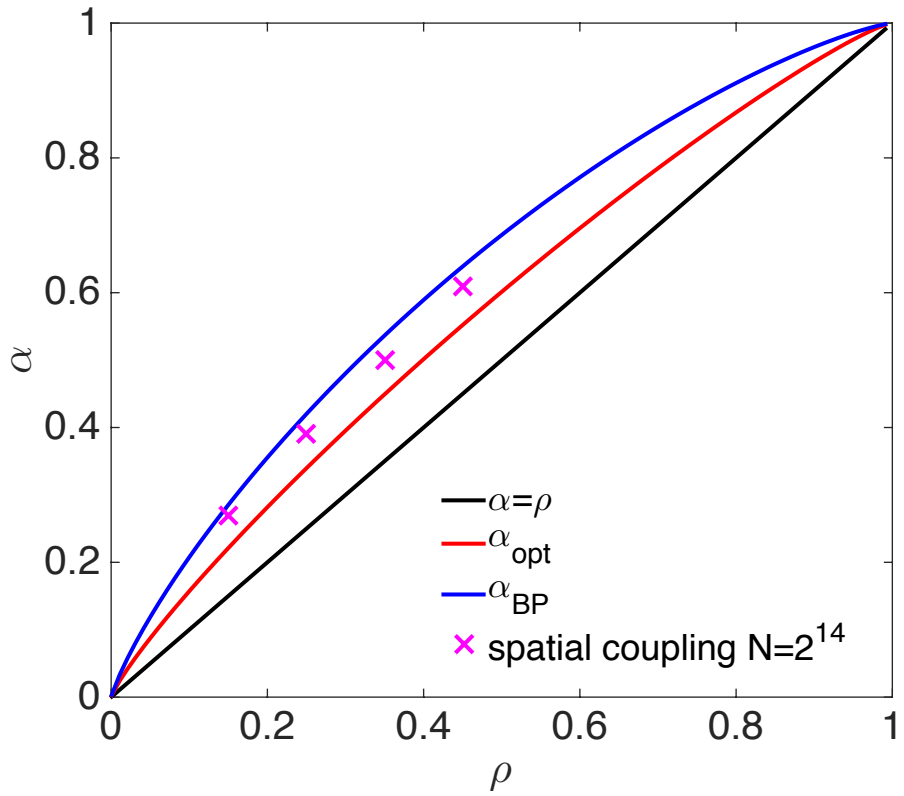

Figure 6.10 – Phase diagram for a small variance $\epsilon = 10^{-6}$ on the (α, ρ) plane, where we added crosses at parameters values where reconstruction have been successful thanks to spatial coupling for signals of size $N = 2^{14}$. As the hard phase between the optimal and BP transitions decreases with ρ and because the gap between the sucessful reconstruction line (the virtual line linking the pink crosses) and $\alpha_{opt}(\rho)$ is close to constant for fixed N, the gain in measurement rate obtained with spatial coupling decreases with ρ. The spatially-coupled random i.i.d Gaussian matrices were drawn from the ensemble ($L_c = 32, L_r = 33, w = 2, \sqrt{J} = 0.4, \alpha, \beta_{seed} = 1.3$).

and MSE performances of reconstruction of Lena at $\alpha = 0.65$ with the approximate sparsity prior are better than the ones at $\alpha = 0.8$ with the sparsity inducing prior. The conclusions are identical with the peppers image. So reconstructing with the approximate sparsity prior images expressed in the wavelet basis appears to be a good strategy. Looking at the pictures, we see that way more details are reconstructed with the approximate sparsity prior. This is due to the fact that these are contained in the high frequency coefficients in the wavelet basis which are hidden in the low energy tail of Fig. 6.11. Because the approximate sparsity model considers this tail as being part of the signal rather than noise, it reconstructs part of these small coefficients, which induce this important gain in detailed imformation.

However, to become competitive with state-of-the-art algorithms [122–126] that can reconstruct wavelet coefficients for images, we also need to find better models for the signal coefficients, likely including the fact that the approximately sparse components are highly

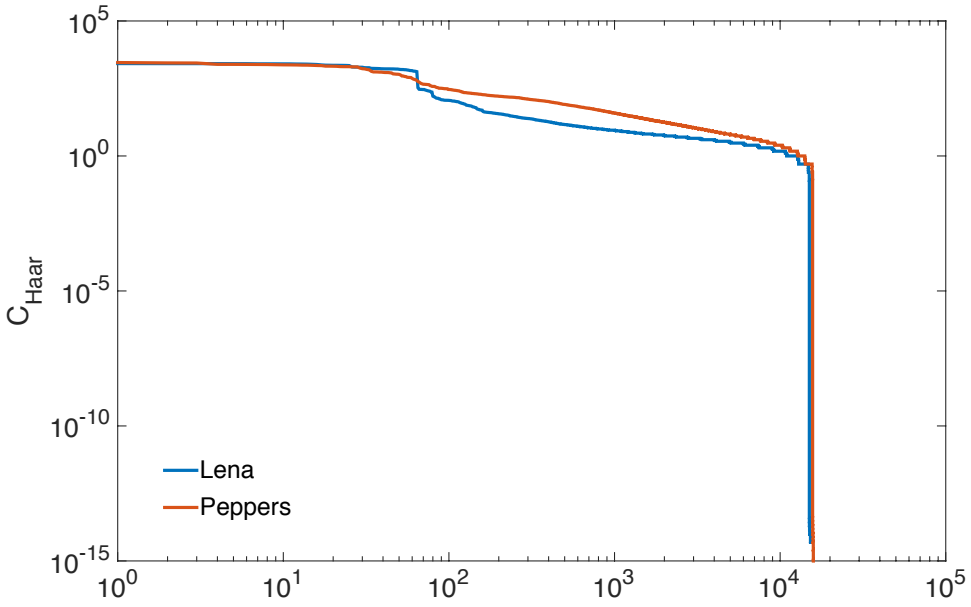

Figure 6.11 – Sorted wavelet spectrum of the 4-steps Haar transformed Lena and peppers images in double logarithmic plot. There are few high amplitude coefficients and a power law tail of smaller coefficients: this is typical of compressible signals.

structured for real images (wavelet coefficients are known to have a tree structure exploited by [122, 123] for example).

6.6 Concluding remarks

At this point we want to state that whereas all our results do depend quantitatively on the statistical properties of the signal, the qualitative features described here (for example the presence and the nature of the phase transitions) are valid for other signal models, distinct from the bi-Gaussian case that we have studied here. This is even the case when the signal model does not match the statistical properties of the actual signal. This was illustrated for example for the noisy compressed sensing of truly sparse signal in [35]. In the same line, we noticed and tested that if AMP corresponding to $\epsilon = 0$ is runned for the approximately sparse signals, then the final MSE is always larger than the one achieved by AMP with the right value of ϵ, as seen on the images.

We studied the case of noiseless measurements, but the measurement noise can be straightforwardly included into the analysis as in [35]. Again, the results would change quantitatively, but not qualitatively. The point was really to understand the influence of the small components alone, as the influence of measurement noise was already extensively studied in compressed sensing [35].

For small variance of the small components of the signal, the AMP algorithm for homoge-

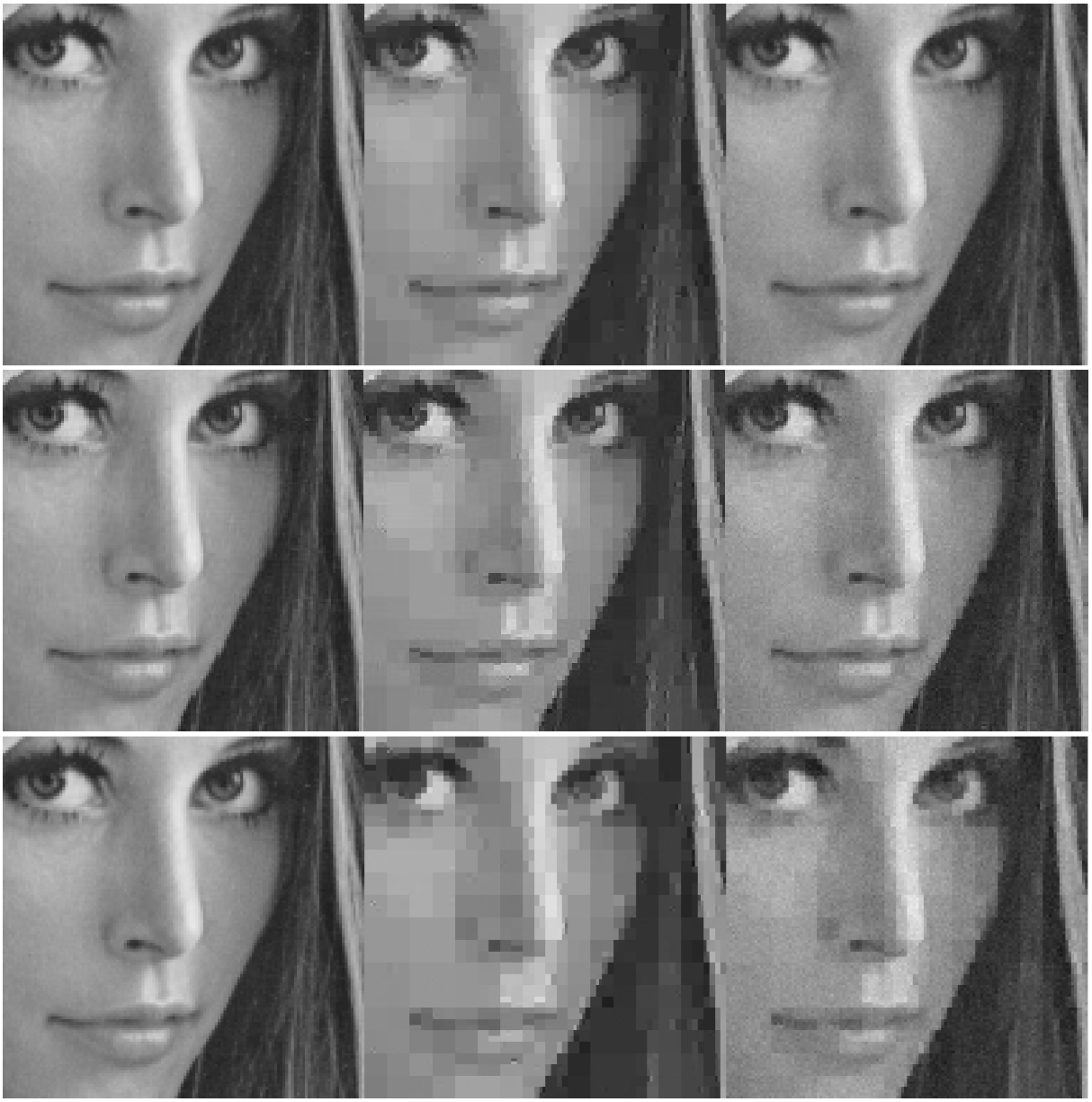

Figure 6.12 – Comparisons between reconstruction results of Lena (left picture). Lena was first expressed in the 4-steps Haar wavelet basis, and the coefficients Fig. 6.11 were then reconstructed with AMP using a strict sparsity inducing prior (center) or the approximate sparsity prior (right) for different measurement rates. The measurement rates and final MSE of the wavelet coefficient are: **Up**: $\alpha = 0.8$, $MSE_{sparse} = 7.2 \times 10^{-4}$, $MSE_{app.sparse} = 2 \times 10^{-4}$, **Middle**: $\alpha = 0.65$, $MSE_{sparse} = 10^{-3}$, $MSE_{app.sparse} = 4.6 \times 10^{-4}$ and **Down**: $\alpha = 0.415$, $MSE_{sparse} = 1.8 \times 10^{-3}$, $MSE_{app.sparse} = 1.3 \times 10^{-3}$.

neous matrices does not reach optimal reconstruction for measurement rates close to the theoretical limit $\alpha_{opt}(\epsilon, \rho)$. The spatial coupling approach, resulting in the design of seeding matrices improves significantly the performances. For diverging system sizes, optimality can

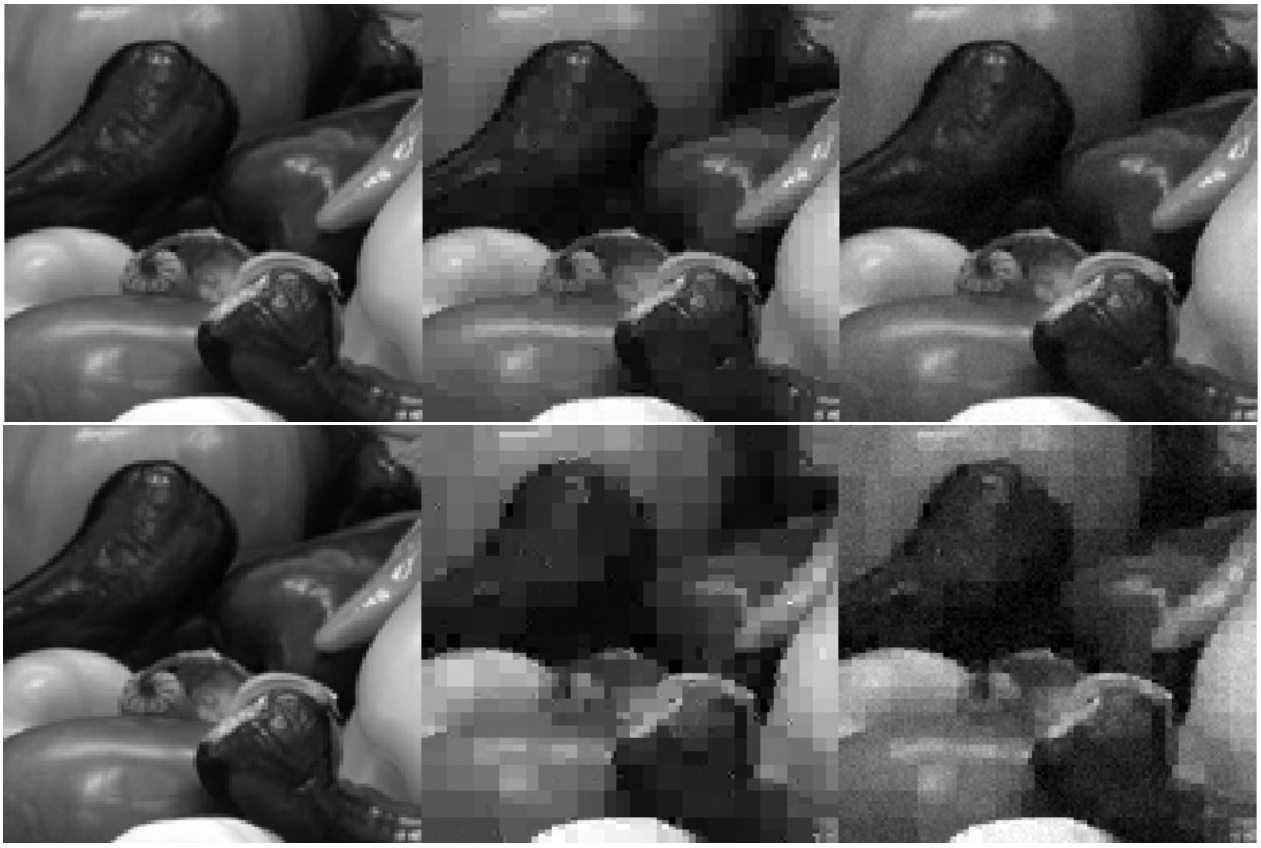

Figure 6.13 – The same experiment as in Fig. 6.12 but with the peppers image. The measurement rates and final MSE of the wavelet coefficient are: **Up**: $\alpha = 0.8$, $MSE_{sparse} = 6.6 \times 10^{-4}$, $MSE_{app.sparse} = 2.4 \times 10^{-4}$ and **Down**: $\alpha = 0.4$, $MSE_{sparse} = 2.7 \times 10^{-3}$, $MSE_{app.sparse} = 2 \times 10^{-3}$.

be restored. We have shown that significant improvement is also reached for sizes of practical interest. There are, however, non negligible finite size effects that should be studied in more details. The optimal design of the seeding matrix for finite system sizes (as studied for instance in depth in the context of error correcting codes [121]) remains an important open question.

7 Approximate message-passing with spatially-coupled structured operators

We now study the behavior of the approximate message-passing algorithm when the i.i.d matrices for which it has been specifically designed are replaced by structured operators, such as Fourier and Hadamard ones. The aim is in one hand to be able to tackle very large single instances of inference problems such as compressed sensing and in the other hand, to reach close to Bayes optimal reconstruction performances.

To work with large signals and matrices, however, one needs fast and memory efficient solvers. Indeed, the mere storage of the measurement matrix in memory can be problematic as soon as the signal size $N > O(10^4)$. A classical trick (see for instance [127]) is thus to replace the random sensing matrix with a structured one, typically random modes of a fast transform such as Fourier-like matrices. We will show empirically that after a proper randomization, the structure of the operators does not significantly affect the performances of the solver.

The use of fast transforms makes matrix multiplications faster ($O(N \log N)$ instead of $O(N^2)$ operations), and thus both speeds up the reconstruction algorithm and removes the need to store the matrix in memory. This is also important for coding applications where $O(N^2)$ operations can be burdensome for the processor.

While using Fourier or Hadamard matrices has often been done with AMP (see for example [128, 129]), we provide here a close examination of its performances with Fourier and Hadamard operators for compressed sensing of complex and real sparse signals respectively. As suggested by the heuristic replica analysis [130, 131], such matrices often lead to better performances than random ones. This will be confirmed through numerical investigation.

Furthermore, inspired by the Gabor construction of [128] that allowed optimal sampling of a random signal with sparse support in frequency domain, we extend the construction of spatially-coupled matrices to a structured form using fast Fourier/Hadamard operators, which allow to deal with large signal sizes and up to the information theoretical limit. Given the lack of theoretical guaranties, we numerically study this strategy on synthetic problems, and compare its performance and behavior with those obtained with random i.i.d Gaussian

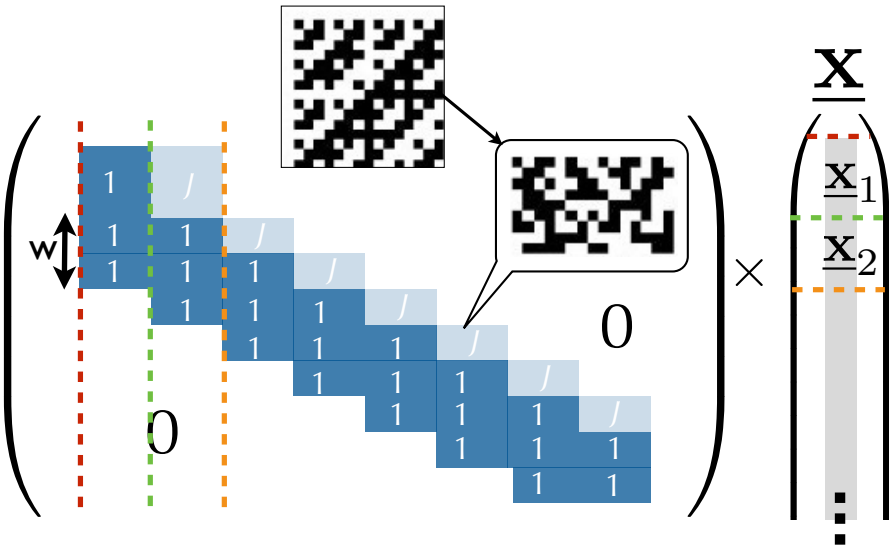

Figure 7.1 – Representation of the spatially-coupled Hadamard sensing matrix used in our study, which structure is the same as Fig. 5.4. The operator is decomposed in $L_r \times L_c$ blocks, each being made of N/L_c columns and $\alpha_{seed} N/L_c$ lines for the blocks of the first block-row, $\alpha_{rest} N/L_c$ lines for the following block-rows, with $\alpha_{seed} > \alpha_{BP} \geq \alpha_{rest} > \alpha_{opt}$. The figure shows how the lines of the original small Hadamard matrix (of size $N/L_c \times N/L_c$) are randomly selected, re-ordered and sign-flipped to form a given block of the final operator. There is a number w (the coupling window) of lower diagonal blocks with elements $\in \{\pm 1\}$ as the diagonal blocks, the upper diagonal blocks have elements $\in \{\pm\sqrt{J}\}$ where $\sqrt{J}$ is the coupling strength, all the other blocks contain only zeros. The colored dotted lines help to visualize the block decomposition of the signal induced by the operator structure: each block of the signal will be reconstructed at different times in the algorithm (see Fig. 7.4 main figure and Fig. 5.4). The procedure is exactly the same for constructing spatially-coupled Fourier operators, replacing the small Hadamard operator from which we construct the blocks by a small Fourier operator. The parameters that define the spatially-coupled operator ensemble are $(L_c, L_r, w, \sqrt{J}, \alpha_{seed}, \alpha_{rest})$.

matrices. The main result is that after some randomization procedure, structured operators appear to be nearly as efficient as random i.i.d matrices. In fact, empirical performances are as good as those reported in [34, 35] despite the drastic improvement in computational time and memory.

7.1 Problem setting

In the following, complex variables will be underlined: $\underline{x}_j = x_{j,1} + i x_{j,2} \in \mathbb{C}$. We will write $\underline{x} \sim \mathcal{CN}(\underline{x}|\underline{\bar{x}}, \sigma^2)$ if the real and imaginary parts x_1 and x_2 of the random variable $\underline{x}$ are independent and verify $x_1 \sim \mathcal{N}(x_1|\Re\underline{\bar{x}}, \sigma^2)$ and $x_2 \sim \mathcal{N}(x_2|\Im\underline{\bar{x}}, \sigma^2)$.

The generic problem we consider is noiseless complex compressed sensing. We consider that the signal components are scalars ($L = N$, $B = 1$). We choose not to consider noise as the point

here is really to study the influence of structure in the matrix. Noise or vectorial components can be trivially included in the model, and do not change the qualitative features presented here. This will be confirmed through the extensive use of the present structured Hadamard operator in the context of the superposition codes chap. 9 where vectorial components signals measured under high noise levels will be reconstructed. In the noiseless complex case (3.18) becomes:

$$\underline{\mathbf{y}} = \underline{\mathbf{F}}\,\underline{\mathbf{s}} \tag{7.1}$$

We will use the following Gauss-Bernoulli distribution to generate ρ-sparse complex random vectors:

$$P(\underline{\mathbf{s}}) = \prod_{j=1}^{N} \left[(1-\rho)\delta(\underline{s}_j) + \rho\, \mathscr{CN}(\underline{s}_j | \underline{\bar{s}}, \sigma^2) \right] \tag{7.2}$$

Here we shall assume that the correct values for ρ, $\underline{\bar{s}}$, σ^2 as well as the empirical signal distribution (7.2) are known and thus we place ourselves under the prior matching condition in the theoretical analyzes. As discussed in sec. 4.3.8, these parameters can be learned efficiently with an expectation maximization procedure if unknown.

7.1.1 Spatially-coupled structured measurement operators

As discussed in sec. 5.5, in order to asymptotically reach the optimal transition α_{opt}, spatial coupling is the strategy of choice. The spatially-coupled structured operator is constructed as Fig. 5.4 with the novelty that the blocks are not made of random i.i.d matrices anymore, but are replaced by sub-sampled fast operators, see Fig.7.1: each of these blocks is constructed from the same original operator of size $N/L_c \times N/L_c$ and the differences from one block to another comes from the selected modes, their permutation and signs that are randomly changed. In the case of an Hadamard construction, all the blocks are generated from the same original small Hadamard operator with the constraint that N/L_c must be a power of two, intrinsic to the Hadamard construction.

7.1.2 The approximate message-passing algorithm for complex signals

In order to avoid confusions with the literature where variations of AMP are already presented, we will refer in the present chapter to the Bayes-optimal AMP by "BP" and "c-BP" for the real and complex case respectively, and to the ℓ_1-minimizing version by "LASSO" and "c-LASSO" respectively. As the thresholding functions are applied component-wise, the time-consuming part of the algorithm are the matrix multiplications in the linear step. Here, we use Fourier and Hadamard operators in order to reduce their complexity from $O(N^2)$ to $O(N \log N)$. The authors of [128] have used a related, yet different way to create spatially-coupled matrices using a set of Gabor transforms.

The generalization of the scalar algorithm Fig. 5.5 to the complex case is not sraightforward as in the derivation of sec. 4.3.3, the real and imaginary parts of the different complex quantities appearing in the computations would couple through non diagonal covariance matrices. But it appears that many simplifications arise due to the independence assumption of the matrix elements, which make the final algorithm Fig. 7.2 look very similar to its scalar version. The derivation and study of this complex version can be found in [132–135]. The four different operators (5.116), (5.117), (5.118), (5.119) are respectively generalized to the complex case as:

$$\tilde{O}_\mu(\mathbf{e}_c) := \sum_{i\in c}^{N/L_c} |\underline{F}_{\mu i}|^2 e_i \tag{7.3}$$

$$O_\mu(\mathbf{e}_c) := \sum_{i\in c}^{N/L_c} \underline{F}_{\mu i} e_i \tag{7.4}$$

$$\tilde{O}_i(\mathbf{f}_r) := \sum_{\mu\in r}^{\alpha_r N/L_c} |\underline{F}_{\mu i}|^2 f_\mu \tag{7.5}$$

$$O_i^*(\mathbf{f}_r) := \sum_{\mu\in r}^{\alpha_r N/L_c} \underline{F}_{\mu i}^* f_\mu \tag{7.6}$$

where $\underline{F}_{\mu i}^*$ is the complex conjugate of $\underline{F}_{\mu i}$. Because the value of $|\underline{F}_{\mu i}|^2$ is either 0, 1 or J $\forall$ (μ, i) depending on the block as we use Hadamard or Fourier operators (it can be read on Fig. 7.1), all these operators do not require matrix multiplications as they are implemented as fast transforms (O_μ and O_i^*) or simple sums ($\tilde{O}_\mu$ and $\tilde{O}_i$). It results in the updates for complex AMP [133] with a generic operator, see Fig. 7.2.

Here, we give the functions $\underline{f_a}$ and f_c that are calculated by Gaussian integration from (7.2) and are thus Bayes-optimal, which is not the case for LASSO and c-LASSO [132] as discussed in sec. 5.1.1. For BP, they are trivially constructed from sec. 4.3.6. For c-BP, the signal is complex and drawn from the distribution (7.2), and the thresholding functions (which give posterior scalarwise estimates of the mean and variance) are given by:

$$\underline{f_a}(\Sigma^2, \underline{R}) = g\rho\chi^2 \underline{M}/Z \tag{7.7}$$

$$f_c(\Sigma^2, \underline{R}) = \left(g\rho\chi^2\left(|\underline{M}|^2 + 2\chi^2\right)/Z - |\underline{f_a}(\Sigma^2, \underline{R})|^2\right)/2 \tag{7.8}$$

together with the following definitions:

$$\underline{M} := (\sigma^2 \underline{R} + \Sigma^2 \underline{\bar{x}})/(\Sigma^2 + \sigma^2) \tag{7.9}$$

$$\chi^2 := \Sigma^2\sigma^2/(\Sigma^2 + \sigma^2) \tag{7.10}$$

$$g := \exp\left(-\frac{1}{2}\left(\frac{|\underline{\bar{x}}|^2}{\sigma^2} + \frac{|\underline{R}|^2}{\Sigma^2} - \frac{|\underline{M}|^2}{\chi^2}\right)\right) \tag{7.11}$$

$$Z := \sigma^2(1-\rho)\exp\left(-\frac{|\underline{R}|^2}{2\Sigma^2}\right) + \rho\chi^2 g \tag{7.12}$$

where $\underline{R}$ and Σ^2 are the AMP fields and we have $\underline{\bar{x}} = \underline{\bar{s}}$. These functions are not identical to

$t \leftarrow 0$
$\delta \leftarrow \epsilon + 1$
while $t < t_{max}$ **and** $\delta > \epsilon$ **do**
  $\Theta_\mu^{t+1} \leftarrow \sum_c^{L_c} \tilde{O}_\mu(\mathbf{v}_c^t)$
  $\underline{w}_\mu^{t+1} \leftarrow \sum_c^{L_c} O_\mu(\underline{\mathbf{a}}_c^t) - \Theta_\mu^{t+1} \frac{\underline{y}_\mu - \underline{w}_\mu^t}{\Delta + \Theta_\mu^t}$
  $\Sigma_i^{t+1} \leftarrow \left[\sum_r^{L_r} \tilde{O}_i\left([\Delta + \mathbf{\Theta}_r^{t+1}]^{-1}\right)\right]^{-1/2}$
  $\underline{R}_i^{t+1} \leftarrow \underline{a}_i^t + (\Sigma_i^{t+1})^2 \sum_r^{L_r} O_i^* \left(\frac{\underline{\mathbf{y}}_r - \underline{\mathbf{w}}_r^{t+1}}{\Delta + \mathbf{\Theta}_r^{t+1}}\right)$
  $v_i^{t+1} \leftarrow f_c\left((\Sigma_i^{t+1})^2, \underline{R}_i^{t+1}\right)$
  $\underline{a}_i^{t+1} \leftarrow \underline{f_a}\left((\Sigma_i^{t+1})^2, \underline{R}_i^{t+1}\right)$
  $t \leftarrow t + 1$
  $\delta \leftarrow < |\underline{\mathbf{a}}^{t+1} - \underline{\mathbf{a}}^t|^2 >$
end while
return $\underline{\mathbf{a}}^t$

Figure 7.2 – The complex AMP algorithm written with operators. Depending on whether it is used on a real or complex signal, with Bayes-optimal or sparsity-inducing thresholding functions $\underline{f_a}$ and f_c, we call it BP, c-BP, LASSO or c-LASSO. ϵ is the accuracy for convergence and t_{max} the maximum number of iterations. A suitable initialization for the quantities is $(\underline{a}_i^{t=0} = \mathbb{E}_{P_0}(x) = 0,\ v_i^{t=0} = \mathrm{Var}_{P_0}(x) = \rho\sigma^2,\ \underline{w}_\mu^{t=0} = \underline{y}_\mu)$ where we have used the prior (7.2). Once the algorithm has converged, i.e. the quantities do not change anymore from iteration to iteration, the estimate of the i^{th} signal component is $\underline{a}_i^t$. The nonlinear thresholding functions $\underline{f_a}$ and f_c take into account the prior distribution. In the case of compressed sensing, applying $\underline{f_a}$ to a $\underline{R}_i^{t+1}$ close to zero will give a result even closer to zero, while bigger inputs will be left nearly unchanged, thus favoring sparse solutions. If needed, the damping scheme of Fig. 4.6 can be used.

the ones for the real case since in the prior distribution (7.2), the real and imaginary parts of the signal are jointly sparse (i.e. have same support but independent values), which can be a good assumption, for instance in MRI. As in c-LASSO [132], because the joint sparsity is more constrained and thus bring more information, it allows to lower the phase transition compared to when the real and imaginary parts of the signal are assumed to be fully independent.

7.1.3 Randomization of the structured operators

The implementation requires caution: the necessary "structure killing" randomization of the fast structured operator used to construct the blocks of the matrix Fig. 7.1 is obtained by applying a permutation of lines after the use of the fast operator. For each block (r,c), we choose a random subset of modes $\Omega^{r,c} = \{\Omega_1^{r,c}, \ldots, \Omega_{N_r}^{r,c}\} \subset \{1, \ldots, N_c\}$. The definition of $O_\mu(\mathbf{e}_c)$ using a standard fast transform FT will be:

$$O_\mu(\mathbf{e}_c) := (\mathrm{FT}(\mathbf{e}_c))_{\Omega_{\mu-\mu_{r_\mu}+1}^{r_\mu,c}} \tag{7.13}$$

where r_μ is the index of the block-row that includes the index μ, μ_{r_μ} is the number of the first line of the block row r_μ and $(\mathbf{u})_\mu$ is the μ^{th} component of $\mathbf{u}$. For $O_i^*(\mathbf{f}_r)$ instead:

$$O_i^*(\mathbf{f}_r) := \left(\text{FT}^{-1}(\tilde{\mathbf{f}}_r)\right)_{i-i_{c_i}+1} \tag{7.14}$$

where c_i is the index of the block-column that includes the index i, i_{c_i} is the number of the first column of the block-column c_i, FT^{-1} is the standard fast inverse operator of FT and $\tilde{\mathbf{f}}_r$ is defined in the following way:

$$\forall \gamma \in \{1,\dots,N_r\}, \quad \left(\tilde{\mathbf{f}}_r\right)_{\Omega_\gamma^{r,c_i}} = (\mathbf{f}_r)_\gamma \quad \text{and} \quad \forall k \notin \Omega^{r,c_i}, \left(\tilde{\mathbf{f}}_r\right)_k = 0 \tag{7.15}$$

The MSE achieved by the algorithm is:

$$E^t := ||\underline{\mathbf{a}}^t - \underline{\mathbf{s}}||_2^2 = < |\underline{\mathbf{a}}^t - \underline{\mathbf{s}}|^2 > \tag{7.16}$$

and measures how well the signal is reconstructed.

7.2 Results for noiseless compressed sensing

When the sensing matrix is i.i.d random, or spatially-coupled with i.i.d random blocks, the evolution of E^t in AMP is predicted in the large signal limit on a rigorous basis by the state evolution [104, 110, 136], see sec. 5.3. For c-BP with i.i.d Gaussian matrices, the derivation goes very much along the same lines and we shall report the results briefly. The generalization to the complex case of the state evolution (5.105) for homogeneous matrices under the prior (7.2) is given by the following recursion:

$$E^{t+1} = \int \mathscr{D}\underline{z} \left[(1-\rho) f_c\left((\Sigma^{t+1})^2, \underline{R}_1^{t+1}(\underline{z})\right) + \rho f_c\left((\Sigma^{t+1})^2, \underline{R}_2^{t+1}(\underline{z})\right)\right] \tag{7.17}$$

together with:

$$\underline{z} := z_1 + i z_2 \tag{7.18}$$

$$(\Sigma^{t+1})^2 := (\Delta + E^t)/\alpha,$$

$$\underline{R}_u^{t+1}(\underline{z}) := \underline{z}\sqrt{\sigma^2 \delta_{u,2} + (\Sigma^{t+1})^2} \tag{7.19}$$

$$\mathscr{D}\underline{z} := dz_1 dz_2 \frac{e^{-\frac{1}{2}(z_1^2+z_2^2)}}{2\pi}. \tag{7.20}$$

with $\Delta = 0$ in the noiseless case. (7.17) has been obtained exactly as in the previous chapter when we derived (6.12). Note that this state evolution equation is the same as given in [132], despite slightly different update rules in the algorithm.

For c-BP with spatially-coupled matrices with i.i.d Gaussian blocks, the expression involves the MSE in each block $c' \in \{1,\dots,L_c\}$, see sec. 5.7. The generalization of (5.175) ,(5.176) and

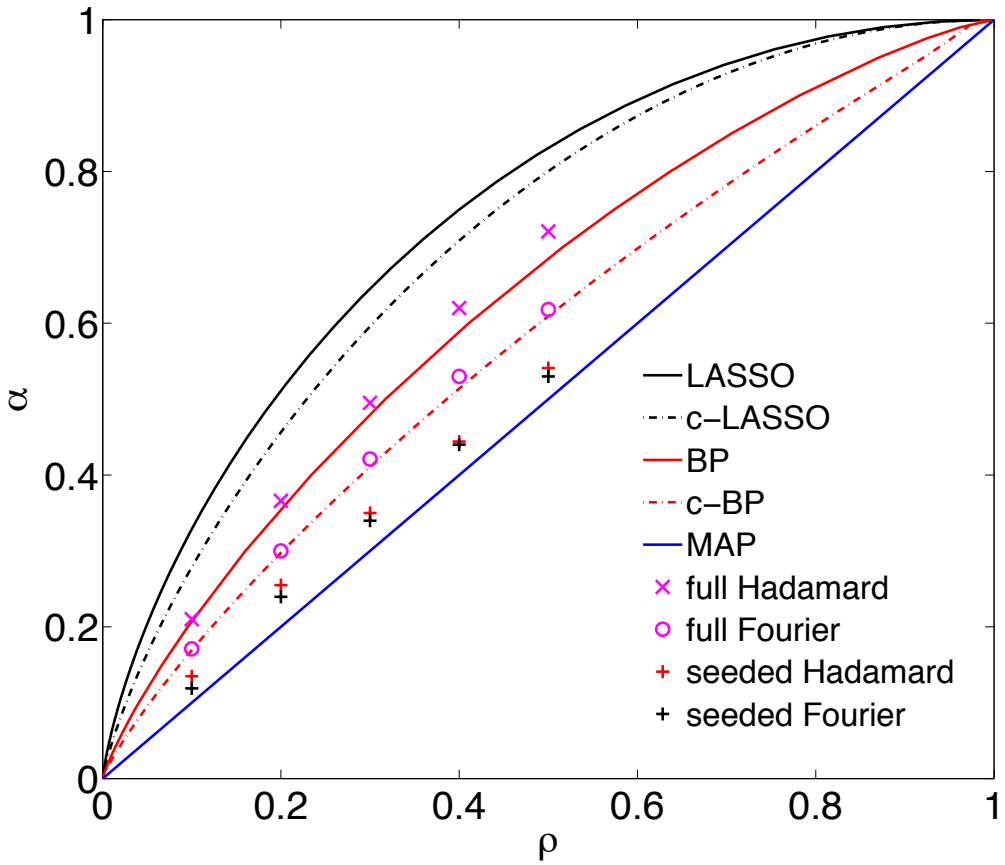

Figure 7.3 – Phase diagram on the (α, ρ) plane for noiseless $\Delta = 0$ compressed sensing. Lines are phase transitions predicted by the state evolution analysis for i.i.d random Gaussian matrices, while markers are points from experiments using structured operators with empirically optimized parameters. Good sets of parameters usually lie in the following sets: $(L_c \in \{8,16,32,64\}, L_r = L_c + \{1,2\}, w \in \{2,\dots,5\}, \sqrt{J} \in [0.2,0.7], \beta_{seed} \in [1.2,2])$, see (5.114). As discussed in the previous chapter, with larger signals, higher values of L_c are better as it allows to get closer to the optimal transition. Just as c-LASSO allows to improve the usual LASSO phase transition when the complex signal is sampled according to (7.2) (thanks to the joint sparsity of the real and imaginary parts), c-BP improves the usual BP transition. The line $\alpha = \rho$ is both the maximum-a-posteriori MAP threshold for noiseless compressed sensing and the (asymptotic) optimal phase transition that can be reached with spatially-coupled matrices. Pink experimental points correspond to perfectly reconstructed instances using homogeneous Hadamard and Fourier operators (on the BP and c-BP phase transition respectively), the black and red crosses using spatially-coupled ones (close to the MAP threshold). Properly randomized structured operators appear to have similar performances as random measurement matrices.

(5.178) to the complex case under the prior (7.2) is given by:

$$E_c^{t+1} = \int \mathscr{D}\underline{z} \left[(1-\rho) f_c\left((\Sigma_c^{t+1})^2, \underline{R}_{c,1}^{t+1}(\underline{z}) \right) + \rho f_c\left((\Sigma_c^{t+1})^2, \underline{R}_{c,2}^{t+1}(\underline{z}) \right) \right] \tag{7.21}$$

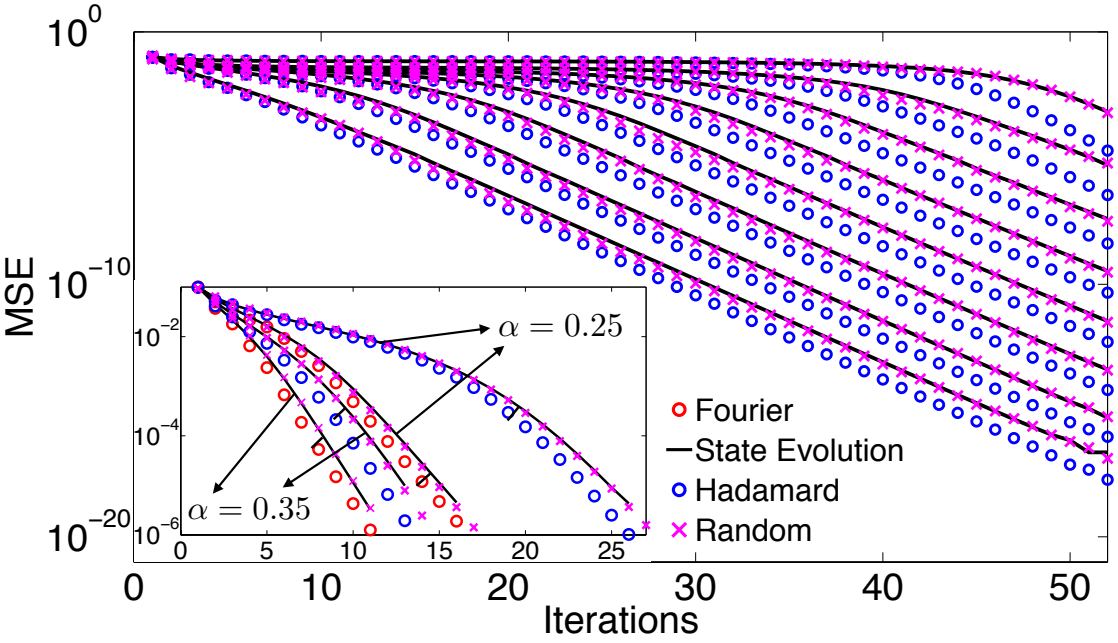

Figure 7.4 – Comparison of the mean square error predicted by the state evolution (black curves) and the actual behavior of the algorithm for spatially-coupled matrices (main figure) and standard homogeneous ones (inset), both with structured operators (circles) and random i.i.d Gaussian matrices (purple crosses) in a noiseless compressed sensing setting. In both plots, the signal size is $N = 2^{14}$ with the random i.i.d Gaussian matrices, and $N = 2^{20} \approx 10^6$ with the operators and are generated with ($\rho = 0.1$, $\underline{\bar{x}} = 0$, $\sigma^2 = 1$). While experiments made with random i.i.d matrices fit very well the asymptotic predictions, those with the structured operators are not described well by the state evolution, although final performances are comparable. **Main:** For an Hadamard spatially-coupled matrix as in Fig. 7.1 with ($L_c = 8$, $L_r = L_c + 2$, $w = 1$, $\sqrt{J} = 0.1$, $\alpha = 0.22$, $\beta_{seed} = 1.36$). Each curve corresponds to the MSE tracked in a different block of the real reconstructed signal $\mathbf{s}$ (see Fig. 7.1). **Inset:** Reconstructions made with structured homogeneous matrices at $\alpha = 0.35$ and $\alpha = 0.25$. The reconstruction with the Fourier operator of a complex signal (instead of real with Hadamard) is faster thanks to the joint sparsity assumption of (7.2). The arrows identify the groups of curves corresponding to same measurement rate α. Both in the Fourier and Hadamard cases, we observe that convergence is slightly faster than in the random i.i.d case as predicted by the state evolution.

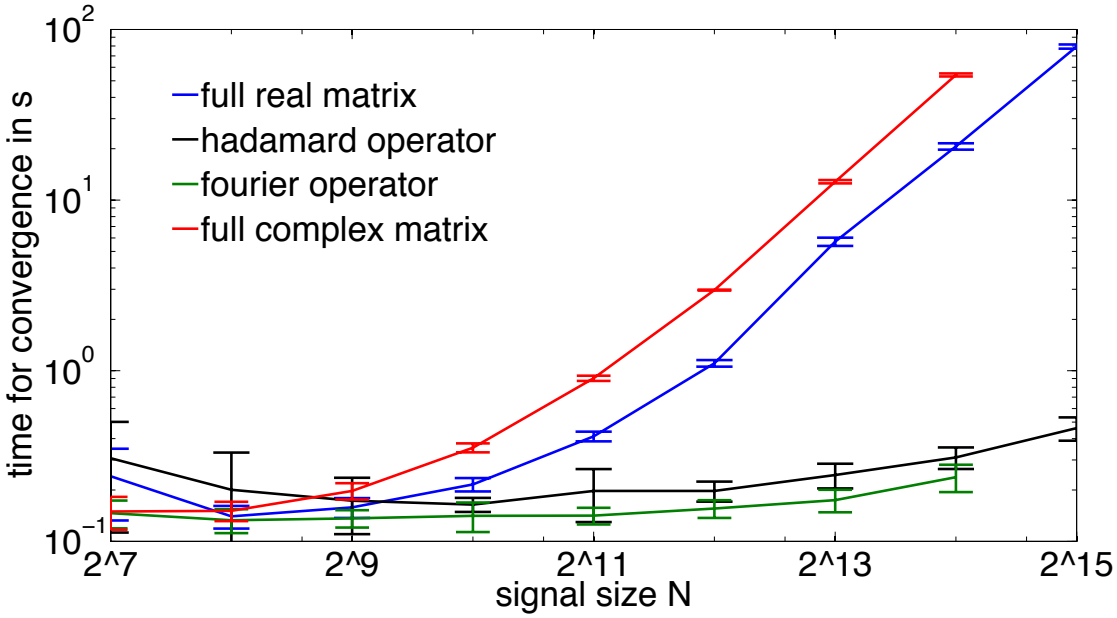

Figure 7.5 – Time required for convergence (i.e. $MSE < 10^{-6}$) of the AMP algorithm in seconds as a function of the signal size, in the homogeneous matrix case for a typical compressed sensing problem. The signal has distribution given by (7.2) and is real (complex) for the reconstruction with real (complex) matrices. The plot compares the speed of AMP with matrices (blue and red lines) to those of AMP using the structured operators (black and green lines). The points have been averaged over 10 random instances and the error bars represent the standard deviation with respect to these. The simulations have been performed on a personal laptop. As the signal size increases, the advantage of using operators becomes quickly obvious.

where:

$$\Sigma_c^{t+1}\left(\{E_{c'}^t\}_{c'}^{L_c}\right) = \left[\sum_r^{L_r} \frac{\alpha_r J_{rc}}{L_c\Delta + \sum_{c'}^{L_c} J_{rc'} E_{c'}^t}\right]^{-1/2} \tag{7.22}$$

$$\underline{R}_{c,u}^{t+1}(\underline{z}) = \underline{z}\sqrt{\sigma^2\delta_{u,2} + (\Sigma_c^{t+1})^2} \tag{7.23}$$

and $\alpha_r = \alpha_{rest} + (\alpha_{seed} - \alpha_{rest})\delta_{r,1}$, J_{rc} is the variance of the elements belonging to the block at the r^{th} block-row and c^{th} block-column (1, J or 0 in Fig. 7.1) and again $\Delta = 0$ in the noiseless case.

We now move to our main point. In the case of AMP with structured (Fourier or Hadamard) operators instead of i.i.d matrices, the state evolution analysis cannot be made. Hence we experimentally compare the performances between AMP with structured operators and i.i.d matrices. The comparison is shown in Fig. 7.4 which presents theoretical results from the state evolution and experimental ones obtained by running AMP on finite size signals. On Fig. 7.3, we show the phase transition lines obtained by state evolution analysis in the (α, ρ) plane, and we added markers showing the position of instances actually recovered by the algorithm

with spatially-coupled structured operators in the noiseless case $\Delta = 0$. It appears that with structured operators, AMP is still able to decode really close to the optimal threshold as with random i.i.d matrices.

7.2.1 Homogeneous structured operators

Let us first concentrate on AMP with homogeneous (or full) structured operators. The first observation is that the state evolution *does not* correctly describe the evolution of the MSE for AMP with full structured operators (inset Fig. 7.4). It is perhaps not surprising, given that AMP has been derived for i.i.d matrices. The difference is small, but clear: E^t decreases faster with structured operators than with i.i.d matrices. However, despite this slight difference in the dynamical behavior of the algorithm, the phase transitions and the final MSE performances for both approaches appear to be extremely close. As seen in Fig. 7.3, for small ρ, we cannot distinguish the actual phase transition with structured operators from the one predicted by state evolution. Thus, the state evolution analysis remains a good predictive tool of the AMP performances with structured operators.

7.2.2 Spatially-coupled structured operators

For spatially-coupled operators, the conclusions are similar (main plot on Fig. 7.4). Again, E_c^t in each of the blocks of the signal, induced by the spatially-coupled structure of the measurement matrix, decreases faster with structured operators than with random i.i.d matrices. But our empirical results are consistent (see Fig. 7.3) with the hypothesis that the proposed scheme, using spatially-coupled Fourier/Hadamard operators, achieves correct reconstruction as soon as $\alpha > \rho$ when N is large. Indeed, we observe that the gap to the MAP threshold (the optimal threshold in the noiseless case) $\alpha_{opt}(\rho) = \rho$ decreases as the signal size increases upon optimization of the spatially-coupled operator structure. The results in Fig. 7.3 and Fig. 7.4 are obtained with spatially-coupled matrices of the ensemble: $(L_c = 8, L_r = L_c + 1, w \in \{1,2\}, \sqrt{J} \in [0.2, 0.5], \beta_{seed} = [1.2, 1.6])$. While these parameters do not quite saturate the bound $\alpha = \rho$ (which is only possible for $L_c \to \infty$ [34, 108, 110], see sec. 5.115), they do achieve near optimal performances. This, as well as the substantial cut in running time (Fig. 7.5) with respect to AMP with i.i.d matrices and the possibility to work with very large systems without saturating the memory strongly supports the advantages of the proposed implementation of AMP.

7.3 Conclusion

We have presented a large empirical study using structured Fourier and Hadamard operators in sparse linear estimation. We have shown that combining these operators with a spatial coupling strategy allows to get very close to the information-theoretical limits. We have tested our algorithm for noiseless compressed sensing of real and complex signals. The resulting algorithm is more efficient than Fig. 5.5 both in terms of memory and running time. This

allows us to deal with signal sizes as high as 10^6 and a measurement rate $\alpha \approx \rho$ on a personal laptop using MATLAB, and achieve perfect reconstruction in about a minute.

8 Approximate message-passing for compressive imaging

In the present chapter, we focus on two distincts applications of the approximate message-passing algorithm to compressive imaging. The first one is the reconstruction of sub-sampled natural images which have a sparse discrete gradient, while the second one focuses on the reconstruction of point like objects, and thus directly sparse in the pixel domain, measured by fluorescence microscopy technique.

8.1 Reconstruction of natural images in the compressive regime by "total-variation-minimization"-like approximate message-passing

Total variation (TV) and the associated gradient-optimization-based algorithms have been the long-standing state-of-the-art approach to the reconstruction of images from compressive measurements. For signals well-modeled by i.i.d priors, many recent works have presented optimal, or near optimal reconstruction algorithms built instead from a statistical framework. Here, we present a method for incorporating a TV-like structured prior in the image domain to allow for $MMSE$ image recovery using the approximate message-passing algorithm.

In gradient-optimization-based methods for image reconstruction, instead of searching for a sparse signal in the pixel domain as in (8.2), we seek for a signal with a sparse discrete gradient. This kind of signals are supposed to represent well natural images which are piecewise constant or smooth. Specific algorithms to tackle image reconstruction in the compressive regime have been designed on the optimization side. For example the TV-AL3 [137] is very robust to noise and reduction of the measurement rate but is also fast. On the probabilistic side, the phase transitions for TV were investigated and a method, TV-AMP [138], was proposed. Furthermore, [139] proposed a structured prior in conjunction with the co-sparse model and developed the GrAMPA algorithm based on the approximate message-passing algorithm. Finally, some recent work have proposed a joint prior defined over nearest-neighbors for 1-d signals in SS-AMP [140]. Motivated by the very good performances of both [140] and [139], we develop here an approximate message-passing algorithm for 2-d piecewise smooth signals reconstruction in the compressive regime.

The methodology presented here is closely related to the GrAMPA algorithm [139] as we will work also with the co-sparse model, defined in the next section, and AMP. The differences with [139] are essentially coming from the learning procedure of the noise that we use which allows for better reconstruction results, and the fact that we use a Gauss-Bernoulli prior for the dual variables (that represent the differences between neighboring pixels), whereas GrAMPA uses the SNIPE prior, the limiting distribution of the Gauss-Bernoulli one when the variance of the Gaussian part goes to infinity. As we will see, this prior does not improve on the Gauss-Bernoulli one, they give similar results. The improvement with respect to GrAMPA really comes from the noise learning which acts as a kind of annealing.

We will show through intensive numerical experiments that when we push the TV-AL3 algorithm (considered as the state-of-the-art optimization algorithm) to its limits by optimizing all of its parameters, it gives almost exactly the same reconstruction results than our implementation which requires way less tuning. We will observe the same when using the SNIPE prior of GrAMPA in our implementation in conjunction with the noise learning. The results are so close that it suggests that we are reaching the limits of classical reconstruction methods based on first neighbors interactions.

8.1.1 Proposed model

We consider the model (3.18) where now **s** is the reshaped image of size N (the original image being of size $\sqrt{N}\times\sqrt{N}$). In order to mimic the total-variation-minimization idea that enforces neighbors pixels to have identical or closeby values, we naturally take a Gauss-Bernoulli prior over the *differences* of the pixel values:

$$P_0(\mathbf{s}|\sigma^2) \propto \prod_{(ij)\in E} \left[(1-\rho)\delta(s_i - s_j) + \rho\mathcal{N}\left(s_i - s_j|0,\sigma^2\right)\right] \tag{8.1}$$

where $E := \{(ij) : i \text{ neighbor to } j \text{ in the picture}\}$ is the set of pairs of pixels that are neighbors in the original image. This natural extension to the bi-dimensionnal case of [140] suffers from convergence issues due to the short loops in the associated factor graph (a grid) that are created by these two-pixels dependent prior terms. A natural idea to face this problem is to use new auxilliary variables. Defining $\{d_{ij} := x_i - x_j : (ij) \in E\}$ as the differences variables, refered as *dual* variables, and $\mathbf{x} := [\mathbf{s}, \mathbf{d}]$ as the concatenation of the vectorized original image and the vector of dual variables, we now get back a factorized prior where the s_i's are independent with uniform prior and the Gauss-Bernoulli prior is over the dual variables:

$$P_0(\mathbf{x}|\sigma^2,\mathscr{S}) \propto \prod_{(ij)\in E} \left[(1-\rho)\delta(d_{ij}) + \rho\mathcal{N}\left(d_{ij}|0,\sigma^2\right)\right] \prod_{i=1}^{N} \mathscr{U}(s_i|\mathscr{S}) \tag{8.2}$$

where $\mathscr{U}(s_i|\mathscr{S})$ is a uniform distribution over the set $\mathscr{S}$. From now on when the index i of the **x** component is such that if $i \leq N$ it correponds to an image pixel variable, else it is a dual

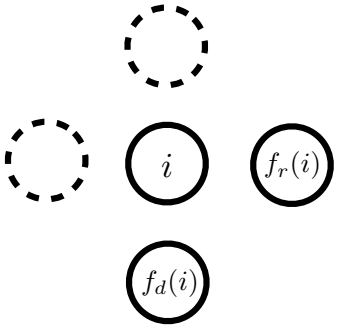

Figure 8.1 – The pixel i with its four closest neighbors, but when considering i in the operator $\tilde{\mathbf{F}}$ construction (8.5), only the interaction with its right neighbor $f_r(i)$ and down one $f_d(i)$ are taken into account not to consider twice each interactions. This choice is arbitrary and considering for example the left and up interactions would be the same, one just needs to design properly the mapping functions f which encode the dependency structure between the image pixels and allow to design the difference matrix $\mathbf{D}$ in (8.5).

variable. Thus the prior can be re-written as:

$$P_0(\mathbf{x}|\sigma^2,\mathscr{S}) \propto \prod_{i=1}^{N+|E|} P_0^i(x_i|\sigma^2,\mathscr{S}) \tag{8.3}$$

$$\propto \prod_{i=1}^{N+|E|} \left[\mathbb{I}(i\leq N)\mathscr{U}(x_i|\mathscr{S}) + \mathbb{I}(i>N)\left[(1-\rho)\delta(x_i) + \rho\mathscr{N}\left(x_i|0,\sigma^2\right)\right]\right] \tag{8.4}$$

and the linear constraints $\{d_{ij}-(s_i-s_j)=0\}_{(ij)\in E}$ are enforced through an extension of the original linear system (3.18) that becomes:

$$\begin{pmatrix} \mathbf{y}_{M,1} \\ \mathbf{0}_{|E|,1} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{M,N} & \mathbf{0}_{M,|E|} \\ \mathbf{D}_{|E|,N} & -\boldsymbol{I}_{|E|,|E|} \end{pmatrix} \begin{pmatrix} \mathbf{s}_{N,1} \\ \mathbf{d}_{|E|,1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\xi}_{M,1} \\ \mathbf{0}_{|E|,1} \end{pmatrix}$$
$$\Leftrightarrow \tilde{\mathbf{y}} = \tilde{\mathbf{F}}\mathbf{x} + \tilde{\boldsymbol{\xi}} \tag{8.5}$$

where the dimension of each vector or matrix has been indicated to avoid confusions and where the tilde stands for these new extended objects. $\mathbf{F}$ is the original operator, $\mathbf{0}_{a,b}$ is a matrix full of zeros of dimensions $a\times b$ and $\boldsymbol{I}_{a,a}$ is the identity matrix if size $a\times a$. $\mathbf{D}$ is the "difference" matrix which is the concatenation of smaller matrices constructed as follows. In the present case we consider for each pixel its four closest neighbors (its left/right and up/down ones on the image). $\mathbf{D} := [\mathbf{D}_r, \mathbf{D}_d]^\intercal$ is the concatenation of $\mathbf{D}_r$ which is made of zeros everywhere except on the diagonal where there are 1's, and -1's on the $\{(i, f_r(i)) : i \in \{1,\dots,N\}, f_r(i)\neq 0\}$ elements where $f_r(i)$ outputs the index (in the vectorized form of $\mathbf{x}$) of the right neighbor of the i^{th} pixel if it actually has a neighbor, 0 else (it is the same for constructing $\mathbf{D}_d$ thanks to the mapping $f_d(i)$), see Fig. 8.1. This new system to be solved is refered as the co-sparse model.

8.1.2 The Hadamard operator and the denoisers

Images are large signals so being able to deal with very large matrices with $O(N^2)$ elements is an issue in itself. We can thus use fast Hadamard-based operators of the form Fig. 5.4, see chap. 7 for a full study of their reconstruction abalities. Because of the fact that each Hadamard mode except the first one (which is a line of ones) have exactly zero mean, the system (8.5) would be invariant by a constant shift in the signal components without this mode. To break this symmetry we enforce the first mode to be present which fixes the mean of the signal, the other ones being selected totally randomly. This issue was never present in the other problems treated in this thesis because the prior was directly on the signal components whereas here, its on the differences (the prior is uniform over the pixels): the prior is invariant by a constant shift in the values of two neighbors and thus of the overall signal.

Once $\mathbf{F}$ is designed (as well as $\mathbf{F}^\intercal, \mathbf{F}^2, (\mathbf{F}^2)^\intercal$ required by AMP, see Fig. 5.5), it is trivial to implement $\tilde{\mathbf{F}}$ (and $\tilde{\mathbf{F}}^\intercal, \tilde{\mathbf{F}}^2, (\tilde{\mathbf{F}}^2)^\intercal$). $\tilde{\mathbf{F}}$ is also fast and does not generate memory issues because all its parts (except $\mathbf{F}$ which is already a fast Hadamard-based operator) are highly sparse.

Now the co-sparse model (8.5) to solve is well defined with a factorized prior (8.2), we can use AMP, Fig. 5.5. Despite the extended measurement matrix $\tilde{\mathbf{F}}$ is sparse and the derivation of AMP in sec. 4.3.3 is based on the high density of the matrix, nothing prevents us to use it anyway and as we will see, it gives very good results. From the definition of the denoising functions (4.115), (4.116) and using the prior (8.4), we obtain:

$$f_{a_i}(\Sigma_i^2, R_i) := \int dx_i\, x_i P_0(x_i|\sigma^2, \mathscr{S}) \mathcal{N}\left(x_i|R_i, \Sigma_i^2\right) \tag{8.6}$$

$$= R_i\, \mathbb{I}(i \leq N) + \tilde{f}_{a_i}(\Sigma_i^2, R_i)\, \mathbb{I}(i > N) \tag{8.7}$$

$$f_{c_i}(\Sigma_i^2, R_i) := \int dx_i\, x_i^2 P_0(x_i|\sigma^2, \mathscr{S}) \mathcal{N}\left(x_i|R_i, \Sigma_i^2\right) - f_{a_i}(\Sigma_i^2, R_i)^2 \tag{8.8}$$

$$= \Sigma_i^2\, \mathbb{I}(i \leq N) + \tilde{f}_{c_i}(\Sigma_i^2, R_i)\, \mathbb{I}(i > N) \tag{8.9}$$

where the explicit form of $\tilde{f}_{a_i}$ and $\tilde{f}_{c_i}$ in this Gauss-Bernoulli case are obtained through the construction of sec. 4.3.6:

$$\tilde{f}_{a_i}(\Sigma^2, R) = \frac{1}{z_i(\Sigma^2, R)} \frac{m\Sigma^2 + R\sigma^2}{\Sigma^2 + \sigma^2} \tag{8.10}$$

$$\tilde{f}_{c_i}(\Sigma^2, R) = \frac{1}{z_i(\Sigma^2, R)} \frac{m^2\Sigma^4 + \Sigma^2(2mR + \Sigma^2)\sigma^2 + (R^2 + \Sigma^2)\sigma^4}{(\Sigma^2 + \sigma^2)^2} - \tilde{f}_{a_i}(\Sigma^2, R)^2 \tag{8.11}$$

$$z_i(\Sigma^2, R) = \frac{\mathcal{N}\left(m|R, \Sigma^2\right)}{\mathcal{N}\left(m|R, \Sigma^2 + \sigma^2\right)} + 1 \tag{8.12}$$

where we have used that the set $\mathscr{S} = \mathbb{R}$ and thus just replaced the uniform distribution in the prior by 1. We could enforce positive values for the pixels, but it complicates the denoising functions expressions and seems to gives exactly the same final results anyway.

8.1.3 The learning equations

The hyperparameters $\mathbf{\Delta} := [\Delta_s, \Delta_d]$ need to be learned, where the noise associated to the μ^{th} measurement is $\Delta_\mu := \Delta_s \mathbb{I}(\mu \leq M) + \Delta_d \mathbb{I}(\mu > M)$. Δ_s is the true noise variance associated to the measure, Δ_d is the artificial noise variance associated to the dual variables which measure the level of relaxation of the constraints that define them: at $\Delta_d = 0$, the dual variables must exactly fulfill their definition, at finite value of Δ_d, these linear constraints are relaxed. We decide to keep ρ free being the control parameter of how smooth the final reconstructed picture is. Furthermore σ^2 can be learned but it appears empirically that fixing its value to $O(10^{-3})$ is sufficient and does not change the performances. For learning $\mathbf{\Delta}$, we will use the expectation maximization learning of sec. 4.3.8 optimizing the Bethe free energy. The learning of the noise is obtained by optimizing (4.202). For the moment we consider that there is a unique noise parameter Δ. The only part F_Δ dependent on it in (4.202) can be rewritten as:

$$F_\Delta = \frac{1}{2} \sum_{\mu}^{M+|E|} \left[\frac{(\tilde{y}_\mu - \sum_i^{N+|E|} \tilde{F}_{\mu i} a_i)^2}{\Delta} + \log\left(\Delta + \sum_i^{N+|E|} \tilde{F}_{\mu i}^2 v_i \right) \right] \tag{8.13}$$

where (a_i, v_i) are the posterior mean and variance of x_i. Then by optimizing it we obtain the fixed point equation for Δ:

$$\frac{\partial F_\Delta}{\partial \Delta} = 0 \tag{8.14}$$

$$\Leftrightarrow \sum_{\mu}^{M+|E|} \left[\frac{1}{\Delta^2} \left(\tilde{y}_\mu - \sum_i^{N+|E|} \tilde{F}_{\mu i} a_i \right)^2 - \left(\Delta + \sum_i^{N+|E|} \tilde{F}_{\mu i}^2 v_i \right)^{-1} \right] = 0 \tag{8.15}$$

A possible solution is to equate the two terms inside the sum for each component, which gives different solutions Δ_μ for each μ. We define the auxiliary functions:

$$\chi_\mu := \left(\tilde{y}_\mu - \sum_i^{N+|E|} \tilde{F}_{\mu i} a_i \right) \tag{8.16}$$

$$g_\mu := \frac{1}{2} \left(\chi_\mu^2 + \chi_\mu \sqrt{\chi_\mu^2 + 4\Theta_\mu} \right) \tag{8.17}$$

reminding the definition of $\Theta_\mu = \sum_i^{N+|E|} \tilde{F}_{\mu i}^2 v_i$ (4.190) at the fixed point. The solutions canceling each term inside (8.15) are given by the second order equations:

$$\Delta_\mu^2 - \chi_\mu^2 \Delta_\mu - \chi_\mu^2 \Theta_\mu = 0 \tag{8.18}$$

which exact solution is simply:

$$\Delta_\mu = g_\mu \tag{8.19}$$

Then we would average over these (the positive solutions are selected among the two possible ones) to get a single parameter Δ. But now we remember that we need to consider two different noise parameters $\{\Delta_s, \Delta_d\}$: the noise learning is not the same for the pixels and the

dual variables. To consider this, the average for the two different noise levels are performed over the proper sets of variables. We get the fixed point equations to which we add the time:

$$\Delta_s^{t+1} = \frac{1}{M}\sum_{\mu=1}^{M} g_\mu^t \tag{8.20}$$

$$\Delta_d^{t+1} = \frac{1}{|E|}\sum_{\mu=M+1}^{M+|E|} g_\mu^t \tag{8.21}$$

where g_μ^t is given by (8.17) with $\chi_\mu^t = \left(\tilde{y}_\mu - \sum_i^{N+|E|} \tilde{F}_{\mu i} a_i^t\right)$, $\Theta_\mu^t = \sum_i^{N+|E|} \tilde{F}_{\mu i}^2 v_i^t$ where (a_i^t, v_i^t) are the AMP posterior estimates at time t. In the implementation, these learnings are weakly damped.

It appears that this learning is essential for the perfomances of the algorithm. It is the main difference with the GrAMPA implementation [139]. Their prior model, refered as the SNIPE prior is different as well but as we will show in the experiments, it does not change the perfomances of AMP: the improvement in our implementation really comes from this noise learning.

8.1.4 Numerical experiments

We now present results of a serie of intensive numerical experiments. We have selected classical test 512×512 images, see Fig. 8.2. For each of them, we have compared the results obtained with the best TV-optimization-based algorithm, namely TV-AL3 [137] and our AMP implementation refered as DC-AMP for dual-constraints AMP. For each measurement rate, we have scanned different noise levels. For each point, 5 instances with different measurement operators and noise realizations have been reconstructed by each algorithm and the result is the averaged normalize mean square error NSNR (the MSE rescaled by the ℓ_2 norm of the picture) in decibels. For TV-AL3 that depends on two slack parameters (see [137] for the details), we have optimized them for *each point* (for every images and for all measurement rates). The same has been done for the unique free parameter of DC-AMP, the smoothness parameter ρ. So these results represent the optimal reconstruction performances of both algorithms. Furthermore, the SNIPE prior have also been implemented in our code for comparison with the Gauss-Bernoulli one and its free w parameter, which is the equivalent of ρ for the Gauss-Bernoulli prior have also been optimized for each point, see [139] for details on this prior.

Looking at all the results and tables, it is striking how close are the performances. There are virtually no differences. This suggests strongly that there exist some bound for the performances of such methods based on gradient optimization. The two priors in AMP appear to give almost the same results as it can be sees from the figures at $\alpha = 0.05$. This similarity remains true at higher measurement rates. When the GrAMPA implementation have been tested with the optimized value of w as well, it appeared to always give worst results by at least one or two decibels than what presented here. This must be due to our noise learning, not present in

their implementation. Furthermore, when we do not use this learning, we get similar results as GrAMPA with our implementation which is normal as the two algorithms are supposed to solve the same problem (8.5). In terms of running time, all algorithms are equivalent despite a small advantage for TV-AL3.

An important remark is that we have tried our implementation also with further interactions, including the 8 first-neighbors in the prior model instead of 4. It appears that it worsen the results by reconstructing too smooth solutions.

8.1.5 Concluding remarks

We have presented an AMP implementation for the reconstruction of natural images based on the co-sparse model, with a Gauss-Bernoulli prior on dual variables representing the differences between pixels. It appears that its optimal performances after optimization of its single free parameter are perfectly equivalent to the ones of the state-of-the-art TV-optimization algorithms which require more parameters to be tuned to get similar results. Furthermore, it seems that our results are weakly sensitive to the prior used with AMP and that Gauss-Bernoulli or its infinite variance limit give similar results. The two main points observed here are that i) the proposed algorithm get better performances that those of the similar GrAMPA algorithm due to a noise learning that exactly minimizes the Bethe free energy at each step and which is essential to reach the best performances and ii) it seems that we are reaching some bound on the perfomances of reconstruction algorithms for natural images in the compressive regime that are based on gradient-based models.

It would be of great interest to study further the fundamental reasons behind this limit. Furthermore, when trying to learn the ρ parameter, it seems to enter in conflict with the noise learning. This issue must be fixed to get a parameter-free algorithm with the best performances.

Figure 8.2 – Images used for the comparisons between TV-AL3 and our AMP implementation refered as DC-AMP for dual constraints AMP. Top, left to right: Lena, Baboon, Barbara. Bottom, left to right: Cameraman and Peppers

Lena Image

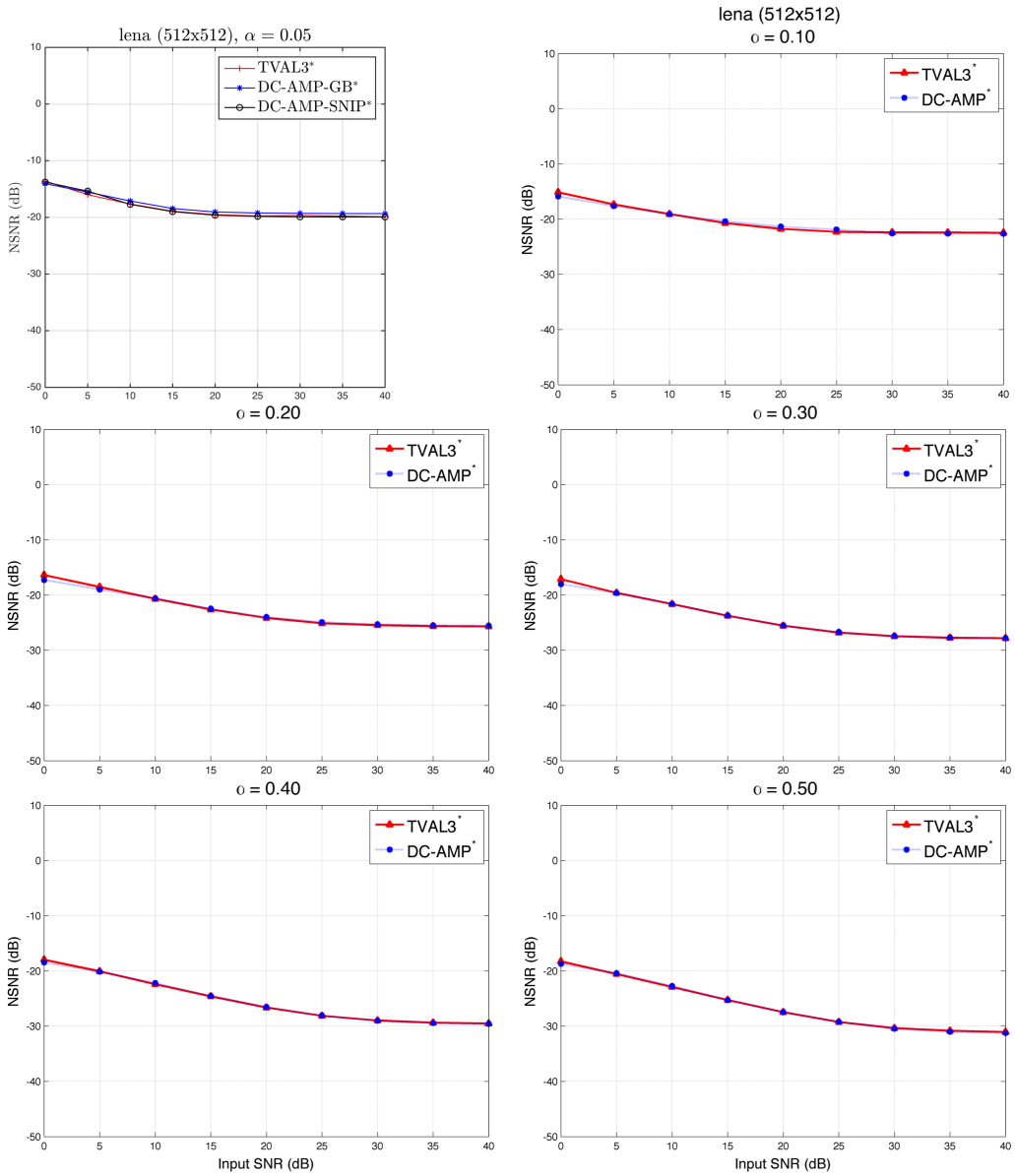

Figure 8.3 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Lena picture.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$	$\alpha = 0.40$	$\alpha = 0.50$
	ISNR = ∞ dB					
TVAL3 Optimal	**-19.74**	-22.51	**-25.71**	-27.80	-29.56	-31.13
DC-AMP Optimal	-19.33	**-22.65**	-25.59	**-27.87**	**-29.64**	**-31.32**
	ISNR = 40 dB					
TVAL3 Optimal	**-19.88**	-22.48	**-25.68**	-27.82	-29.51	-31.06
DC-AMP Optimal	-19.33	**-22.65**	-25.57	**-27.83**	**-29.58**	**-31.23**
	ISNR = 30 dB					
TVAL3 Optimal	**-19.66**	-22.40	**-25.46**	**-27.46**	-28.97	-30.34
DC-AMP Optimal	-19.29	**-22.58**	-25.34	-27.41	**-29.03**	**-30.46**
	ISNR = 25 dB					
TVAL3 Optimal	**-19.75**	**-22.31**	**-25.11**	**-26.81**	**-28.13**	**-29.25**
DC-AMP Optimal	-19.22	-21.90	-24.93	-26.71	-28.07	-29.24
	ISNR = 20 dB					
TVAL3 Optimal	**-19.53**	**-21.78**	**-24.15**	**-25.56**	**-26.63**	**-27.46**
DC-AMP Optimal	-19.05	-21.36	-24.01	-25.50	-26.54	-27.46
	ISNR = 15 dB					
TVAL3 Optimal	**-18.91**	**-20.71**	**-22.63**	**-23.77**	**-24.60**	-25.25
DC-AMP Optimal	-18.43	-20.40	-22.48	-23.70	-24.53	**-25.27**
	ISNR = 10 dB					
TVAL3 Optimal	**-17.64**	-19.08	**-20.69**	**-21.64**	**-22.40**	**-22.91**
DC-AMP Optimal	-17.10	**-19.09**	-20.61	-21.62	-22.20	-22.75
	ISNR = 5 dB					
TVAL3 Optimal	**-15.98**	-17.34	-18.54	-19.61	-20.05	**-20.55**
DC-AMP Optimal	-15.53	**-17.65**	**-19.02**	**-19.63**	**-20.13**	-20.42
	ISNR = 0 dB					
TVAL3 Optimal	-13.59	-15.14	-16.37	-17.13	-17.96	-18.26
DC-AMP Optimal	**-14.05**	**-15.90**	**-17.28**	**-18.03**	**-18.45**	**-18.72**

Table 8.1 – Table of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Lena picture.

barbara Image

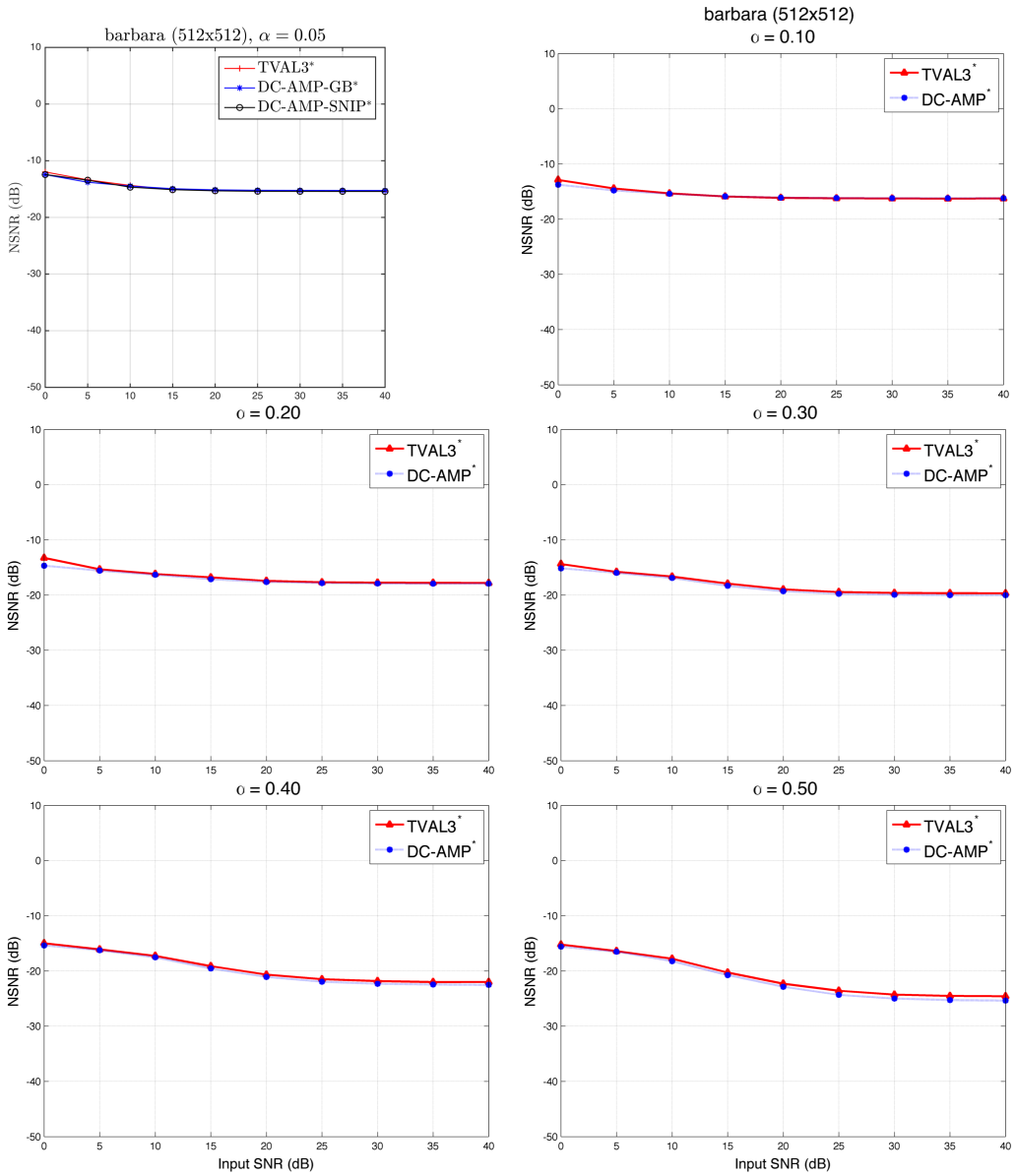

Figure 8.4 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Barbara picture.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$	$\alpha = 0.40$	$\alpha = 0.50$
	ISNR = ∞ dB					
TVAL3 Optimal	**-15.31**	**-16.27**	-17.80	-19.70	-22.00	-24.63
DC-AMP Optimal	-15.25	-16.22	**-17.94**	**-20.05**	**-22.54**	**-25.41**
	ISNR = 40 dB					
TVAL3 Optimal	**-15.38**	**-16.26**	-17.81	-19.69	-21.99	-24.59
DC-AMP Optimal	-15.25	-16.22	**-17.94**	**-20.03**	**-22.51**	**-25.37**
	ISNR = 30 dB					
TVAL3 Optimal	**-15.28**	**-16.26**	-17.77	-19.62	-21.83	-24.28
DC-AMP Optimal	-15.23	-16.21	**-17.91**	**-19.95**	**-22.33**	**-25.00**
	ISNR = 25 dB					
TVAL3 Optimal	**-15.35**	**-16.24**	-17.69	-19.45	-21.49	-23.58
DC-AMP Optimal	-15.22	-16.19	**-17.84**	**-19.78**	**-21.95**	**-24.32**
	ISNR = 20 dB					
TVAL3 Optimal	**-15.23**	**-16.16**	-17.44	-18.97	-20.67	-22.31
DC-AMP Optimal	-15.14	-16.11	**-17.63**	**-19.31**	**-21.07**	**-22.86**
	ISNR = 15 dB					
TVAL3 Optimal	**-15.09**	**-15.91**	-16.79	-17.93	-19.13	-20.29
DC-AMP Optimal	-14.94	-15.88	**-17.13**	**-18.33**	**-19.55**	**-20.75**
	ISNR = 10 dB					
TVAL3 Optimal	-14.36	-15.34	-16.20	-16.66	-17.27	-17.79
DC-AMP Optimal	**-14.46**	**-15.42**	**-16.32**	**-16.91**	**-17.53**	**-18.23**
	ISNR = 5 dB					
TVAL3 Optimal	-13.38	-14.45	-15.34	-15.81	-16.10	-16.40
DC-AMP Optimal	**-13.80**	**-14.80**	**-15.58**	**-15.98**	**-16.25**	**-16.52**
	ISNR = 0 dB					
TVAL3 Optimal	-11.98	-12.90	-13.28	-14.39	-14.98	-15.23
DC-AMP Optimal	**-12.45**	**-13.77**	**-14.69**	**-15.16**	**-15.37**	**-15.59**

Table 8.2 – Table of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Barbara picture.

baboon Image

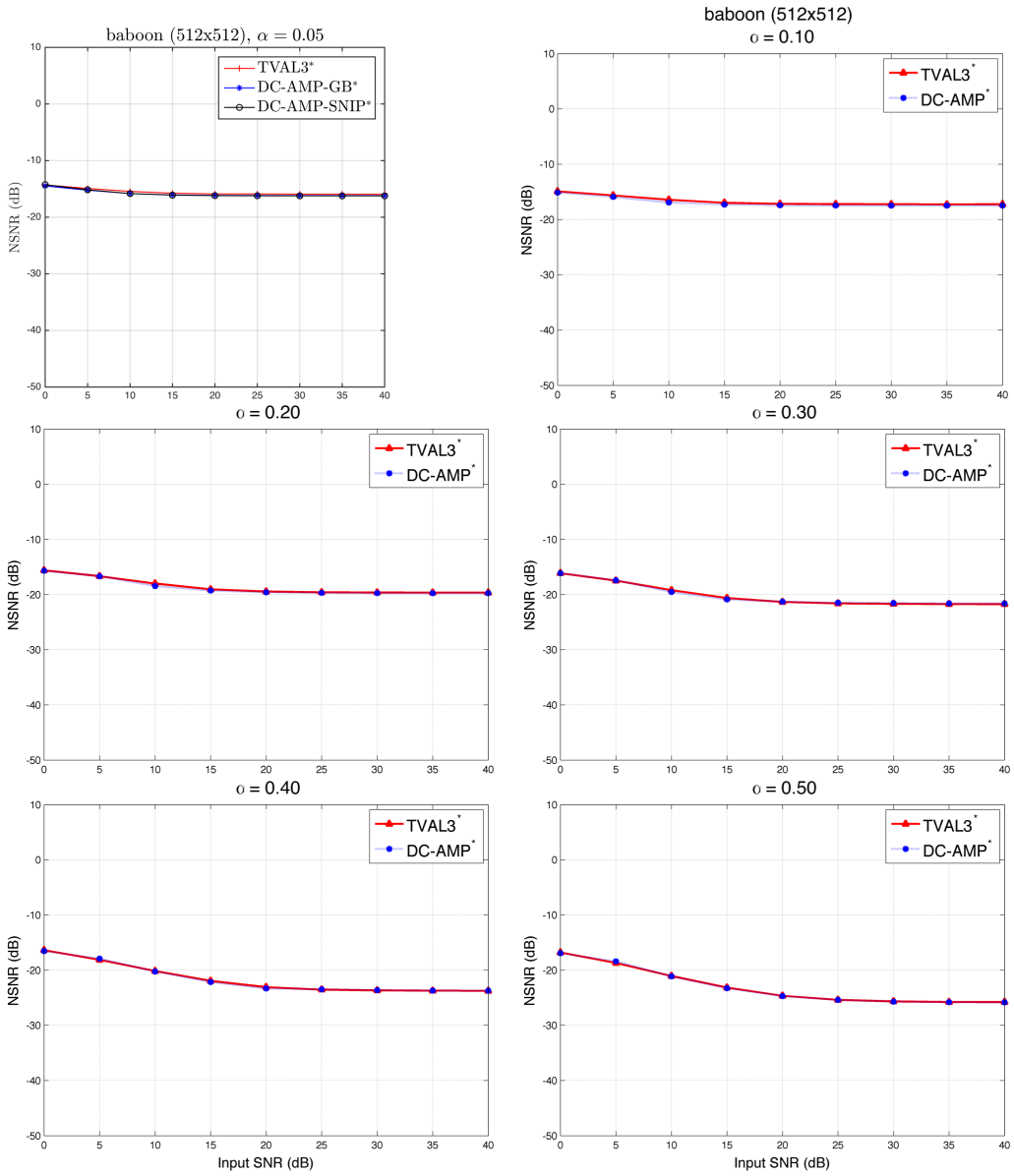

Figure 8.5 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Baboon picture.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$	$\alpha = 0.40$	$\alpha = 0.50$
	ISNR = ∞ dB					
TVAL3 Optimal	-15.94	-17.28	-19.65	**-21.74**	**-23.75**	-25.79
DC-AMP Optimal	**-16.18**	**-17.50**	**-19.76**	-21.63	-23.74	**-25.85**
	ISNR = 40 dB					
TVAL3 Optimal	-15.93	-17.26	-19.65	**-21.75**	**-23.74**	-25.78
DC-AMP Optimal	**-16.18**	**-17.50**	**-19.75**	-21.63	-23.71	**-25.84**
	ISNR = 30 dB					
TVAL3 Optimal	-15.92	-17.24	-19.64	**-21.69**	**-23.67**	-25.67
DC-AMP Optimal	**-16.18**	**-17.49**	**-19.74**	-21.57	-23.64	**-25.71**
	ISNR = 25 dB					
TVAL3 Optimal	-15.88	-17.24	-19.58	**-21.60**	**-23.53**	-25.38
DC-AMP Optimal	**-16.17**	**-17.48**	**-19.70**	-21.49	-23.48	**-25.39**
	ISNR = 20 dB					
TVAL3 Optimal	-15.88	-17.18	-19.45	**-21.34**	-23.08	-24.65
DC-AMP Optimal	**-16.15**	**-17.44**	**-19.60**	-21.30	**-23.32**	**-24.70**
	ISNR = 15 dB					
TVAL3 Optimal	-15.76	-16.99	-19.05	-20.63	-21.95	-23.15
DC-AMP Optimal	**-16.06**	**-17.30**	**-19.28**	**-20.89**	**-22.15**	**-23.27**
	ISNR = 10 dB					
TVAL3 Optimal	-15.44	-16.44	-18.00	-19.21	-20.15	-21.06
DC-AMP Optimal	**-15.82**	**-16.93**	**-18.46**	**-19.53**	**-20.24**	**-21.13**
	ISNR = 5 dB					
TVAL3 Optimal	-14.92	-15.66	-16.66	**-17.49**	**-18.14**	**-18.70**
DC-AMP Optimal	**-15.22**	**-15.93**	**-16.70**	-17.42	-17.93	-18.40
	ISNR = 0 dB					
TVAL3 Optimal	-14.29	-14.93	-15.61	-16.13	-16.39	-16.80
DC-AMP Optimal	**-14.48**	**-15.14**	**-15.69**	**-16.13**	**-16.53**	**-16.93**

Table 8.3 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Baboon picture.

***cameraman* Image**

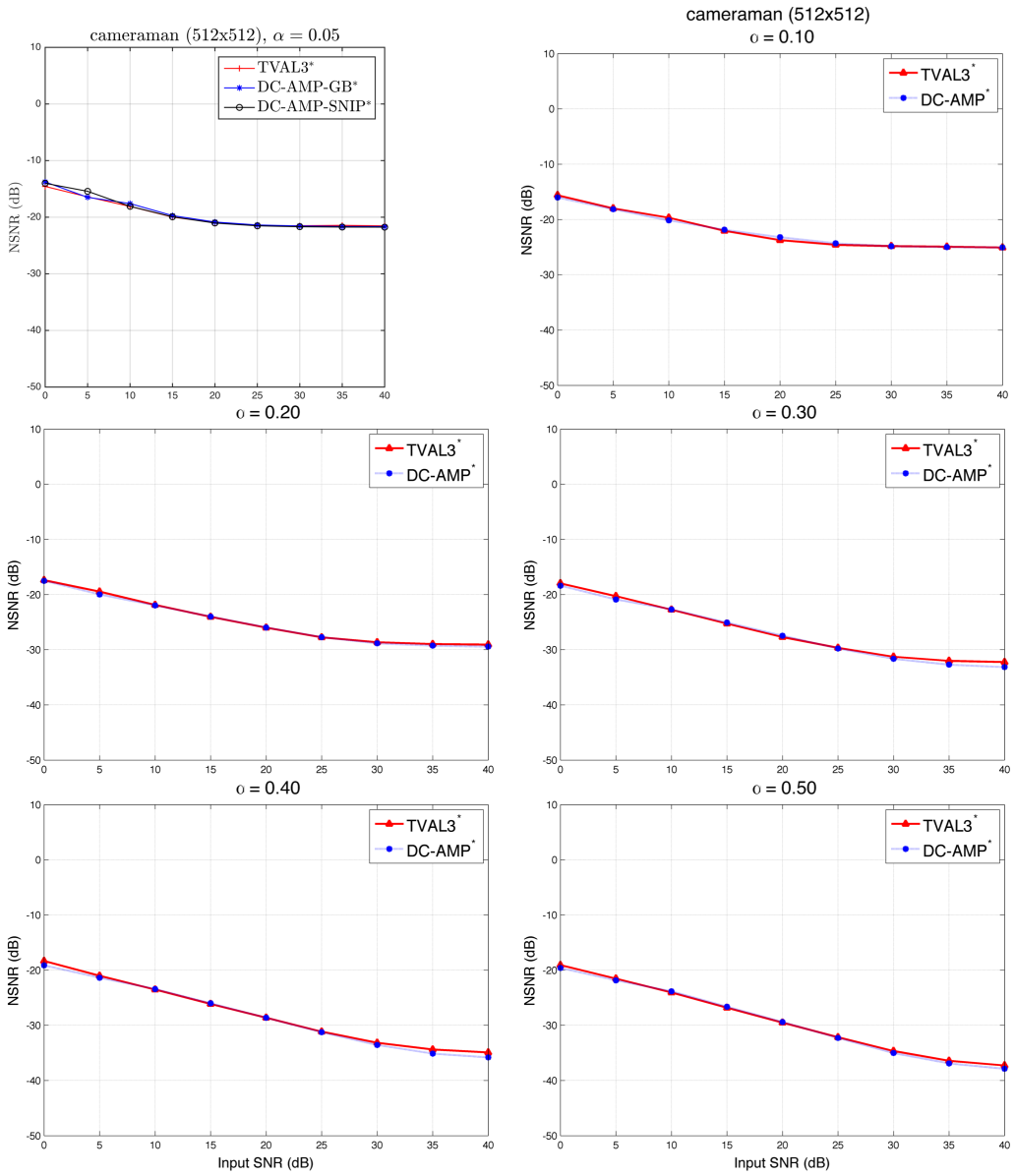

Figure 8.6 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Cameraman picture.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$	$\alpha = 0.40$	$\alpha = 0.50$
	ISNR = ∞ dB					
TVAL3 Optimal	-21.43	-25.02	-29.11	-32.42	-35.12	-37.67
DC-AMP Optimal	**-21.65**	**-25.10**	**-29.50**	**-33.35**	**-36.18**	**-38.50**
	ISNR = 40 dB					
TVAL3 Optimal	-21.52	**-25.09**	-29.09	-32.24	-34.91	-37.27
DC-AMP Optimal	**-21.64**	-25.07	**-29.44**	**-33.16**	**-35.81**	**-37.88**
	ISNR = 30 dB					
TVAL3 Optimal	-21.50	-24.83	-28.66	-31.28	-33.17	-34.67
DC-AMP Optimal	**-21.54**	**-24.85**	**-28.84**	**-31.65**	**-33.57**	**-35.03**
	ISNR = 25 dB					
TVAL3 Optimal	**-21.51**	**-24.60**	**-27.77**	-29.66	-31.15	-32.15
DC-AMP Optimal	-21.37	-24.33	-27.73	**-29.78**	**-31.25**	**-32.28**
	ISNR = 20 dB					
TVAL3 Optimal	**-20.99**	**-23.75**	**-26.02**	**-27.72**	-28.62	**-29.52**
DC-AMP Optimal	-20.85	-23.21	-25.98	-27.47	**-28.62**	-29.41
	ISNR = 15 dB					
TVAL3 Optimal	**-19.97**	**-22.06**	**-24.05**	**-25.28**	**-26.14**	**-26.82**
DC-AMP Optimal	-19.71	-21.87	-24.00	-25.08	-26.01	-26.64
	ISNR = 10 dB					
TVAL3 Optimal	**-18.09**	-19.69	-21.87	**-22.75**	**-23.53**	**-24.04**
DC-AMP Optimal	-17.59	**-20.17**	**-21.96**	-22.69	-23.43	-23.84
	ISNR = 5 dB					
TVAL3 Optimal	-16.40	-18.00	-19.47	-20.29	-21.01	-21.54
DC-AMP Optimal	**-16.48**	**-18.15**	**-19.99**	**-20.90**	**-21.38**	**-21.85**
	ISNR = 0 dB					
TVAL3 Optimal	**-14.56**	-15.62	-17.41	-17.98	-18.34	-19.08
DC-AMP Optimal	-13.77	**-16.03**	**-17.51**	**-18.40**	**-19.13**	**-19.61**

Table 8.4 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Cameraman picture.

peppers Image

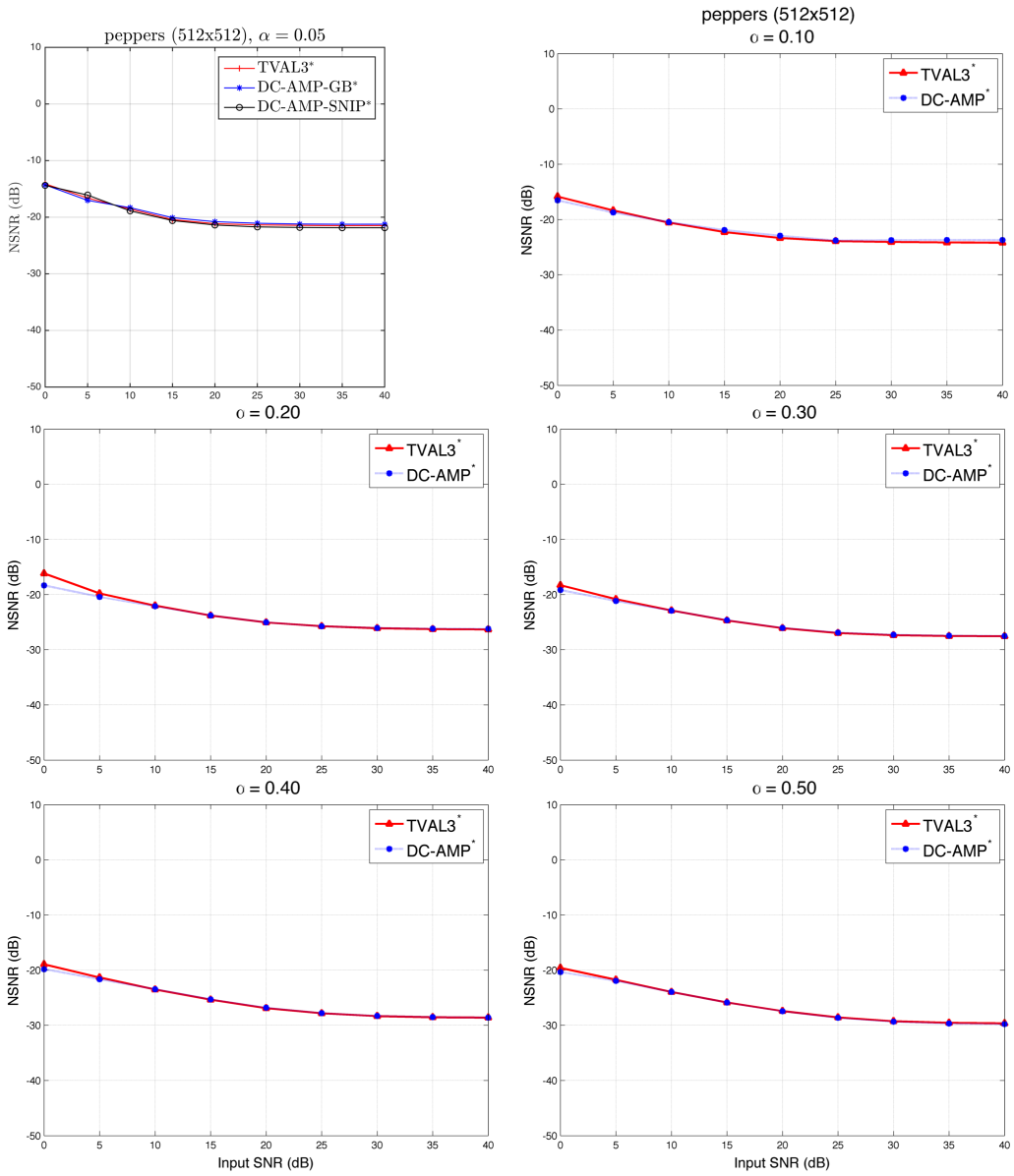

Figure 8.7 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Peppers picture.

	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.30$	$\alpha = 0.40$	$\alpha = 0.50$
	ISNR = ∞ dB					
TVAL3 Optimal	**-21.45**	**-24.22**	**-26.32**	**-27.59**	**-28.65**	-29.68
DC-AMP Optimal	-21.25	-23.74	-26.24	-27.55	-28.65	**-29.84**
	ISNR = 40 dB					
TVAL3 Optimal	**-21.48**	**-24.21**	**-26.34**	**-27.56**	-28.61	-29.65
DC-AMP Optimal	-21.23	-23.73	-26.22	-27.52	**-28.61**	**-29.79**
	ISNR = 30 dB					
TVAL3 Optimal	**-21.43**	**-24.08**	**-26.12**	**-27.38**	**-28.34**	-29.26
DC-AMP Optimal	-21.20	-23.76	-26.07	-27.30	-28.29	**-29.34**
	ISNR = 25 dB					
TVAL3 Optimal	**-21.35**	**-23.92**	-25.74	**-26.99**	**-27.82**	-28.56
DC-AMP Optimal	-21.07	-23.81	**-25.76**	-26.90	-27.79	**-28.64**
	ISNR = 20 dB					
TVAL3 Optimal	**-21.14**	**-23.37**	**-25.06**	**-26.10**	**-26.90**	-27.42
DC-AMP Optimal	-20.79	-22.94	-25.05	-26.05	-26.80	**-27.44**
	ISNR = 15 dB					
TVAL3 Optimal	**-20.42**	**-22.30**	**-23.81**	**-24.70**	**-25.36**	-25.86
DC-AMP Optimal	-20.09	-21.93	-23.79	-24.66	-25.30	**-25.88**
	ISNR = 10 dB					
TVAL3 Optimal	**-18.56**	**-20.59**	-22.01	-22.89	**-23.51**	**-23.96**
DC-AMP Optimal	-18.32	-20.48	**-22.11**	**-22.92**	-23.46	-23.95
	ISNR = 5 dB					
TVAL3 Optimal	-16.61	-18.37	-19.80	-20.84	-21.33	-21.74
DC-AMP Optimal	**-17.02**	**-18.74**	**-20.43**	**-21.17**	**-21.65**	**-21.94**
	ISNR = 0 dB					
TVAL3 Optimal	-14.11	-15.86	-16.19	-18.32	-18.94	-19.61
DC-AMP Optimal	**-14.29**	**-16.55**	**-18.35**	**-19.18**	**-19.84**	**-20.35**

Table 8.5 – Comparison of the final reconstruction results in NSNR as a function of the noise level in dB between TV-AL3 and DC-AMP for different measurement rates α for the Peppers picture.

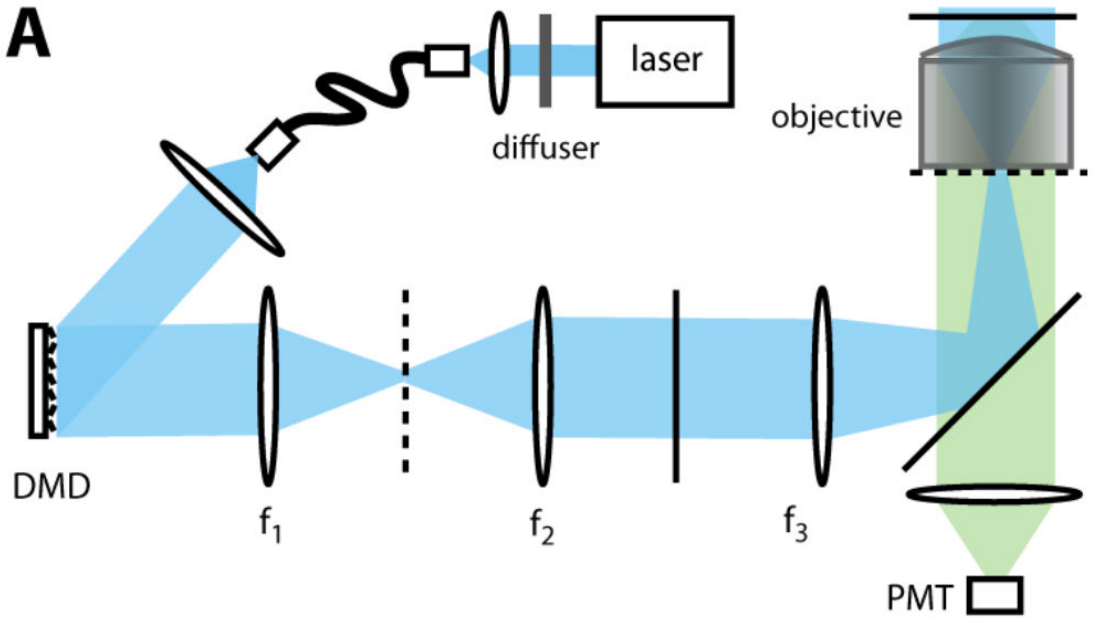

Figure 8.8 – Image taken from [141]. The experimental setup for compressive measurements of the fluorescent beads. A random selection of 2-d binary $\{0,1\}$ (illumination or not) Hadamard patterns generated by a laser beam and a digital micromirror device DMD are successively projected onto the sample of interest. For each pattern, the beads that are illuminated are excited. These then emit light by fluorescence. Each measure (one per Hadamard bi-dimensional pattern) corresponds to perform the sum of all the resulting photons by converging all this emitted light on a single point, which intensity is measured by the objective that outputs a single scalar value y_μ proportional to the number of photons received.

8.2 Image reconstruction in compressive fluorescence microscopy

We now present an application of approximate message-passing inference to image reconstruction in fluoresence microscopy. The present work is part of an ongoing collaboration with Vincent Studer, Makhlad Chahid and Maxime Dahan who performed the experimental part of [141]. All the data analysed in sec. 8.2.7 has been generated by them.

8.2.1 Introduction

For the next, we call a *measure* (or measurement) the process of measuring the overall light intensity emitted by the beads after the excitation by one single 2-d Hadamard pattern thanks to the setup Fig. 8.8, and an *acquisition* the full process of getting M different measures: one can try to reconstruct the image of the beads from one vector of M measurements $\mathbf{y}$ obtained thanks to one acquisition.

In the present problem, the aim is to locate fluorescent point-like beads on a plane (thus the signal is directly sparse in the pixel domain) thanks to compressive measurements i.e. from

$M < N$ measurements, where N is the number of pixels of the image to reconstruct.

Fluorescence microscopy has a great potential, especially in biological applications. Unfortunatly measurements may be costly in time making a single acquisition very long and thus compressed sensing is highly relevant here. As the image of the beads is sparse as seen from Fig. 8.10, compressed sensing theoretically allows to reconstruct it from far fewer measurements than usual methods and thus to drastically speed up the acquisition. If acquisitions were fast enough thanks to compressed sensing, one could think of observing the dynamical behavior of small objects such as proteins and cells. This would require that their typical evolution time scale would be smaller than the acquisition time. Indeed, in order to reconstruct an image from an acquisition, it is essential that the measured system does not evolve from one measurement to the next or they would be incoherent and reconstruction impossible.

For example, even at very low temperatures (not even speaking of biologically relevant temperatures), a protein tertiary structure may evolve due to the thermal noise, and if one aim at oberving it in a particular configuration, fast acquisitions processes are essential or it will have time enough to relax to a new configuration during the acquisition.

Compressed sensing could also allow to increase the spatial resolution of the images from a computational point of view. As the number of pixels N fixes the maximum spatial resolution for the location of the beads and the classical sensing procedures require a number of measures that scale with N, an increase in resolution cost many more measures whereas with compressed sensing, only $O(\rho N)$ measurements are required, and thus the resolution for very sparse images with $\rho \ll 1$ can be improved at low cost.

8.2.2 Experimental setup and algorithmic setting

The signal processing problem treated in this chapter is the reconstruction of fluorescent beads randomly placed on a plane. The experimental setup of Fig. 8.8 is used for obtaining the compressive measurements of the beads: a random selection of M 2-d binary $\{0,1\}$ (illuminated or not) Hadamard patterns successively generated by a laser beam and a digital micromirror device are projected onto the plane where the beads are. For each pattern, the beads that are illuminated absorbe this light and then emit it back by fluorescence. The sum of all the resulting photons is performed by converging all this emitted light on a single point measured by an accurate intensity detector that output a scalar y_μ for each measurement, $\mu \in \{1,\dots,M\}$. We refer to [141] for a detailed description of the experimental setup. A random 2-d Hadamard pattern is represented at the center of Fig. 8.9 and its projection on a biological sample is on the right part.

In the present setting, we assume that the linear model describing this experimental setup is given by (3.18). Indeed the 2-d Hadamard transform H_{2d} of a matrix $\mathbf{x}$ of size $\sqrt{N} \times \sqrt{N}$ is

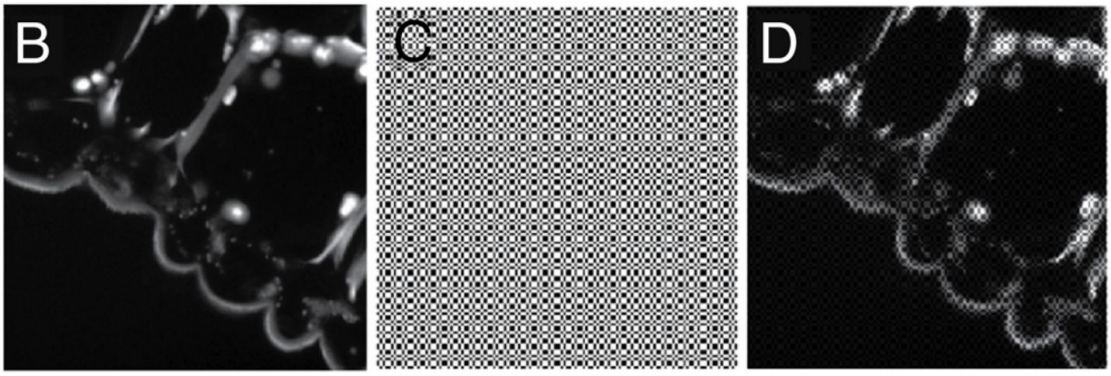

Figure 8.9 – Image taken from [141]. The left part is a biological sample, the center one is a typical randomly selected 2-d Hadamard pattern (or mode) used for the compressive measurements of the fluorescent beads and the right one is the same pattern projected onto the biological sample: the lighter points are the ones that are excited.

defined in function of the Hadamard matrix H as:

$$H_{2d}(\mathbf{x})_{ij} := \left(H(H\mathbf{x})^{\mathsf{T}}\right)_{ij} = \sum_{k,u}^{\sqrt{N},\sqrt{N}} H_{ik}H_{ju}x_{uk} \tag{8.22}$$

This operator can be easily written as a single matrix by rasterizing each 2-d mode made of $\sqrt{N}$ lines of size $\sqrt{N}$ (see Fig. 8.9) into one line of size N of $\mathbf{F}$. Doing this for all the M modes and considering the properly rasterized signal vector $\mathbf{x}$ which is now of size N, we get back the system (3.18).

But for the algorithm implementation, we want to use the fast Hadamard transform and thus the form (8.22) of the operator H_{2d} instead of its matrix representation as we are working with large data sets. Thus the operator $H_{2d}(\mathbf{x})$ which outputs the same vector as the direct matrix product $\mathbf{Fx}$ (which thus can be directly used in Fig. 5.5 without having to change anything) but that is constructed using the fast Hadamard transform is implemented by the following

pseudo-code:

$$
\begin{aligned}
&H_{2d}\Big(\textbf{input : } \mathbf{x} \text{ of size } N, \text{ sets of 1-d Hadamard modes indices } (I_1, I_2)\Big)\\
&\quad 1) \text{ "Derasterize" } \mathbf{x} \text{ to make it of size } \sqrt{N}\times\sqrt{N}\\
&\quad 2) \text{ Apply the 2-d fast Hadamard transform : } \mathbf{p} = FT(FT(\mathbf{x})^{\mathsf{T}})\\
&\quad 3) \text{ Rasterize } \mathbf{p} \text{ to make it of size } N\\
&\quad 4) \text{ Rescale the result to take into account the } \{0,1\} \text{ nature of the measurement}\\
&\quad \text{operator while the } FT \text{ operator is } \{-1,1\} : \mathbf{z} = \frac{1}{2}\left(\mathbf{p} + \sum_i^N x_i\right)\\
&\quad 5) \text{ Select the proper } M \text{ firsts modes : } \mathbf{u} = \left[\mathbf{z}\left(\sqrt{N}(I_1(i)-1) + I_2(i)\right)\right]_i^M\\
&\textbf{output : } \mathbf{u} \text{ of size } M
\end{aligned}
\tag{8.23}
$$

where FT is the usual fast Hadamard transform. The required backward operator construction is related to what has been presented in sec. 7.1.3 and is implemented here as:

$$
\begin{aligned}
&H_{2d}^{\mathsf{T}}\Big(\textbf{input : } \mathbf{y} \text{ of size } M, \text{ sets of 1-d Hadamard mode indices } (I_1, I_2)\Big)\\
&\quad 1) \text{ Create a vector with the } M \text{ firsts 1-d mode indices : } \mathbf{v} = \left[\sqrt{N}(I_1(i)-1) + I_2(i)\right]_i^M\\
&\quad 2) \text{ Define } \mathbf{f} \text{ of size } N \text{ such that : } f_{v_i} = y_i \ \forall i \in \{1,\dots,M\}, f_k = 0 \ \forall k \notin \mathbf{v}\\
&\quad 3) \text{ Derasterize } \mathbf{f} \text{ to make it of size } \sqrt{N}\times\sqrt{N}\\
&\quad 4) \text{ Apply the 2-d fast Hadamard transform : } \mathbf{p} = FT(FT(\mathbf{f})^{\mathsf{T}})\\
&\quad 5) \text{ Rasterize } \mathbf{p} \text{ to make it of size } N\\
&\quad 6) \text{Rescale the result to take into account the } \{0,1\} \text{ nature of the measurement}\\
&\quad \text{operator while the } FT \text{ operator is } \{-1,1\} : \mathbf{u} = \frac{1}{2}\left(\mathbf{p} + \sum_\mu^M y_\mu\right)\\
&\textbf{output : } \mathbf{u} \text{ of size } N
\end{aligned}
\tag{8.24}
$$

The modes-indices sets (I_1, I_2) are selected by the experimentalists and define the scrambled sub-sampled 2-d Hadamard transform used for acquisition (they contain respectively the M $\{i\}$ and $\{j\}$ indices selected when using (8.22), with $i, j \in \{1,\dots,\sqrt{N}\}$). 8192 modes were selected for the data used in sec. 8.2.7. Only a sub-set of them can be selected for the reconstruction tests, changing M in the previous operators. The fast implementation of the homogeneous approximate message-passing algorithm Fig. 5.5 (with $L = N, L_c = L_r = 1, B = 1$) is thus defined replacing the forward operator O_μ (5.117) by the fast operator H_{2d} and the backward one O_i (5.119) by H_{2d}^{T}. The two others operators (5.116), (5.118) are also defined repectively by H_{2d} and H_{2d}^{T} as 0 or 1 are invariant by the square operation. Let us now discuss a proposal for the denoisers that appear to have very good performances in the present problem.

8.2.3 A proposal of denoisers for the reconstruction of point-like objects measured by compressive fluorescence microscopy

In order to perform inference of the beads locations thanks to the approximate message-passing algorithm Fig. 4.6, we propose here an exponential prior with approximate Gaussian sparsity for designing the denoisers following the procedure given in sec. 4.3.6. The exponential part approximates the discrete Poisson distribution associated with the number of informative photons emitted by the actual beads. This law is typical of sources which light emission is due to desexcitation processes, such as in fluorescence. The full prior that we assume factorizable over the pixels is thus given by:

$$P_0(\mathbf{x}|\boldsymbol{\theta}) = \prod_i^N P_0(x_i|\boldsymbol{\theta}) = \prod_i^N \left[\rho\lambda e^{-\lambda x_i}\mathbb{I}(x_i > 0) + \mathcal{N}\left(x_i|m,\sigma^2\right)\right] \tag{8.25}$$

where $\boldsymbol{\theta} := [\lambda, \rho, m, \sigma^2]$. ρ is the density of pixels of the picture on which beads are standing and $\mathbb{I}(x_i > 0)$ enforces the pixels to have positive values. Of course this model is an approximation of the true signal generating process, but it appears empirically to reach very good performances. More complex prior models including correlations between pixels could be considered but this cannot be done with AMP, at least in its canonical form Fig. 4.6 which requires that the prior is factorizable over some subsets of signal components. Nevertheless, we will see in sec. 8.2.6 how to include "a posteriori" some effective interaction between closeby pixels.

In the present context, we consider the measurement noise variance $\Delta \to 0$ as the background photon noise is already included into the Gaussian part of the prior. We could think also of using a strictly sparse prior replacing the Gaussian by a Dirac distribution and letting the noise variance have a finite value instead, but it appears empirically that it gives worst results than this approximate sparsity prior. This may come from the fact that the learning rules of the Gaussian parameters (m, σ^2) are different that of the noise variance Δ and there are two instead of one. This remark is actually quite general, and we observed empirically in many situations that approximate sparsity without measurement noise gets better results than a sparse prior with noise.

8.2.4 Optimal Bayesian decision for the beads locations

A great advantage of the Bayesian framework with respect to convex optimization procedures is that it allows to directly estimate the probability that a pixel supports a bead or not, i.e. if this pixel is informative or belongs to the background noise. The posterior "noise" probability $P(x_i \in \mathcal{N})$ for a pixel x_i to be pure noise (denoted by $\mathcal{N}$ in reference to the Gaussian part for the noise in the prior) is proportional to the prior probability of belonging to the background

noise re-weighted by the AMP Gaussian field:

$$P(x_i \in \mathcal{N}|\boldsymbol{\theta}, \Sigma_i^2, R_i) = \frac{1}{z(\Sigma_i^2, R_i|\boldsymbol{\theta})} \int dx_i \mathcal{N}\left(x_i|m, \sigma^2\right) \mathcal{N}\left(x_i|R_i, \Sigma_i^2\right) \tag{8.26}$$

$$z(\Sigma_i^2, R_i|\boldsymbol{\theta}) = \int dx_i \mathcal{N}\left(x_i|R_i, \Sigma^2\right) P_0(x_i|\boldsymbol{\theta}) \tag{8.27}$$

where $z(\Sigma_i^2, R_i|\boldsymbol{\theta})$ is the posterior partition function of pixel i, P_0 is given by (8.25) and (Σ_i^2, R_i) are the usual moments controlling the Gaussian AMP field summarizing the likelihood constraints on x_i. The (Σ_i^2, R_i) values are iteratively computed by AMP Fig. 4.6. This expression is analytical and using (8.25), (8.26) it becomes:

$$P(x_i \in \mathcal{N}|\boldsymbol{\theta}, \Sigma_i^2, R_i) = \left[1 + \rho\lambda\sqrt{\frac{\pi(\Sigma_i^2 + \sigma^2)}{2}} e^{\frac{\lambda}{2}(\lambda\Sigma_i^2 - 2R_i) + \frac{(m-R_i)^2}{2(\Sigma_i^2+\sigma^2)}} \operatorname{erfc}\left(\frac{\lambda\Sigma_i^2 - R_i}{\sqrt{2\Sigma_i^2}}\right)\right]^{-1} \tag{8.28}$$

After convergence of the algorithm, in order to obtain the final estimate $\hat{x}_i$ for each pixel, we thus cancel all the final AMP posterior pixels estimates a_i which final posterior noise probability is more than 0.5. Doing this we keep only the supposed informative pixels from the AMP point of view:

$$\hat{x}_i = a_i^t \, \mathbb{I}\left(P(x_i \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_i^t)^2, R_i^t) < 0.5\right) \tag{8.29}$$

where t is the final time step. This kind of decisions cannot be taken with ℓ_1-minimization based solvers as they are not probabilistic algorithms, and arbitrary thresholding functions *must* be applied at the end or the results are very poor in this kind of highly noisy problems. In the experiments of sec. 8.2.7, the thresholding function applied to the ℓ_1-minimization based solvers final estimates is such that we keep only the pixels that have an amplitude approximately 4 times higher than the mean of the recovered overall picture, we cancel the other ones. This value of 4 has been selected *empirically* to obtain the best possible match between the reconstructed and original pictures in the high measurement rate regime ($M = 8192$ measurements in the results of sec. 8.2.7). Despite being probably suboptimal for other measurement rates, this thresholding function appears to output results close to the best ones at any rate, i.e. that are obtained when the optimization of this thresholding function is performed for every measurement rate and ℓ_1-minimization solver independently.

8.2.5 The learning equations

In order to find the optimal values of the free parameters of the prior (8.25), we use the expectation maximization strategy discussed in sec. 4.3.8. In the present case, all the quantities can be simply obtained directly from the posterior estimates and noise probability. Appropriate

recursions that are very stable numerically and have simple interpretations are given by:

$$\rho^{t+1} = \frac{1}{N}\sum_{i}^{N} \mathbb{I}\left(P(x_i \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_i^t)^2, R_i^t) < 0.5\right) \tag{8.30}$$

$$\lambda^{t+1} = \left[\frac{1}{|\mathbf{a}_{supp.}^t|}\sum_{i}^{|\mathbf{a}_{supp.}^t|} a_{i,supp.}^t\right]^{-1} = \frac{1}{<\mathbf{a}_{supp.}^t>} \tag{8.31}$$

$$m^{t+1} = \frac{1}{|\mathbf{a}_{noise}^t|}\sum_{i}^{|\mathbf{a}_{noise}^t|} a_{i,noise}^t = <\mathbf{a}_{noise}^t> \tag{8.32}$$

where $\mathbf{a}_{supp.}^t := \left[a_i^t : P(x_i \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_i^t)^2, R_i^t) < 0.5\right]$ are the posterior estimates of the estimated support pixels of the beads at time t, $\mathbf{a}_{noise}^t := \left[a_i^t : P(x_i \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_i^t)^2, R_i^t) > 0.5)\right]$ are the posterior estimates of the noise pixels (that are not in the support) at time t. The variance of the Gaussian could be learned as well in the same way but it appears empirically that fixing its value is more efficient. All these equalities are just coming from the very definitions of the different quantities. For example, the parameter λ in the exponential distribution must be equal to the inverse of the mean of this distribution, and we naturally take only into account the values of the pixels that are considered in the support as the exponential is here to model the beads.

8.2.6 Improvement using small first neighbor mean field interactions between pixels

It appears empirically that using a small first neighbor mean field interaction between pixels improve the results and allows AMP to recover perfectly the beads locations at smaller measurement rates. The trick is empirically done by adding to the moment R_i^{t+1} controling the AMP field felt by the pixel x_i at time $t+1$ a perturbation ϵh_i^{t+1} where ϵ is a small $\in O(10^{-1})$ auxiliary parameter and h_i^{t+1} is the mean field that takes into account the (up to) 4 x_i's neighbors states. It is defined as an extension of the so-called bilateral denoiser in statistical image processing and is defined as:

$$h_i^{t+1} = \frac{\sum_{j\in\partial i} a_j^t\, w_{ij}^t\, \mathbb{I}\left(P(x_j \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_j^t)^2, R_j^t) < 0.5\right)}{\sum_{j\in\partial i} w_{ij}^t\, \mathbb{I}\left(P(x_j \in \mathcal{N}|\boldsymbol{\theta}^t, (\Sigma_j^t)^2, R_j^t) < 0.5\right)} \tag{8.33}$$

where the weight given to the neighbor pixel x_j of pixel x_i is a Gaussian proportional to their posterior estimates difference, i.e. it gives higher weight to similar pixels:

$$w_{ij}^t = \mathcal{N}\left(a_i^t | a_j^t, \sigma_w^2\right) \tag{8.34}$$

This field thus weakly shifts the AMP field of the i^{th} pixel to higher values when its neighbors are considered being part of the support: $R_i^{t+1} = R_i^{t+1} + \epsilon h_i^{t+1}$. This mean field thus mimics in some sense the behavior of the algorithm discussed in sec. 8.1 but without changing the prior

that remains fully factorizable over the pixels. It appears empirically that the improvement thanks to this strategy is only very weakly dependent on the σ_w^2 parameter and we take a uniform weight $\sigma_w^2 \to \infty$ for the results presented in the next section.

8.2.7 Reconstruction results on experimental data

We now present some results which compare the reconstruction performances of AMP and of two state-of-the-art ℓ_1-minimization based solvers (NESTA [52] and fast iterative hard thresholding [80]) and where the Bayesian optimal and heuristic thresholding functions discussed in sec. 8.2.4 are applied to the final reconstructions made with AMP and the ℓ_1-minimization based solvers respectively. Furthermore, each reconstructed support pixel is "highlighted" a posteriori by giving a positive value to its closest neighbors as well. The data used here has been obtained by the authors of [141] using the experimental setup described in Fig. 8.8.

A first remark is that independently of the algorithm used or the measurement rate α, there are always artefacts appearing on the up-left corner (two beads are always missing in the reconstruction) and for the positions of some beads as well. This must come from the experimental data set used here rather than the algorithms, as even at the higher rate α, there remain these errors. So it is considered that these systematic errors are actually not errors for the algorithmic reconstructions.

The results Fig. 8.10-Fig. 8.15 all show a clear advantage for AMP: its reconstruction and location of the beads is perfect (up to these systematic data-dependent errors) until $M = 512$ whereas comparably good results (yet not perfect, whereas the AMP results are) are obtained with NESTA only for $M \geq 4096$ or with FastIHT at $M \geq 8192$. So the gain with AMP is substantial, and the location is way more accurate. Furthermore, the speed of convergence of AMP is always 2 to 10 times faster than the convex optimization solvers used here. For $M < 512$ approximatively, the AMP performances start to worsen continuously, and actual beads disappear while new "fake" ones start to appear as seen on the last figure Fig. 8.15.

An important remark is that when the previously discussed "TV-like" AMP algorithm of sec. 8.1 have been tried on the data, it appeared that the reconstruction performances were not comparable with the AMP implementation presented here. This is due to the point-like nature of the beads: the gradient-minimizing prior of sec. 8.1.1 tends to smoothen too much the background and makes totally disappear the beads for low measurement rates, whereas the present specifically designed prior (8.25) does consider the precense of such ponctual objects.

8.2.8 Concluding remarks and open questions

We have studied how the AMP algorithm can be used for reconstruction of sparse images in the pixel domain measured by fluoresence microscopy in the compressive regime. This study is a proof-of-concept, which naturally extended the work of [141] where convex optimization

is used for the reconstruction. A natural continuation is to try the algorithm on real biological samples, where the beads are put inside membranes of cells for example as in [141]. Furthemore, in the data set used for the present results, the beads were strongly excited and thus emitted a lot of photons. The question of wether the reconstruction by AMP is robust to a net lowering of the beads emission intensity is of great interest as stronger excitations means a longer exposition time of the sample to the light field, and thus an overall longer acquisition time. Another natural idea would be to try spatial coupling combined with Hadamard patterns as in chap. 7. But as the noise is high (typically a relative intensity of $O(10^{-2}/10^{-3})$ with respect to the beads intensity in the present experiments), it is probable that there is no first order transition and thus that this strategy does not improve the performances as discussed in chap. 6. Nevertheless, the results are very encouraging and AMP seems to be a good option in this context.

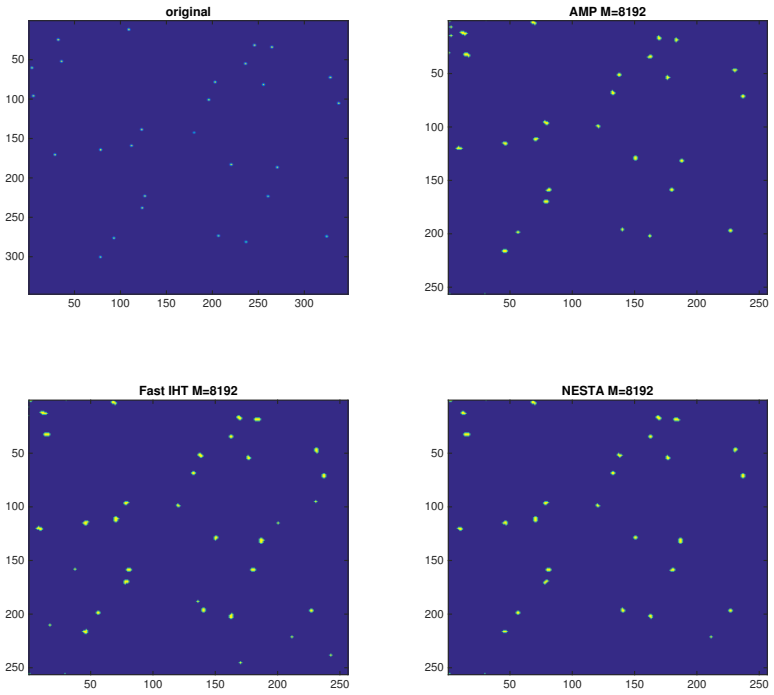

Figure 8.10 – Comparison of the reconstruction results of the 3 algorithms used here: AMP, NESTA and Fast Iterative Hard Thresholding with the original picture. The number of measurements is here $M = 8192$, $\alpha = 0.125$.

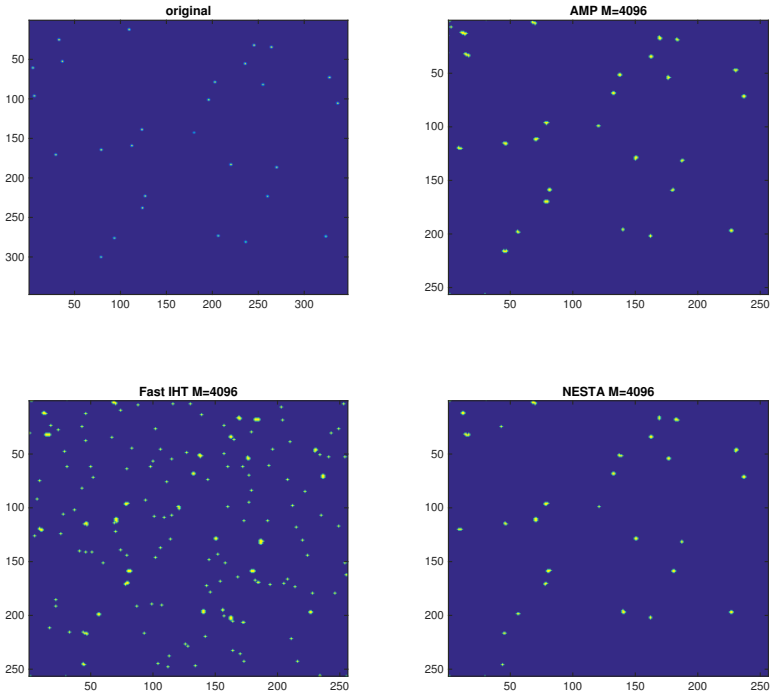

Figure 8.11 – Same as Fig. 8.10 with $M = 4096$, $\alpha = 0.0625$.

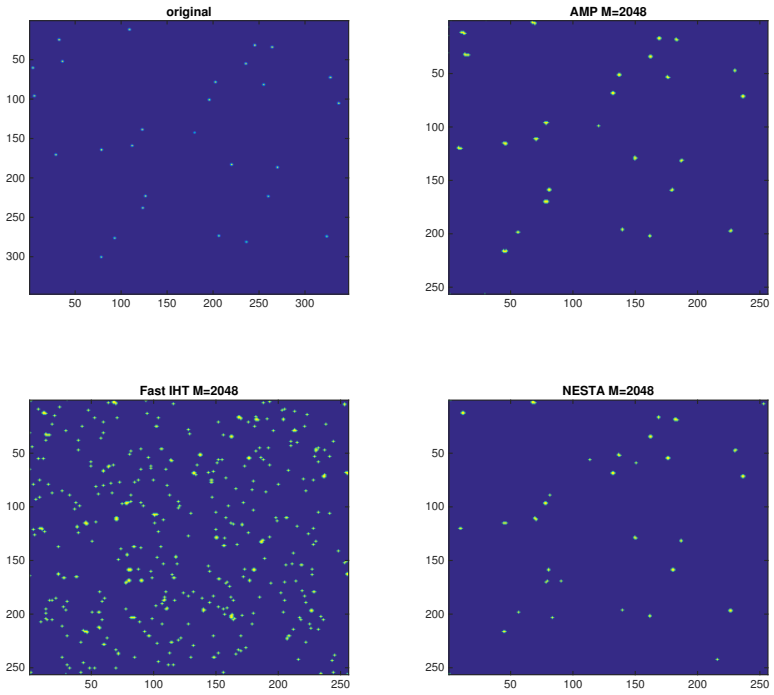

Figure 8.12 – Same as Fig. 8.10 with $M = 2048$, $\alpha = 0.031$.

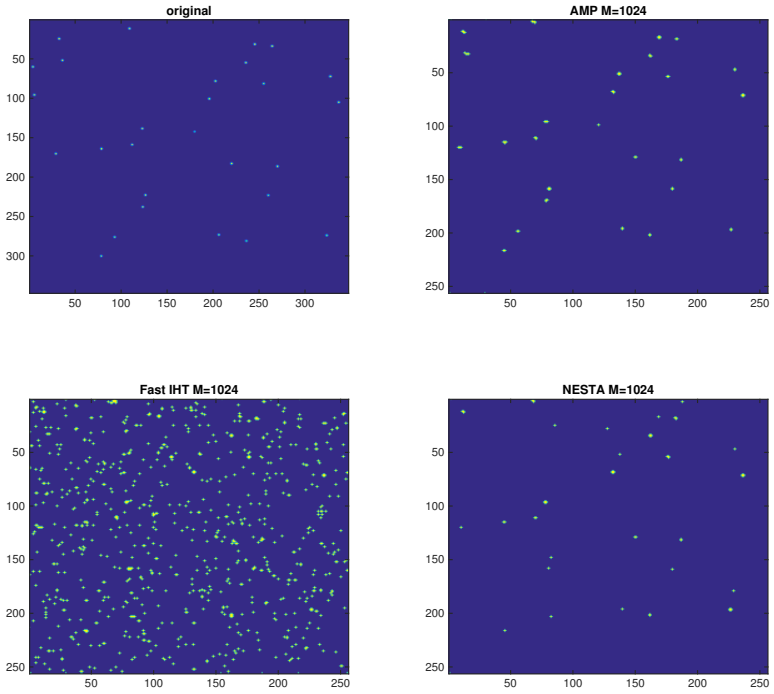

Figure 8.13 – Same as Fig. 8.10 with $M = 1024$, $\alpha = 0.016$.

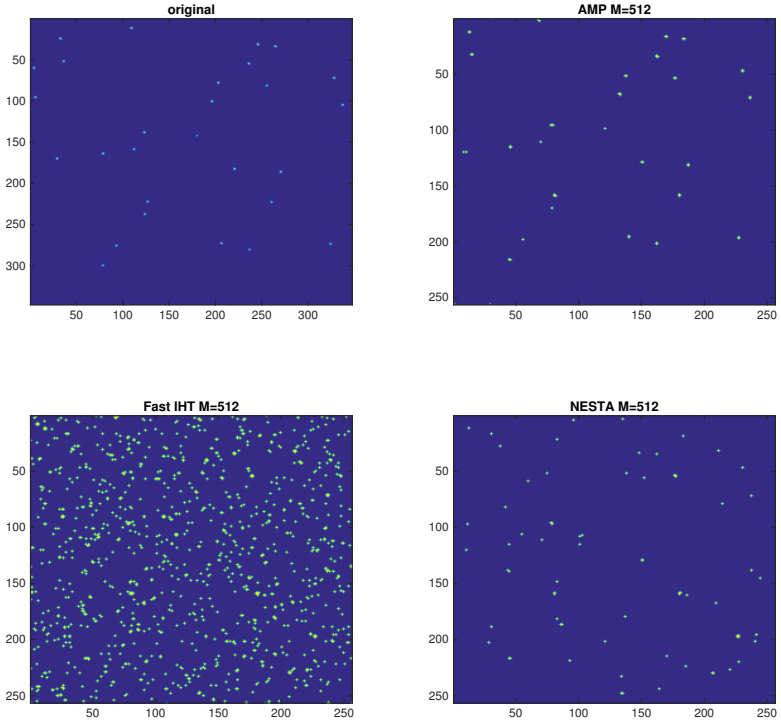

Figure 8.14 – Same as Fig. 8.10 with $M = 512$, $\alpha = 0.0078$. This α is the limit of "perfect" beads location using the AMP algorithm.

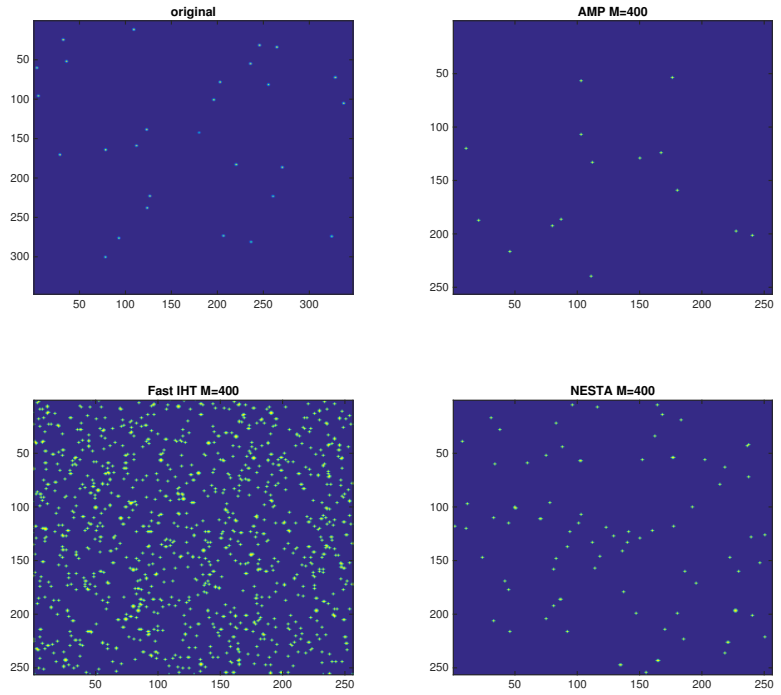

Figure 8.15 – Same as Fig. 8.10 with $M = 400$, $\alpha = 0.0061$. We observe a continuous worsening of the AMP results.

Coding theory Part IV

9 Approximate message-passing decoder and capacity-achieving sparse superposition codes

We study the approximate message-passing decoder for sparse superposition coding over the additive white Gaussian noise channel. While this coding scheme asymptotically reaches the Shannon capacity, we show that the AMP iterative decoder is limited by the BP phase transition, similar to what happens in low density parity check LDPC codes. We present and study two solutions to this problem, that both allow to reach the Shannon capacity: i) a non constant power allocation and ii) the use of spatially-coupled codes. We also present extensive simulations that suggest that spatial coupling is more robust and allows for better correction at finite code lengths. Finally, we show empirically that the use of a fast Hadamard-based operator allows for an efficient reconstruction, both in terms of computational time and memory allocation, and the ability to deal with very large signals.

9.1 Introduction

The error correction scheme called sparse superposition codes has originally been introduced and studied in [142–144] by Barron and Joseph who proved the scheme to be capacity achieving over the additive white Gaussian noise AWGN channel under maximum-a-posteriori MAP decoding. In [142–144], an iterative decoder called *adaptive successive decoder* was presented, which was later improved in [145, 146] by soft thresholding methods. The idea is to decode a sparse vector with a special block structure over the AWGN channel, represented in Fig. 9.1. With these decoders together with the use of power allocation, the scheme was proved to be capacity achieving in a proper limit. However, finite blocklength performances were far from ideal. In fact, it seemed that the asymptotic results could be reproduced at any reasonable finite lengths.

We propose instead an approximate message-passing decoder for sparse superposition codes. We will show that this decoder have much better performances. In fact it allows better decoding than the iterative successive decoder at any reasonable finite length, and this even without power allocation. We present two modifications of sparse superposition codes that allow

AMP to be asymptotically capacity achieving as well, while retaining good finite block length properties. The first one is the addition of power allocation to sparse superposition codes as done for the iterative successive decoder, and the second one, which is specific to the message-passing decoder, is the use of spatial coupling which appears to be even more promising.

We also present extensive numerical simulations and a study of a practical scheme with Hadamard operators. The overall scheme is computationally efficient and allows to practically reach near-to-capacity transmission rates with low error floors.

9.1.1 Related works

The phenomenology of these codes under AMP decoding, in particular the sharp BP phase transition happening before the optimal threshold, has many similarities with what appears in LDPC codes [147]. It is not a priori trivial because LDPC are codes over finite fields, the sparse superposition codes scheme works in the continuous framework and LDPC codes are decoded by loopy belief propagation whereas sparse superposition codes are decoded by AMP, a Gaussian approximation of loopy BP, see sec. 4.3.3. However, they arise due to a deep connection to compressed sensing where these phenomena (phase transition, spatial coupling, ...) are well known [34, 35, 110, 119] as discussed in sec. 5.1.1, and we shall make use of this connection extensively.

As we will see, the AMP algorithm is naturally applied to sparse superposition codes as this scheme can be interpreted as a compressed sensing problem with structured sparsity. This scheme is actually the first example of error correction of a signal that is directly mapped to a compressed sensing problem. In the chap. 10, the approach is different as it is the noise that is reconstructed. The state evolution technique [104] is unfortunatly not rigorous for the present AMP approach because of the structured sparsity of the signal, but in spite of that, we conjecture that it is exact.

Note that reconstruction of structured signals is a new trend in compressed sensing theory that aims at going beyond simple sparsity by introducing more complex structures in the vector that is to be reconstructed. Other examples include group sparsity or tree structure in the wavelet coefficients in image reconstruction [28].

A recent work of Rush, Greig and Venkataramanan [148] also studied AMP decoding in superposition codes combined with power allocation. Using the same technics as in [104], they proove rigorously that sparse superposition codes under AMP deconding is capacity-achieving if a proper power allocation is used. This strengthen the claim that AMP is the tool of choice in the present problem. We will see, however, that spatial coupling leads to even better decoding results at finite size.

9.1.2 Main contributions of the present study

The main original results of the present study are listed below.

• A detailed derivation of the AMP decoder for sparse superposition codes for a generic power allocation. The derivation is self-contained and starts all the way from the canonical loopy BP equations, see sec. 4.3.3.

• An analysis of the performance of the AMP decoder from the state evolution recursions. It is done in full generality with and without power allocation, and with and without spatial coupling. It is shown in particular that AMP, for simple sparse superposition codes, suffers from a phenomenon similar to those of BP with LDPC codes: there exist a sharp BP transition different from the optimum one of the code itself beyond which the decoder performance suddenly drops.

• An analysis of the optimum performance of sparse superposition codes using the non-rigorous replica method. This leads in particular to a single-letter formulation of the $MMSE$ estimate which we conjecture to be exact. The connection and consistency with the results coming from the state evolution approach is also underlined, see sec. 5.4.

• The large section limit for the behavior of AMP is studied, and we compute its limit rate, the asymptotic BP threshold $R_{BP}^{\infty} < C$ where C is the Shannon capacity of the channel. Studying as well the optimal threshold in this limit, we reconfirm using the replica method that these codes are Shannon capacity achieving.

• We also show that, with a proper power allocation, the BP threshold that was blocking the AMP decoder disappears so that AMP becomes asymptotically capacity achieving over the AWGN in the large section limit.

• Building on the connection with compressed sensing [34, 35, 110] we also show that the use of spatial coupling [109] for sparse superposition codes is an alternative way to obtain capacity achieving performances with AMP.

• We present an extensive numerical study at finite blocklength, showing that despite improvements of the scheme thanks to power allocation, a properly designed spatially-coupled coding matrix seems to allow better performances and robustness to noise for decoding over finite size signals.

• Furthermore we discuss a more practical scheme where the random coding operators are replaced by fast ones based on an Hadamard construction, see chap. 7. We show that this allows a close to linear time algorithm able to deal with very large signals, yet performing very well at large rate for finite signals. We study the efficiency of these operators combined with sparse superposition codes with or without spatial coupling.

Finally, we note that this work differs from the mainstream of existing literature. While a large part of the existing coding theory literature provides theorems, part of this work, that

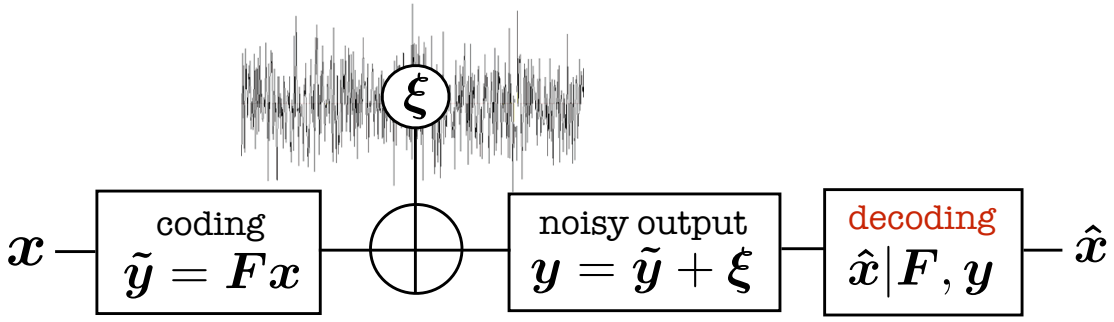

Figure 9.1 – Sending information through the AWGN channel with superposition codes: the message $\mathbf{x}$, created such that it has only a single non zero element in each of its L sections, is first coded by a linear transform, $\tilde{\mathbf{y}} = \mathbf{F}\mathbf{x}$. The resulting codeword is then sent through the AWGN channel that adds an i.i.d Gaussian noise $\boldsymbol{\xi}$ with zero mean and a given variance Δ to each components. The receptor gets the corrupted codeword $\mathbf{y}$ and must estimate $\hat{\mathbf{x}}$ as close as possible from $\mathbf{x}$ from the knowledge of $\mathbf{F}$ and $\mathbf{y}$. Perfect decoding happens if $\hat{\mathbf{x}} = \mathbf{x}$.

using the replica method, is based on statistical physics methods that are conjectured to give exact results. While many results obtained with these methods on a variety of problems have indeed been proven later on, a general proof that these methods are rigorous is not known yet. Note, however, that the state evolution technique has been turned into a rigorous tool under control in many similar cases [104, 149]. The present approach does not verify the assumptions required for the proofs to be valid because of the structured sparsity of the signal, but nevertheless we conjecture that the analysis remains exact. We thus expect that both the replica and state evolution analyzes are exact and believe it is only a matter of time before they are fully proven.

9.2 Sparse superposition codes

Suppose you want to send a generic message $\mathbf{s}$ made of L symbols through an AWGN channel, where each symbol belongs to an alphabet composed of B letters : $\mathbf{s} := [s_l : s_l \in \{1, \dots, B\}]_{l=1}^{L}$. Starting from a standard binary representation of $\mathbf{s}$, it is of course trivial to encode it in this form.

An alternative and highly *sparse* representation is given by the sparse superposition codes scheme: the representation $\mathbf{x}$ of this message $\mathbf{s}$ is made of L sections of size B, where only a *unique* value is $\neq 0$ in each section at the location corresponding to the original symbol. We will consider each non zero value to be positive as it can be interpreted as an input energy in the channel. Thus if the i^{th} component of the original message $\mathbf{s}$ is the k^{th} symbol of the alphabet, the i^{th} section of $\mathbf{x}$ contains only zeros, except at the position k where there is a positive value (which amplitude depends on the power allocation).

As an example, in the simplest setting where the *power allocation* is $c_l = 1 \; \forall \; l \in \{1, \dots, L\}$ (where c_l is the positive constant appearing in the l^{th} section), if $\mathbf{s} = [a, c, b, c]$, where the alphabet has

only three symbols $\{a,b,c\}$, then its sparse representation $\mathbf{x}$ is made of four sections which are $\mathbf{x} = [[100],[001],[010],[001]]$. The l^{th} section of $\mathbf{x}$ will be denoted $\mathbf{x}_l := [x_i]_{i\in l}$ where l is both the set of indices corresponding to the 1-d components of $\mathbf{x}$ in the l^{th} section or the index of the section depending on the context.

In sparse superposition codes, $\mathbf{x}$ is then encoded through a linear transform by application of an operator $\mathbf{F}$ of dimension $M \times N$ to obtain a codeword $\tilde{\mathbf{y}}$ of dimension M, $\tilde{\mathbf{y}} = \mathbf{F}\mathbf{x}$ which is then sent through the Gaussian noisy channel and the receiver gets a corrupted version of it. This is summarized in Fig. 9.1.

The dimension of the coding operator $\mathbf{F}$ is linked to the section size B and the coding (or transmission) rate in bits per-channel use R. Defining $K := \log_2(B^L)$ as the number of informative bits carried by the signal $\mathbf{x}$ made of L sections of size B (i.e. its entropy (3.34) considering that all the messages are equiprobable, see sec. 3.5), we have:

$$
\begin{aligned}
R &:= K/M && (9.1)\\
&= L\log_2(B)/(\alpha N) && (9.2)\\
&= \log_2(B)/(B\alpha) && (9.3)\\
\Leftrightarrow \alpha &:= M/N = \log_2(B)/(RB) && (9.4)
\end{aligned}
$$

In what follows, we will concentrate on i.i.d Gaussian $\mathbf{F}$ elements with 0 mean and variance $v_{\mathbf{F}}$ in order to be able to obtain analytical results. We always fix to 1 the total power sent through the channel $P := ||\tilde{\mathbf{y}}||_2^2 = <\tilde{\mathbf{y}}^2> = 1$ by a proper rescaling of the variance of the elements of $\mathbf{F}$. The only relevant parameter is thus the signal-to-noise ratio:

$$
\mathrm{snr} := P/\Delta = 1/\Delta \tag{9.5}
$$

where Δ is the variance of the Gaussian noise of the AGWN channel. It allows to define the Shannon capacity $C = \log_2(1+\mathrm{snr})/2$ of the channel (3.91), see sec. 3.7.1 for its derivation.

The transmitted codeword $\tilde{\mathbf{y}}$ is corrupted through the model (3.18) by the AWGN channel. Let us now turn our attention to the decoding task. It is essentially a sparse linear estimation problem where we know $\mathbf{y}$ and need to estimate a sparse solution of $\mathbf{y} = \mathbf{F}\mathbf{x} + \boldsymbol{\xi}$. However the problem is different from the canonical compressed sensing problem [32] in that the elements of $\mathbf{x}$ are strongly correlated by the constraint that only a single element in each section is non-zero, see Fig. 9.2. We thus prefer to think about the problem as a multidimensional one: each section $l \in \{1,2,\ldots,L\}$ made of B 1-d variables in $\mathbf{x}$ is interpreted as *a single* B-d variable for which we have a strong prior information: it is zero in all dimensions *but* one where there is a fixed positive known value. Given its length, we thus know the vector must point in only one direction of the hypercube of dimension B. In this new setting, instead of dealing with a N-d vector $\mathbf{x}$ with elements $\{x_i\}_i^N$, we deal with a L-d vector $\mathbf{x}$ which elements $\{\mathbf{x}_l\}_l^L$ are B-d sections.

In this framework, the decoding problem becomes exactly of the kind considered in the

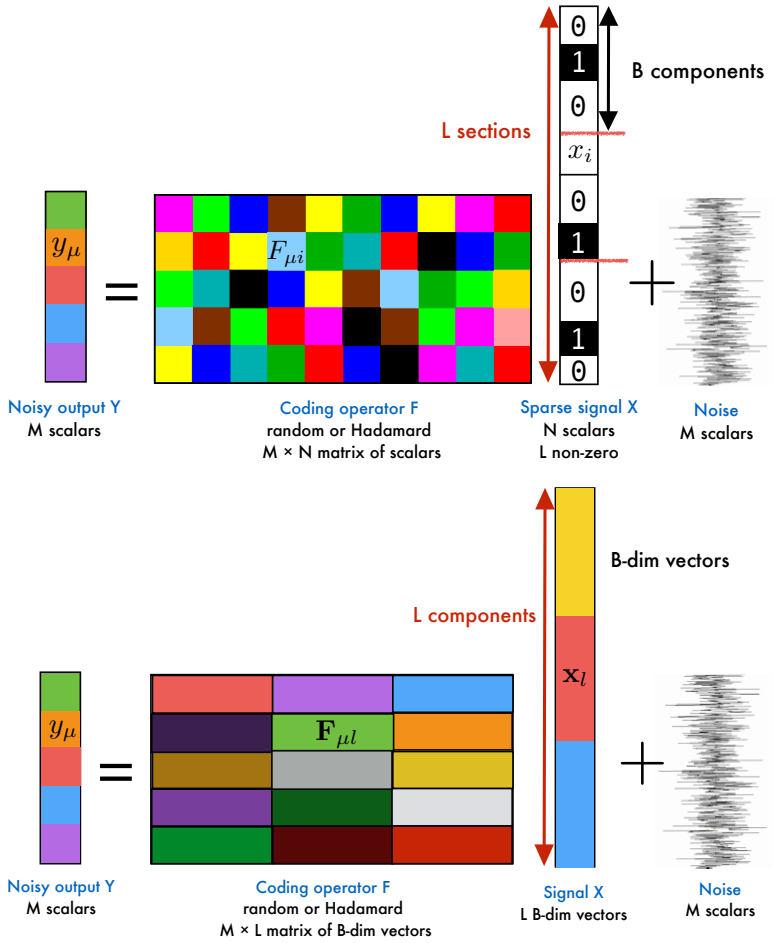

Figure 9.2 – **Up** : Representation of the estimation problem associated to the decoding of the sparse signal over the AWGN channel. All 1-d variables in the same section $\mathbf{x}_l = [x_i]_{i\in l}$ are strongly correlated due to the hard constraint that only one of them can be positive (or 1 in this example). **Down** : Reinterpreting the same problem in terms of B-d variables. Now, the matrix elements of the previous figure are concatenated to form B-d vectors $\{\mathbf{F}_{\mu l} := [F_{\mu i}]_{i\in l}\}$ that are applied (using the usual scalar product for vectors) on the associated B-d vectors representing the new components of the signal, the sections. In this new setting, all the sections are uncorrelated.

Bayesian approach to compressed sensing, see sec. 3.6.1 and for example [2, 34, 35, 89, 117]. We can thus directly apply these techniques to the present problem. From now on, we always consider that the true snr is accessible to the channel user, and thus can be used in the algorithm.

We will be interested in two error estimators, the MSE E and the section error rate SER. They are defined respectively as the MSE of the 1-d variables and the fraction of wrongly decoded

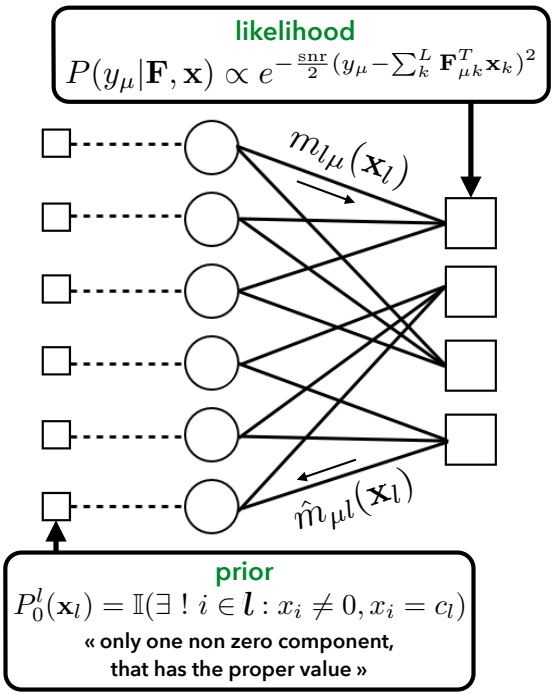

Figure 9.3 – Factor graph associated to the sparse superposition codes. It is a bipartite graph where the variable estimates $\{\mathbf{x}_l\}_l^L$ are represented by circles, the constraints (or factors) by squares. The variables are constrained by the M likelihood factors that enforce the system $\mathbf{y} = \mathbf{F}\mathbf{x}$ to be fulfilled up to Gaussian fluctuations due to the Gaussian noise of the AWGN channel. The prior constraints enforce each section to have only one non-zero component, that must be the value fixed by the power allocation. In the homogeneous operator case, the variables are connected to all the likelihood factors and vice versa and in the spatially-coupled case, only to a finite fraction that depends on the spatial coupling ensemble, see Fig. 5.4. The factor-to-node $\hat{m}_{\mu l}(\mathbf{x}_l)$ and node-to-factor $m_{l\mu}(\mathbf{x}_l)$ cavity messages are represented. They should stand on the same edge as they depend on the same variable but we put them on distinct edges for readibility purpose.

sections:

$$E := \frac{1}{N}\sum_i^N (x_i - \hat{x}_i)^2 \tag{9.6}$$

$$SER := \frac{1}{L}\sum_l^L \mathbb{I}(\mathbf{x}_l \neq \hat{\mathbf{x}}_l) \tag{9.7}$$

where $\mathbb{I}(A)$ is the indicator function of the event A which is one if A happens to be true, zero else and $\hat{\mathbf{x}} := [\hat{\mathbf{x}}_l]_l^L = [\hat{x}_i]_i^N$ is the final estimate of the signal by the decoder.

9.3 Approximate message-passing decoder for superposition codes

The only problem-dependent objects in the AMP Fig. 5.5 are the denoising functions $\{f_{a_i}, f_{c_i}\}$. Let us derive them for any power allocation $\{c_l > 0\}_l^L$. Here, the prior that matches the signal distribution by enforcing the constraint of having only one known value $c_l > 0$ per section is:

$$P_0^l(\mathbf{x}_l) := \frac{1}{B}\sum_{i\in l}^{B}\delta(x_i - c_l)\prod_{j\in l: j\neq i}^{B-1}\delta(x_j) \tag{9.8}$$

The denoisers which generic expressions are given by (4.115) and (4.116) are easily derived. We obtain the posterior average a_i^t and variance v_i^t of x_i at step t of the algorithm:

$$a_i^t := f_{a_i}((\mathbf{\Sigma}_{l_i}^t)^2, \mathbf{R}_{l_i}^t) = c_{l_i}\frac{\exp\left(-\frac{c_{l_i}(c_{l_i}-2R_i^t)}{2(\Sigma_i^t)^2}\right)}{\sum_{j\in l_i}^{B}\exp\left(-\frac{c_{l_i}(c_{l_i}-2R_j^t)}{2(\Sigma_j^t)^2}\right)} \tag{9.9}$$

$$v_i^t := f_{c_i}((\mathbf{\Sigma}_{l_i}^t)^2, \mathbf{R}_{l_i}^t) = a_i^t(c_{l_i} - a_i^t) \tag{9.10}$$

where $\mathbf{\Sigma}_{l_i}, \mathbf{R}_{l_i}$ are the AMP fields of the section l_i to which the i^{th} 1-d component of the signal belongs to. Combined with Fig. 5.5, we thus get the full AMP algorithm for sparse superposition codes with associated graphical model given by Fig. 9.3.

9.3.1 The fast Hadamard-based coding operator

In the present study, we use spatially-coupled operators constructed as in Fig. 7.1. Fig. 9.4 shows that when the signal sparsity increases, i.e. when the section size B increases, using Hadamard-based operators becomes quickly equivalent to using random i.i.d Gaussian ones in terms of performances (as observed in chap. 7). For the figure, we have fixed the snr = 100 and then plotted the distance in dB to the BP threshold $R_{BP}(\text{snr} = 100, B)$ at which the decoder starts to decode perfectly with Hadamard-based or random i.i.d Gaussian operators. We remind that R_{BP} is defined as the highest rate until which AMP decoding is optimal without the need of non constant power allocation or spatial coupling. It appears that at low section size, it is advantageous to use random operators but as B increases, structured operators quickly match their performances. The BP threshold is predicted by the state evolution analysis presented in Sec. 9.4.

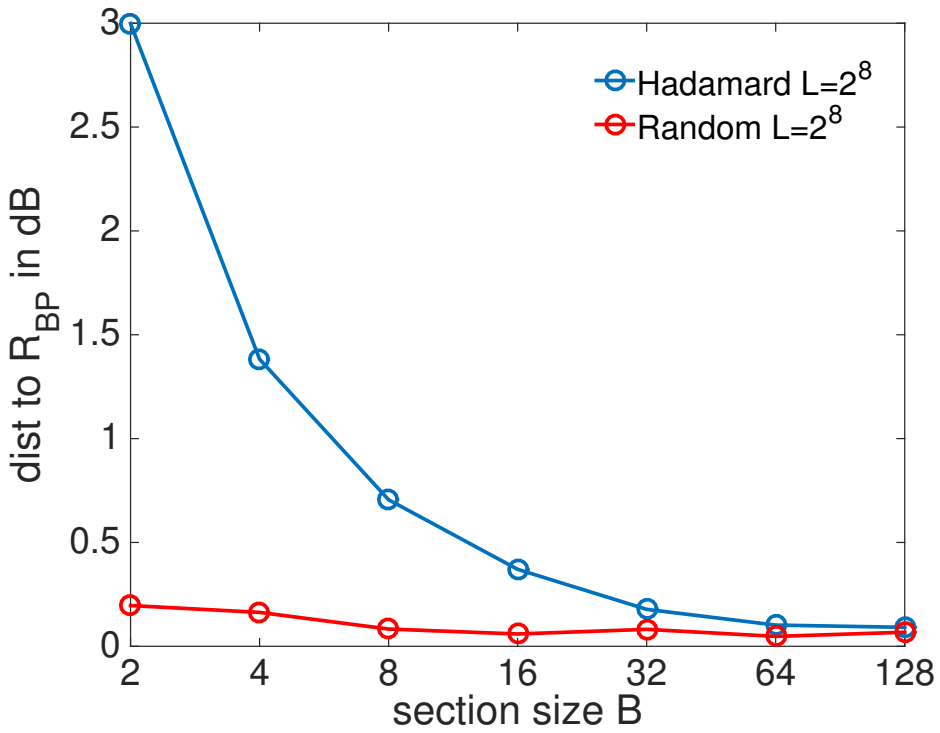

Figure 9.4 – Comparison between the distance in dB to the asymptotic BP threshold $R_{BP}(\text{snr} = 100, B)$ at which the AMP decoder with homogeneous Hadamard-based coding operators (blue) or random i.i.d Gaussian ones (red) starts to reach an $SER < 10^{-5}$, which essentially means perfect decoding for most of the instances. This is done for a fixed number of sections $L = 2^8$ and snr = 100. Each point have been averaged over 100 random instances. The BP threshold $R_{BP}(\text{snr} = 100, B)$ is obtained by state evolution analysis for each B. Decoding with Hadamard-based operators works poorly when the signal density increases (i.e. when B decreases), but matches quickly the random matrix performances as it decreases. Decoding with random Gaussian i.i.d matrices has a performance that is close to constant as a function of B at fixed L as it should (the relevant signal size is L).

9.4 State evolution analysis for random i.i.d homogeneous operators with constant power allocation

As usual, for the state evolution analysis we consider the case of an i.i.d Gausian matrix $\mathbf{F}$, such that the recursions obtained in sec. 5.3 are valid. Here we consider the matrix to be homogeneous and a constant power allocation. We will use the state evolution to predict the results with the Hadamard operator as well, as we have shown in chap. 7 that the analysis derived in the random i.i.d case is a good predictive tool of the behavior of the AMP decoder with structured operators, despite not perfect nor rigorous.

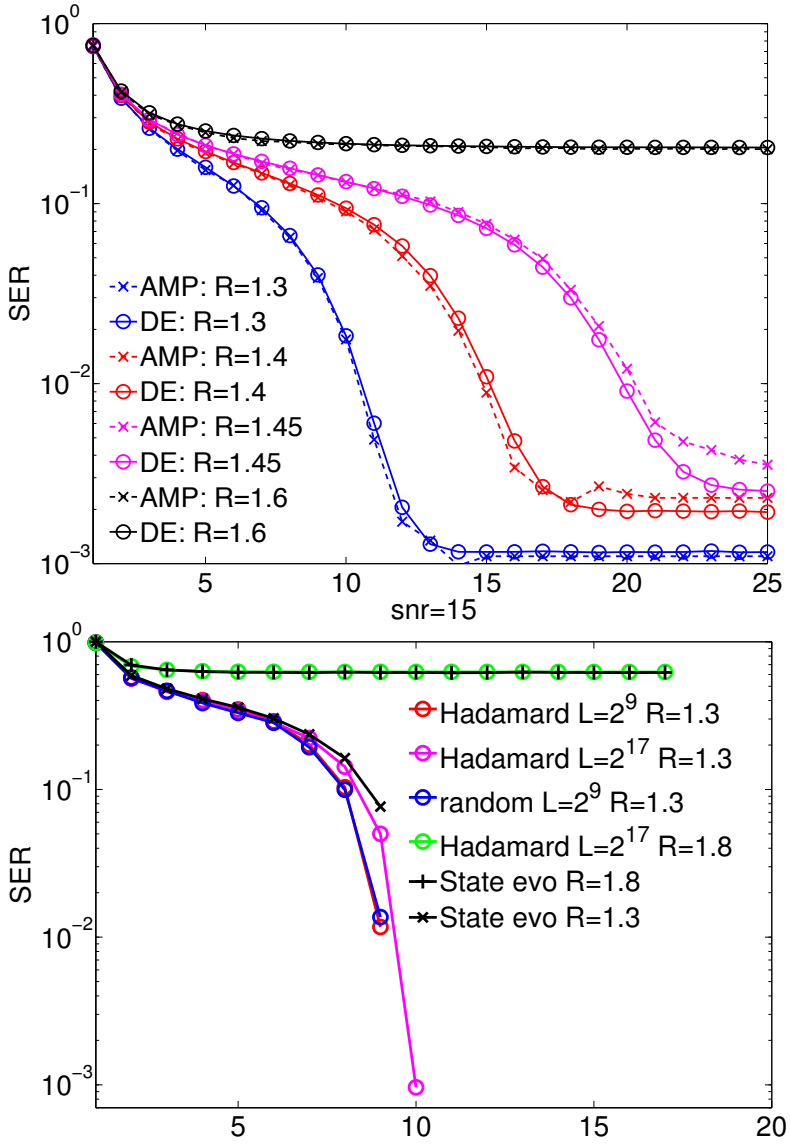

Figure 9.5 – **Up** : The state evolution prediction (solid lines) of the SER^t compared to the actual one of the algorithm on single instances for snr = 15, different rates and a section size $B = 4$ (to be compared to the BP threshold at $R_{BP}(\text{snr} = 15, B = 4) = 1.55$). The matrix is i.i.d Gaussian and we consider a constant power allocation $\{c_l = 1\}_l^L$. The state evolution is computed by monte carlo with a sample size of 10^7 and the signal size for AMP is $L = 2^{13}$. **Down** : The same as the upper plot with snr = 15, different rates R (one above and one below R_{BP}) but with a larger section size $B = 64$. The BP threshold is here $R_{BP}(\text{snr} = 15, B = 64) = 1.47$. We consider different signal sizes L and homogeneous Hadamard-based or purely random i.i.d Gaussian operators. The state evolution is computed by monte carlo with a sample size of 10^6 as B is larger and thus the monte carlo computation requires more time. The finite size curves stop without reaching a noise floor because the recovery is actually perfect and the final $SER = 0$ which is due to the finite size effects. The same happens for the theoretical curves that reach 0 due to finite numerical precision. We observe that when the signal size L is big (and thus that we must use the Hadamard operator), the algorithm behavior follows the theoretical predictions closely.

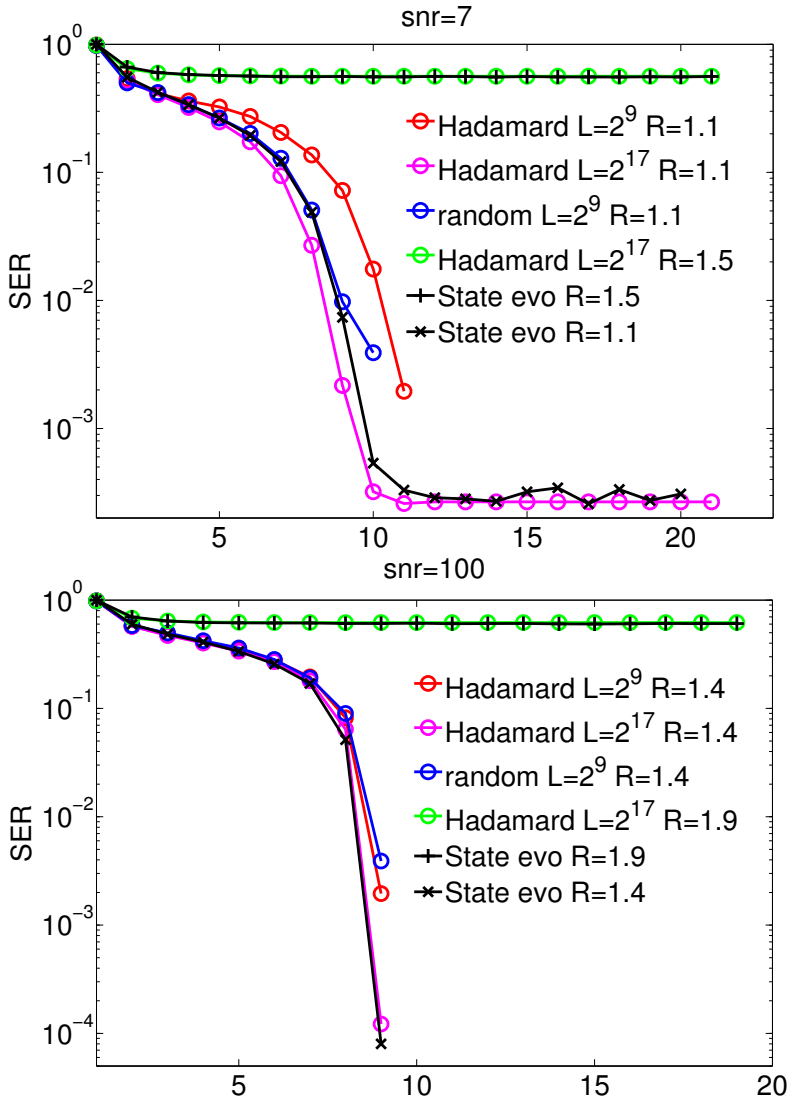

Figure 9.6 – The state evolution prediction of the section error rate SER^t (black curves), compared to the actual one of the AMP decoder for snr $= 7/100$, a section size $B = 64$, different rates R (one above and one below the BP threshold $R_{BP}(\text{snr} = 7, B = 64) = 1.275, R_{BP}(\text{snr} = 100, B = 64) = 1.625$), different signal sizes L in the homogeneous Hadamard-based or purely random i.i.d Gaussian operator case with constant power allocation $\{c_l = 1\}_l^L$. The state evolution is computed by monte carlo with a sample size of 10^6.

We define $\tilde{\Sigma}^{t+1}(E^t) := \Sigma^{t+1}(E^t)\sqrt{\log(B)}$ as this expression will be more convenient. $\Sigma^{t+1}(E^t)$ is given by (5.74) where we use (9.4) to express α in function of the rate. Starting from (5.73), using (5.75) and the prior (9.8) for sparse superposition codes, we get:

$$E^{t+1} = \frac{1}{B}\int \mathscr{D}\mathbf{z}\Big([f_{a_{1|1}}(\tilde{\Sigma}^{t+1})^2, \mathbf{z}) - 1]^2 + (B-1) f_{a_{2|1}}((\tilde{\Sigma}^{t+1})^2, \mathbf{z})^2\Big) \tag{9.11}$$

where we define:

$$f_{a_{i|i}}(\tilde{\Sigma}^2,\mathbf{z}) := \left[1+e^{-\frac{\log(B)}{\tilde{\Sigma}^2}}\sum_{1\leq j\leq B:j\neq i}^{B-1} e^{\frac{\sqrt{\log(B)}(z_j-z_i)}{\tilde{\Sigma}}}\right]^{-1} \tag{9.12}$$

$$f_{a_{j|i}}(\tilde{\Sigma}^2,\mathbf{z}) := \left[1+e^{\frac{\log(B)}{\tilde{\Sigma}^2}+\frac{\sqrt{\log(B)}(z_j-z_i)}{\tilde{\Sigma}}}+\sum_{1\leq k\leq B:k\neq i,j}^{B-2} e^{\frac{\sqrt{\log(B)}(z_k-z_i)}{\tilde{\Sigma}}}\right]^{-1} \tag{9.13}$$

Under the constant power allocation assumption, the quantity $f_{a_{i|i}}$ ($f_{a_{j|i}}$) can be interpreted as the asymptotic AMP posterior probability estimate of the i^{th} component (j^{th} component) to be the 1 given that it is indeed the 1 in the signal (given that it is actually the i^{th} component that is the 1 in the signal). In this approach, there is a one to one correspondance from the value of the MSE to the SER thanks to the mapping:

$$SER^{t+1} = \int \mathscr{D}\mathbf{z}\, \mathbb{I}\left(\exists\, j\in\{2,\dots,B\} : f_{a_{j|1}}((\tilde{\Sigma}^{t+1})^2,\mathbf{z}) > f_{a_{1|1}}((\tilde{\Sigma}^{t+1})^2,\mathbf{z})\right) \tag{9.14}$$

From this equation, we can exactly predict the asymptotic $L\to\infty$ evolution in time of the algorithm, such as in Fig. 9.5 and Fig. 9.6. The state evolution on these plots represent the iteration of (9.14), (combined with (9.11), (5.74)) for different experimental settings (snr, R, B) using homogeneous Hadamard-based or random i.i.d Gaussian matrices. (9.14) and (9.11) are computed at each time step by monte carlo technique. We observe that for the snr = 15/100, $B=64$ cases, the experimental and theoretical curves stop at some iteration without reaching a noise floor. For the experimental curves, this is due to the fact that in order to observe an $SER\in O(\epsilon)$, there must be at least $L\approx 1/\epsilon$ sections which is not the case for signals of reasonnable sizes, when the asymptotic SER is very small. This finite size effect is actually in favor of the reconstruction performances. In fact, when the rate is below the BP threshold, the decoding is usually perfect and is found to reach with high probability $SER=0$. The black asymptotic curves should anyway reach a finite error floor but they do not because it is so low that the sample size used in the monte carlo computation sould be way too large to deal with by the same argument. But on Fig. 9.5, for smaller $B=4$ the error floor is higher than for $B=64$ and thus we can see it numerically.

Another observation, natural from the definition of the state evolution as an asymptotic analysis, is that the theoretical and experimental results match better for larger signals. At rate $R>R_{BP}$ (green or black experimental curves and associated theoretical ones), we see that the AMP decoder does not reconstruct the signal and converges to an high SER solution well predicted by the state evolution. On the contrary, below the threshold, the reconstruction works fine up to an error floor dependent on the parameters (B, snr, R). We also observe as in chap. 7 that the state evolution predicts well the final performances of AMP with Hadamard-based operators.

9.5 State evolution analysis for spatially-coupled i.i.d operators or with power allocation

Thanks to the spatial-coupling we can asymptotically reach the optimal rate. The total rate is a function of the rate of the blocks. From (5.114) combined with (9.4) we deduce it:

$$R = \frac{L_c R_{rest} R_{seed}}{(L_r - 1)R_{seed} + R_{rest}} \underset{L_c, L_r \to \infty}{\longrightarrow} R_{rest} < R_{opt}(\text{snr}, B) \tag{9.15}$$

where R_{rest} can be asymptotically as large as the Bayes optimal rate $R_{opt}(\text{snr}, B)$ defined as the highest rate until which the superposition codes scheme allows to decode, see sec. 5.1.1. Let us now derive the state evolution recursions in the spatially-coupled operator case. As in the homogeneous case, defining the rescaled $\tilde{\Sigma}_c^{t+1} := \Sigma_c^{t+1}\sqrt{\log(B)}$, from (5.174), (5.175) and (5.176), we obtain the following state evolution for the MSE E_c inside the block c in the $L \to \infty$ limit:

$$E_c^{t+1} = \frac{1}{B}\int \mathscr{D}\mathbf{z}\left([f_{a_{1|1}}((\tilde{\Sigma}_c^{t+1})^2, \mathbf{z}) - 1]^2 + (B-1) f_{a_{2|1}}((\tilde{\Sigma}_c^{t+1})^2, \mathbf{z})^2\right) \tag{9.16}$$

$$\tilde{\Sigma}_c^{t+1}\left(\{E_{c'}^t\}_{c'}^{L_c}\right) = \sqrt{\log(B)}\left[B\sum_r^{L_r} \frac{\alpha_r J_{rc}}{L_c/\text{snr} + B\sum_{c'}^{L_c} J_{rc'} E_{c'}^t}\right]^{-1/2} \tag{9.17}$$

where the f_a functions (9.12), (9.13) are defined in the previous section and where the mapping to the SER_c^{t+1} per block is given by:

$$SER_c^{t+1} = \int \mathscr{D}\mathbf{z}\, \mathbb{I}\left(\exists\, j \in \{2,\dots,B\} : f_{a_{j|1}}((\tilde{\Sigma}_c^{t+1})^2, \mathbf{z}) > f_{a_{1|1}}((\tilde{\Sigma}_c^{t+1})^2, \mathbf{z})\right)\ \forall\ c \in \{1,\dots,L_c\} \tag{9.18}$$

Thanks to this analysis, we can now predict the asymptotic SER per block in the signal estimate by the AMP decoder. Fig. 9.7 shows a comparison of the SER per block $\{SER_c^t\}_c^{L_c}$ predicted by state evolution (black curves) with the actual SER per block of the superposition codes with the AMP decoder combined with an Hadamard-based spatially-coupled operator on a single instance. The discrepancies between the theoretical and experimental curves come from the fact that state evolution is derived for random i.i.d Gaussian matrices, but the final error using these Hadamard operators is the same as predicted by state evolution as observed in the snr = 7 case. In the high snr regime, the curves stop for the same reasons as the Fig. 9.6 of the previous section and it means that the decoding was perfect. As noted in chap. 7, structured operators converge faster to the predicted final error than purely i.i.d matrices as predicted by the state evolution.

9.5.1 State evolution for power allocated signals

We now observe that we can trivially obtain the state evolution for any power allocation of the signal encoded with a random i.i.d Gaussian matrix from the previous analysis, thanks to the transformation of Fig. 9.8: starting from an homogeneous matrix and a given power allocated

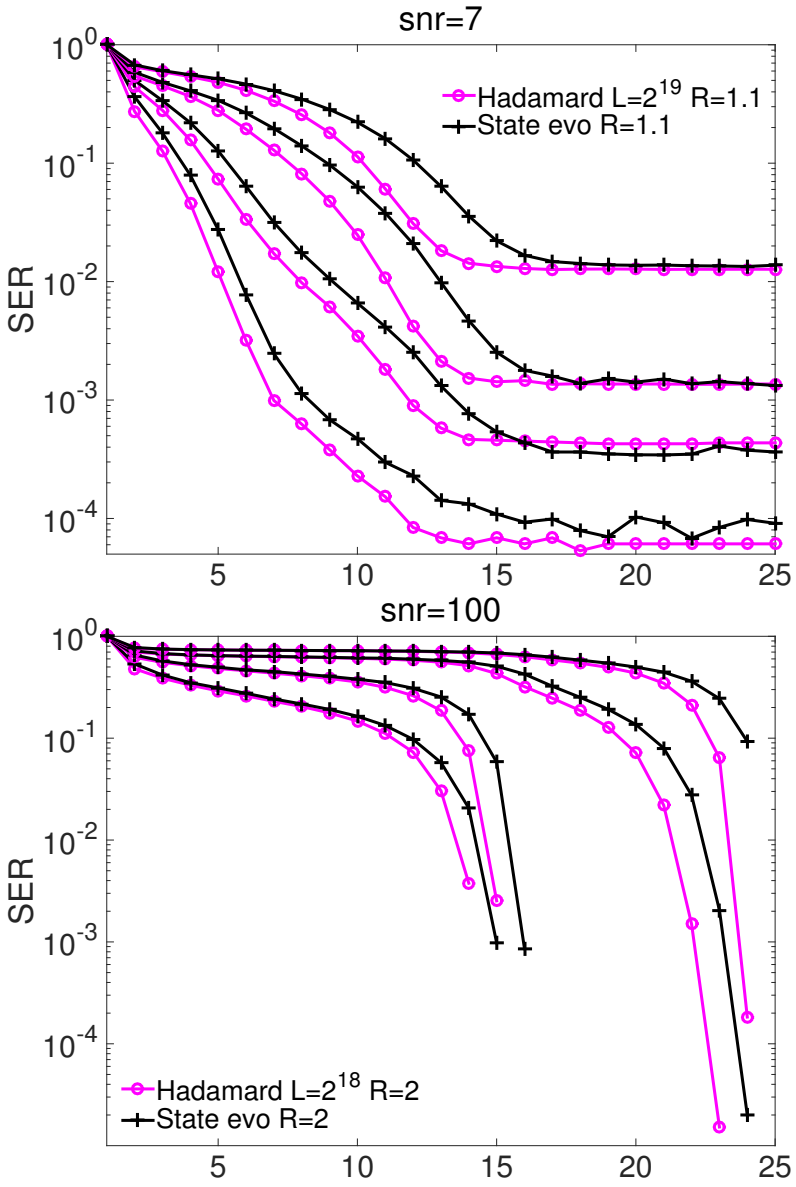

Figure 9.7 – The state evolution prediction of the section error rate $\{SER_c^t\}_c^{L_c=4}$ for each of the four induced block of the signal (see Fig. 7.1) as a function of time (black curves), compared to the actual error of the algorithm for snr = 7/100, different rates R, a section size $B = 32$, different signal size L values with a spatially-coupled Hadamard-based operator. The operator is drawn from the ensemble $(L_c = 4, L_r = 5, w = 2, \sqrt{J} = 0.6, R, \beta_{\text{seed}} = 1.5)$. The power allocation is constant. The state evolution is computed by monte carlo with a sample size of 10^6. The finite size curves at high snr stop without reaching a noise floor because the recovery is actually perfect due to the finite size effects, the same happens for the theoretical curves due to finite numerical precision. In the low snr = 7 case, the error floor (different in each block) is well predicted by state evolution.

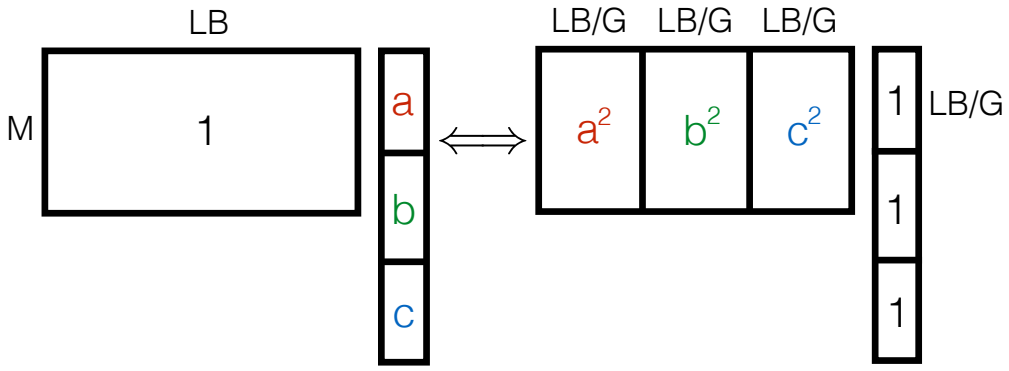

Figure 9.8 – The figure shows how to convert a system with any power allocated signal encoded through an homogeneous operator with elements of variance $1/L$ into an equivalent system with constant power allocation encoded by a structured operator. The values on the matrix represent the variance of the elements of the matrix (we drop here the rescaling by $1/L$), the values on the signal represent the non zero values of the sections that belong to a given group: here the signal is decomposed into $G = 3$ groups, and all the sections inside the first group have a non zero value equal to a, and so on. The transformation is done by structuring the operator into block-columns, with as many blocks as different values in the power allocation, or groups: if a column of the original matrix acts on a component of a section where the non zero value is u, then this column variance is multiplied by u^2 in the new structured operator (such that the elements of this column are multiplied by u). The different sizes of the matrix blocks and signal groups are represented.

signal, we convert the system into a structured matrix with a constant power allocated signal.

Suppose the signal is decomposed into G groups, where inside the group g, the power allocation is the same for all the sections belonging to this group and equals c_g. Now one must create a structured operator starting from the original homogeneous one, decomposing it into blocks with LB/G columns and multiply all the elements of the block-column g by c_g, as shown in Fig. 9.8. The system with this new operator acting on a constant power allocated signal is totally equivalent to the original system and we have the state evolution of this new system from the previous analysis. Using (9.17) in the present setting, one has to be careful with the value of α_r defined as the number of lines over the number of columns of the block-line r (which is unique). Here there is a unique value that equals $M/(N/G) = G\alpha$ where α is defined as the original measurement rate (9.4). Given that, $L_c = G$ we finally obtain:

$$E_g^{t+1} = \frac{1}{B}\int \mathscr{D}\mathbf{z}\left([f_{a_{1|1}}((\tilde{\Sigma}_g^{t+1})^2,\mathbf{z})-1]^2 + (B-1)f_{a_{2|1}}((\tilde{\Sigma}_g^{t+1})^2,\mathbf{z})^2\right) \tag{9.19}$$

$$\tilde{\Sigma}_g^{t+1}\left(\{E_{g'}^t\}_{g'}^G\right) = \sqrt{\log(B)}\left[B\frac{\alpha c_g^2}{1/\mathrm{snr} + B/G\sum_{g'}^G c_{g'}^2 E_{g'}^t}\right]^{-1/2} \tag{9.20}$$

The square c_g^2 appears because multiplying the matrix elements by c_g multiply their variance by its square.

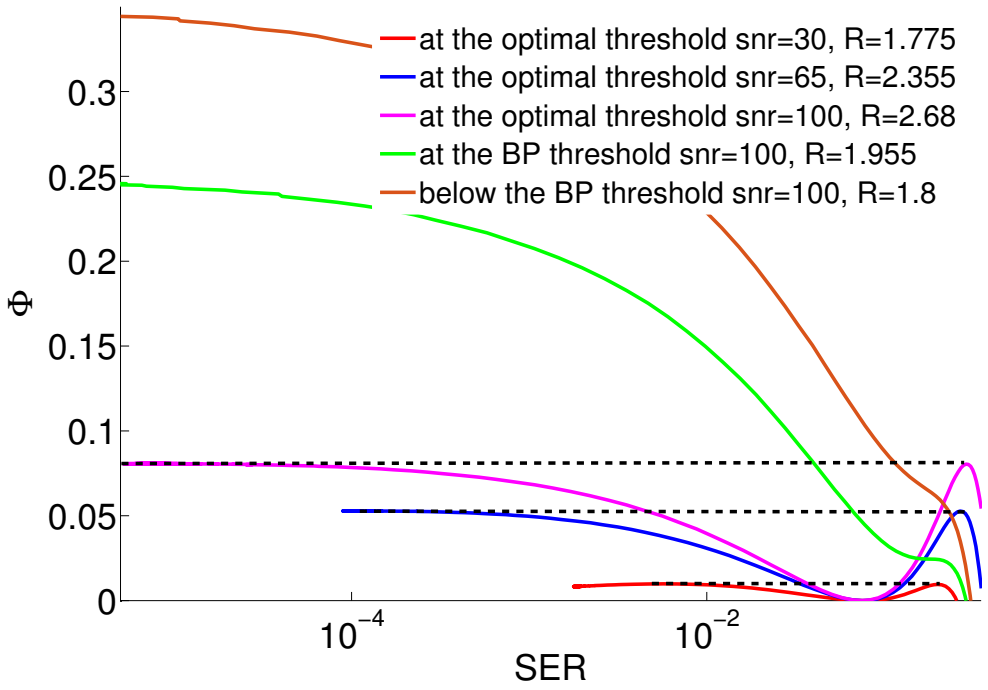

Figure 9.9 – The Bethe free entropy (or potential) $\Phi(SER)$ for $B = 2$, different rates and snr. The maxima of the curves correspond to the typical SER which are fixed points of the state evolution equations (9.14) for a given set of parameters (R, B, snr). The global maximum is the (exponentially) most probable SER solution, i.e. the equilibrium state refered also as the optimal SER. The curves are obtained by numerical integration of (9.21). The optimal threshold $R_{opt}(B, \text{snr})$ is defined as the rate where the high and low error maxima have same height (i.e. same probability), see pink, blue and red curves. The BP threshold $R_{BP}(B, \text{snr})$ is the rate at which the metastable state a high error that blocks the convergence of AMP appears, see green curve. The plot illustrates how the gap at the optimal threshold between the two maxima increases with the snr. **snr = 100** : Here for rates larger than $R > 2.68$, the optimal SER jumps from a low value to a large $O(1)$ one (pink curve). This defines the maximum possible rate (to compare here to $C = 3.3291$) below which acceptable performance can be obtained with AMP combined with spatial coupling or non constant power allocation. For $R < 2.68$, the SER is much lower (and decay with R). The AMP decoder Fig. 5.5 allows to perform an ascent of this function. As long as the maximum is unique (i.e. for $R < 1.955$, see green curve), it will be able to achieve the predicted optimal performance with no need of spatial coupling or non constant power allocation in the large size limit, as in the case of the brown curve.

9.6 Replica analysis and phase diagram

Let us now compute the asymptotic $L \to \infty$ free entropy by the replica method. This potential of the SER is derived in the constant power allocation case. We are more interested in this case as we will show later with Fig. 9.16 that anyway, the most efficient reconstruction scheme is with constant power allocation combined with spatial coupling. Pluging the prior (9.8) in the

general free entropy expression under the matching prior condition (5.48), we directly obtain:

$$\Phi_B(E) = -\frac{\log_2(B)}{2R}\left(\log(1/\text{snr}+BE)+\frac{1-BE}{1/\text{snr}+BE}\right) + \int \mathscr{D}\mathbf{z}\log\left(e^{\frac{\log(B)}{2\tilde{\Sigma}(E)^2}+\frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}(E)}}+\sum_{i=2}^{B} e^{-\frac{\log(B)}{2\tilde{\Sigma}(E)^2}+\frac{\sqrt{\log(B)}z_i}{\tilde{\Sigma}(E)}}\right) \tag{9.21}$$

where E is the MSE and $\tilde{\Sigma}(E)^2 := \log(B)\Sigma(E)^2$ together with (5.49). Going from this expression to $\Phi_B(SER)$ is possible thanks to (9.14) at the fixed point.

As discussed in sec. 5.1.1, the SER values associated with the maxima of this potential correspond to fixed points of the state evolution equations (9.14). The information brought by this analysis that is not explicitly included in the state evolution analysis is the identification of the phase in which the system is (easy/hard/impossible inference) for a given set of parameters (R, B, snr). In particular, the hard phase can only be identified knowing where is situated the global maxima. This infomation cannot be extracted from the state evolution: in the hard phase, (9.14) will converge to a local maxima that depends on its initialization, but without telling which of the two is the global one. The state evolution can thus identify the appearance of the hard phase at the BP transition, when the second fixed point appear but it cannot identify the optimal transition as it requires to know the relative height of the maxima.

An example of this potential (9.21) in the $(B = 2, \text{snr} = 100)$ case for various R is shown on Fig. 9.9. As discussed in sec. 5.1.1, the AMP algorithm follows a dynamic that can be *interpreted* as a gradient ascent of this free entropy, which starts the ascent from an high error state, i.e. a random guess for the signal estimate. The brown curve thus corresponds to an easy case as the global maximum is unique and corresponds to a low error state. The green curve corresponds to the BP threshold, which marks the appearance of the hard inference phase. For higher rate, the problem is to reach the global maximum despite the high error metastable state: it is achieved using spatial coupling. The pink curve is the optimal transition which marks the entrance in the impossible inference phase. Below this rate, the AMP algorithm combined with spatial coupling or well designed power allocation is theoretically able to decode.

The blue and red curves also correspond to optimal transitions at higher noise levels and we notice that as the snr decreases the relative height between the maxima decreases and the basin of attraction of the maxima tends to be more flat. This explains why it is easier to decode finite size signals closer to the optimal threshold with spatial coupling for larger snr: as the basin of attraction of the equilibrium has a more pronounced gradient and the global maximum is higher, the dynamic climbs more easily to the maximum. The solutions associated to the maxima of this potential are exponentially more probable than the other ones (the probability of any state is proportional to the exponential of its free entropy times the system size) but at finite size, the factors gained thanks to an higher maximum can help a lot the algorithm convergence. For example let us assume that we transmit information at a rate $R_{BP} < R < R_{opt}$ which is such that the difference in the free entropy of the equilibrium and metastable states is $\Delta\Phi > 0$. Furthemore the system size is L. It implies that the ratio of the

probability $P_{eq.}$ of the equilibrium state over the metastable state's one $P_{meta.}$ is $P_{eq.}/P_{meta.} \propto \exp(L\Delta\Phi)$. If the difference $\Delta\Phi \to \Delta\Phi/\tau$ is divided by $\tau > 0$ because we increase the rate to get closer to R_{opt} or because the snr decreases, the system size must be multiplied by τ as well to keep this ratio constant, thus this $\Delta\Phi$ does matter at small finite size.

Furthermore, the SER gap separating the high and low error states decreases as well with the snr which implies that the error floor of the decoding increases. At some value of snr, there are no more two maxima for any rate and the transition becomes continuous as we observed in chap. 6.

9.6.1 Large section limit of the superposition codes by analogy with the random energy model

In order to get the asymptotic behavior in the section size of this potential, we need to compute the asymptotic value $I := \lim_{B\to\infty} I_B$ of the integral I_B that appears in (9.21). We shall drop the dependency of $\tilde{\Sigma}$ in E to avoid confusions and compute:

$$I_B := \int \mathscr{D}\mathbf{z} \log \underbrace{\left(e^{\frac{\log(B)}{2\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}}} + \sum_{i=2}^{B} e^{-\frac{\log(B)}{2\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_i}{\tilde{\Sigma}}} \right)}_{:=K_B(\mathbf{z})} \tag{9.22}$$

$$= \mathbb{E}_{\mathbf{z}}\{\log(K_B(\mathbf{z}))\} \tag{9.23}$$

We shall adopt here the vocabulary of statistical mechanics [24]: this is formally a problem of computing the average of the logarithm of a partition function $K_B(\mathbf{z})$ of a system with B (disordered) states. Indeed, one can re-write (9.23) as:

$$I_B = -\frac{\log(B)}{2\tilde{\Sigma}^2} + \int \mathscr{D}\mathbf{z} \log\left(e^{\frac{\log(B)}{\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}}} + \sum_{i=2}^{B} e^{\frac{\sqrt{\log(B)}z_i}{\tilde{\Sigma}}} \right) \tag{9.24}$$

$$= -\frac{\log(B)}{2\tilde{\Sigma}^2} + \int \mathscr{D}\mathbf{z} \log\left(\mathscr{Z}_1(z_1) + \mathscr{Z}_2\left(\{z_i\}_{i=2}^{B}\right) \right) \tag{9.25}$$

where:

$$\mathscr{Z}_1(z_1) := \exp\left(\frac{\log(B)}{\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}} \right) \tag{9.26}$$

$$\mathscr{Z}_2\left(\{z_i\}_{i=2}^{B}\right) := \sum_{i=2}^{B} \exp\left(\frac{\sqrt{\log(B)}z_i}{\tilde{\Sigma}} \right) \tag{9.27}$$

In fact $\mathscr{Z}_2$ is formally known as a random energy model in the statistical physics literature [24, 150], a statistical physics model where i.i.d energy levels are drawn from some given distribution. This analogy can be further refined by writing the energy as $u_i = -\sqrt{\log(B)}z_i$ and by denoting $\tilde{\Sigma}$ as the temperature. In this case, a standard result [24, 150, 151] is:

• The asymptotic limit for large B of $\mathscr{J} := \log_B(\mathscr{Z}_2)$ exists, and is concentrated (i.e. it does not

depend on the disorder $\mathbf{z}$ realization).

• It is equal to $\mathscr{J} = \sqrt{2}/\tilde{\Sigma}\,\mathbb{I}\left(\tilde{\Sigma} < 1/\sqrt{2}\right) + \left(1/(2\tilde{\Sigma}^2) + 1\right)\,\mathbb{I}\left(\tilde{\Sigma} > 1/\sqrt{2}\right)$.

We can thus now obtain the value of the integral by comparing $\mathscr{Z}_1$ and $\mathscr{Z}_2$ and keeping only the dominant term. First let us consider the case where $\tilde{\Sigma} < 1/\sqrt{2}$ and B is large:

$$\log_B(\mathscr{Z}_1 + \mathscr{Z}_2) = \log_B(\mathscr{Z}_1) + \log_B\left(1 + \frac{\mathscr{Z}_2}{\mathscr{Z}_1}\right) \tag{9.28}$$

$$\approx \log_B(\mathscr{Z}_1) + \frac{1}{\log(B)}\exp\left(-\frac{\log(B)}{\tilde{\Sigma}^2} - \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}} + \frac{\log(B)\sqrt{2}}{\tilde{\Sigma}}\right) \tag{9.29}$$

$$\approx \log_B(\mathscr{Z}_1) \tag{9.30}$$

$$\Rightarrow \lim_{B\to\infty}\frac{I_B}{\log(B)} = -\frac{1}{2\tilde{\Sigma}^2} + \frac{1}{\log(B)}\int \mathscr{D}\mathbf{z}\left(\frac{\log(B)}{\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}}\right) \tag{9.31}$$

$$= \frac{1}{2\tilde{\Sigma}^2} \tag{9.32}$$

using (9.25) and where $\log_B$ is the base B logarithm. If, however, $\tilde{\Sigma} > 1/\sqrt{2}$ and B is large, then:

$$\log_B(\mathscr{Z}_1 + \mathscr{Z}_2) = \log_B(\mathscr{Z}_2) + \log_B\left(1 + \frac{\mathscr{Z}_1}{\mathscr{Z}_2}\right) \tag{9.33}$$

$$\approx \log_B(\mathscr{Z}_2) + \frac{1}{\log(B)}\exp\left(\frac{\log(B)}{\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}} - \frac{\log(B)}{2\tilde{\Sigma}^2} - \log(B)\right) \tag{9.34}$$

$$\approx \log_B(\mathscr{Z}_2) \tag{9.35}$$

$$\Rightarrow \lim_{B\to\infty}\frac{I_B}{\log(B)} = -\frac{1}{2\tilde{\Sigma}^2} + \frac{1}{2\tilde{\Sigma}^2} + 1 = 1 \tag{9.36}$$

This leads to:

$$\lim_{B\to\infty}\frac{I_B}{\log(B)} = \frac{1}{2\tilde{\Sigma}^2}\,\mathbb{I}\left(\tilde{\Sigma} < 1/\sqrt{2}\right) + \mathbb{I}\left(\tilde{\Sigma} > 1/\sqrt{2}\right) \tag{9.37}$$

From these results combined with (9.21), we now can give the asymptotic value of the potential:

$$\phi(E) := \lim_{B\to\infty}\left(\frac{\Phi_B(E)}{\log(B)}\right) \tag{9.38}$$

$$= -\frac{1}{2R\log(2)}\left(\log(1/\mathrm{snr} + BE) + \frac{1 - BE}{1/\mathrm{snr} + BE}\right) + \max\left(1, \frac{1}{2\tilde{\Sigma}^2(E)}\right) \tag{9.39}$$

where (5.49) with (9.4) implies:

$$\tilde{\Sigma}^2(E) = R\log(2)(1/\mathrm{snr} + BE) \tag{9.40}$$

We define $\tilde{E} := BE$. Let us now look at the extrema of this potential. We see that we have to distinghish between the high error case ($\tilde{\Sigma} > 1/\sqrt{2}$ so that $\tilde{E} > 1/(2R\log(2)) - 1/\mathrm{snr}$) and the low error one ($\tilde{\Sigma} < 1/\sqrt{2}$, so that $\tilde{E} < 1/(2R\log(2)) - 1/\mathrm{snr}$).

In the high error case, the derivative $\partial_{\tilde{E}}\phi(\tilde{E})$ of the potential is zero when:

$$\frac{1}{2R\log(2)}\left(\frac{1}{1/\text{snr}+\tilde{E}}-\frac{1/\text{snr}+1}{(1/\text{snr}+\tilde{E})^2}\right)=0 \tag{9.41}$$

which happens when $\tilde{E}=1$. Therefore, if both the condition $\tilde{E}=1$ and $\tilde{E}>1/(2R\log(2))-1/\text{snr}$ are met, there is a stable extremum of the replica potential at $\tilde{E}=1$. The existence of this high-error extremum thus requires $1/(2R\log(2))-1/\text{snr}<1$, and we thus define the critical rate beyond which the state at $\tilde{E}=1$ is stable:

$$R_{BP}^{\infty}(\text{snr}):=[(1/\text{snr}+1)2\log(2)]^{-1} \tag{9.42}$$

Since we initilize the recursion at $\tilde{E}=1$ when we attempt to reconstruct the signal with AMP, we see that $R_{BP}^{\infty}(\text{snr})$ is a crucial limit for the reconstruction ability by message-passing.

In the low error case, the derivative of the potential is zero when:

$$\frac{1}{2R\log(2)}\left(\frac{1}{1/\text{snr}+\tilde{E}}-\frac{1/\text{snr}+1}{(1/\text{snr}+\tilde{E})^2}\right)=-\frac{1}{2R\log(2)}\frac{1}{(1/\text{snr}+\tilde{E})^2} \tag{9.43}$$

which happens when $\tilde{E}=0$. Hence, there is another extremum with zero error. Let us determine which of these two is dominant. We have:

$$\phi(\tilde{E}=0)=-\frac{1}{2R\log(2)}\big(\log(1/\text{snr})+\text{snr}\big)+\frac{\text{snr}}{2R\log(2)} \tag{9.44}$$

$$=\frac{\log_2(\text{snr})}{2R} \tag{9.45}$$

$$\phi(\tilde{E}=1)=-\frac{\log_2(1/\text{snr}+1)}{2R}+1 \tag{9.46}$$

The perfect reconstruction extremum is dominant as long as:

$$\log_2(\text{snr})>2R-\log_2(1+1/\text{snr}) \tag{9.47}$$

or equivalently when:

$$R<\frac{1}{2}\log_2(1+\text{snr})=C \tag{9.48}$$

where we recognize the expression of the Shannon capacity (3.91). As discussed in sec. 5.1.1, the optimal transition of the code is defined as the rate where the two maxima have same height, i.e. at $R=C$: these results are thus confirming that, at large value of B, the correct Bayes optimal value of the section error rate tends to zero and to a perfect reconstruction, at least as long as the rate remains below the Shannon capacity after which, of course, this could not be true anymore. This confirms, using the replica method, the results by [142, 143] that these codes are capacity achieving.

9.6.2 Alternative derivation of the large section limit via the replica method

We now re-derive the results of the previous section, that the superposition codes are capacity achieving, using the replica method to compute $I := \lim_{B\to\infty} I_B$ that appears in (9.21). The computation is performed at fixed $\tilde{\Sigma}$ which plays again the role of a temperature. The replica method is appropriate because we have to average the logarithm of the partition function (9.23) over the disorder $\mathbb{E}_{\mathbf{z}}\{\log(K_B(\mathbf{z}))\}$, here the Gaussian i.i.d vector $\mathbf{z}$. Starting from (9.23), we can re-write K_B as:

$$K_B(\mathbf{z}) := \exp\left(\frac{\log(B)}{2\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_1}{\tilde{\Sigma}}\right) + \sum_{i=2}^{B}\exp\left(-\frac{\log(B)}{2\tilde{\Sigma}^2} + \frac{\sqrt{\log(B)}z_i}{\tilde{\Sigma}}\right) \tag{9.49}$$

$$= \sum_{i}^{B}\exp\left(-\frac{1}{\tilde{\Sigma}}\left(\frac{\log(B)}{2\tilde{\Sigma}}(1-2\delta_{i,1}) - \sqrt{\log(B)}z_i\right)\right) \tag{9.50}$$

$$=: \sum_{i}^{B}\exp\left(-\frac{h_i(z_i)}{\tilde{\Sigma}}\right) \tag{9.51}$$

Meanwhile Z given by (5.2) is the full (random) partition function of the overall signal, K_B can be interpreted as the partition function of one single section of size B. An important difference with the random energy model (9.27) of the previous section is that here there is a favored section state distinct from the other ones (noted state 1), corresponding to the actual state of the section in the original signal. It has been treated apart in the previous section but we kept it here in the "energy states" $\{h_i\}_i^B$. From the statistics of z_i we get the one of h_i:

$$z_i \sim \mathcal{N}(z_i|0,1) \tag{9.52}$$

$$\Rightarrow h_i \sim \mathcal{N}\left(h_i\middle|\frac{(1-2\delta_{i,1})\log(B)}{2\tilde{\Sigma}}, \log(B)\right) \tag{9.53}$$

The average of K_B with respect to $\mathbf{z}$ can thus be replaced by the average over $\mathbf{h}$, the vector of independent energy states (independent because the $\{z_i\}_i^B$ are). We use again the replica trick for computing $I_B = \mathbb{E}_{\mathbf{h}}\{\log(K_B(\mathbf{h}))\}$ as B diverges. We thus need the average replicated partition function as in the section sec. 5.2:

$$I := \lim_{B\to\infty}\mathbb{E}_{\mathbf{h}}\{\log K_B(\mathbf{h})\} \tag{9.54}$$

$$= \lim_{B\to\infty}\lim_{n\to 0}\frac{\mathbb{E}_{\mathbf{h}}\{K_B^n\}-1}{n} \tag{9.55}$$

$$\mathbb{E}_{\mathbf{h}}\{K_B^n\} = \mathbb{E}_{\mathbf{h}}\left\{\sum_{i_1,..,i_n}^{B,..,B}\exp\left(-\frac{1}{\tilde{\Sigma}}(h_{i_1}+\ldots+h_{i_n})\right)\right\} \tag{9.56}$$

$$= \mathbb{E}_{\mathbf{h}}\left\{\sum_{i_1,..,i_n}^{B,..,B}\prod_{j}^{B}\exp\left(-\frac{h_j}{\tilde{\Sigma}}\sum_{a}^{n}\delta_{j,i_a}\right)\right\} \tag{9.57}$$

$$= \sum_{i_1,..,i_n}^{B,..,B} \prod_j^B \mathbb{E}_{h_j} \left\{ \exp\left(-\frac{h_j}{\tilde{\Sigma}} \sum_a^n \delta_{j,i_a} \right) \right\} \tag{9.58}$$

$$= \sum_{i_1,..,i_n}^{B,..,B} \exp\left(\frac{\log(B)}{2\tilde{\Sigma}^2} \sum_j^B \left(\sum_{a,b}^{n,n} \delta_{j,i_a}\delta_{j,i_b} - (1-2\delta_{j,1}) \sum_a^n \delta_{j,i_a} \right) \right) \tag{9.59}$$

$$= \sum_{i_1,..,i_n}^{B,..,B} \exp\left(\frac{\log(B)}{2\tilde{\Sigma}^2} \left(\sum_{a,b}^{n,n} \delta_{i_a,i_b} - \sum_j^B \sum_a^n \delta_{j,i_a}(1-2\delta_{j,1}) \right) \right) \tag{9.60}$$

$$= \sum_{i_1,..,i_n}^{B,..,B} \exp\left(\frac{\log(B)}{2\tilde{\Sigma}^2} \left(\sum_{a,b}^{n,n} \delta_{i_a,i_b} - n + 2\sum_a^n \delta_{1,i_a} \right) \right) \tag{9.61}$$

We now define new macroscopic order parameters:

$$q_{ab} := \delta_{i_a,i_b} \ \forall \ (a,b) \tag{9.62}$$

$$m_a := \delta_{i_a,1} \ \forall \ a \tag{9.63}$$

The first one indicates if two replicas are in the same state or not, the second one if a given replica is in the favored state 1. As in sec. 5.2, we replace the microscopic sums over the single replica states by sums over the macroscopic replica order parameters (9.62), (9.63) which become the new variables. The definitions of these must be fulfilled so the sums are restricted over the subspace matching the order parameters definitions (9.62), (9.63). In the sec. 5.2, this condition was enforced by the introduction of Dirac delta functions in the integral through (5.30), here it is simpler because we are in a discrete case. We deduce from (9.61):

$$\mathbb{E}_{\mathbf{h}}\{K_B^n\} = \sum_{\mathbf{q},\mathbf{m}} \exp\left(\frac{\log(B)}{2\tilde{\Sigma}^2} \left(\sum_{a,b}^{n,n} q_{ab} + 2\sum_a^n m_a - n + 2\tilde{\Sigma}^2 s_{\mathbf{q},\mathbf{m}} \right) \right) \tag{9.64}$$

where we have introduced the entropy associated to these new order parameters: $s_{\mathbf{q},\mathbf{m}} := S_{\mathbf{q},\mathbf{m}}/\log(B)$ where $S_{\mathbf{q},\mathbf{m}}$ is the logarithm of the number of microscopic configurations (or states) of the replicas compatible with $\mathbf{q}$ and $\mathbf{m}$ definitions at the same time, where $\mathbf{q} := \left[q_{ab}\right]_{a,b}^{n,n}$ and $\mathbf{m} := [m_a]_a^n$. We use as before the replica symmetric ansatz where each replica is considered equivalent:

$$q_{ab} = q + (1-q)\delta_{a,b} \ \forall \ (a,b) \tag{9.65}$$

$$m_a = m \ \forall \ a \tag{9.66}$$

It allows to simplify the average replicated partition function:

$$\mathbb{E}_{\mathbf{h}}\{K_B^n\} = \sum_{q,m} \exp\left(n\log(B) \underbrace{\left[\frac{(n-1)q + 2m + \frac{2\tilde{\Sigma}^2}{n} s_{q,m}}{2\tilde{\Sigma}^2} \right]}_{:=\tilde{I}(q,m)} \right) \tag{9.67}$$

$$=: \sum_{q,m} \exp\left(n\log(B)\tilde{I}(q,m) \right) \tag{9.68}$$

Looking at (9.62), (9.63), there are a priori four different possible ansatz, corresponding to four different macroscopic states of the section: $(q = m = 0), (q = m = 1), (q = 0, m = 1)$ and $(q = 1, m = 0)$ but actually, only three possibilities remain as the state $(q = 0, m = 1)$ has no meaning: the replicas cannot be all in different states $(q = 0)$ and all in the favored one $(m = 1)$ at the same time. Thus it remains:

- $(q = m = 0)$: all the replicas are in different states but none of them are in the favored one 1.
- $(q = m = 1)$: all the replicas are in the favored state 1.
- $(q = 1, m = 0)$: all the replicas are in the same state, which is not the favored one.

The last ansatz can be forgotten as the computation shows that it always leads to lower free entropy than the two other ansatz. This is understandable as there should be a symmetry among all the "wrong" states (different from 1) as none of them is special with respect to the other ones, so the replicated system should not choose a particular one spontaneously. It leaves two ansatz. The previous sum $\sum_{q,m}$ is performed by the saddle point method as $B \to \infty$, assuming as previously (see sec. 5.2) the commutativity of the limits in (9.55). From (9.55), the "section free entropy density" $I/\log(B)$ associated to each state is thus $\tilde{I}(q^*, m^*)$ where (q^*, m^*) is choosen among the two possible ansatz. The integral I is thus:

$$I = \log(B) \max_{(q^*,m^*)} \left[\tilde{I}(q^*, m^*)\right] \tag{9.69}$$

Let's compute the value of $\tilde{I}$ for the two remaining ansatz as $n \to 0$ in order to find the maximum:

$$(q^* = m^* = 0) \Rightarrow s_{0,0} = \log((B-1)^n)/\log(B) \approx n \tag{9.70}$$

$$\Rightarrow \tilde{I}(E|q^* = m^* = 0) \approx 1 \tag{9.71}$$

$$(q^* = m^* = 1) \Rightarrow s_{1,1} = \log(1)/\log(B) = 0 \tag{9.72}$$

$$\Rightarrow \tilde{I}(E|q^* = m^* = 1) = (2\tilde{\Sigma}(E)^2)^{-1} \tag{9.73}$$

where $\tilde{\Sigma}(E)^2$ is given by (9.40). From these results combined with (9.21), defining $\tilde{E} = BE$, the rescaled potential $\phi(\tilde{E})$ and the function $g(\tilde{E})$ as:

$$g(\tilde{E}) := -\frac{1}{2R\log(2)}\left(\log(1/\text{snr} + \tilde{E}) + \frac{1-\tilde{E}}{1/\text{snr} + \tilde{E}}\right) \tag{9.74}$$

$$\phi(\tilde{E}|q^*, m^*) := \lim_{B\to\infty} \frac{\Phi_B(\tilde{E}|q^*, m^*)}{\log(B)} \tag{9.75}$$

$$= g(\tilde{E}) + \tilde{I}(\tilde{E}|q^*, m^*) \tag{9.76}$$

we get the final potential of the sparse superposition codes in the large section size limit:

$$\phi_0(\tilde{E}) := \phi(\tilde{E}|q^* = m^* = 0) = g(\tilde{E}) + 1 \tag{9.77}$$

$$\phi_1(\tilde{E}) := \phi(\tilde{E}|q^* = m^* = 1) = g(\tilde{E}) + \big(2\log(2)R(1/\text{snr} + \tilde{E})\big)^{-1} \tag{9.78}$$

$$\phi(\tilde{E}) = \max\big(\phi_0(\tilde{E}), \phi_1(\tilde{E})\big) \tag{9.79}$$

where the actual potential $\phi(\tilde{E})$ for a given error $\tilde{E}$, rate R and snr is the maximum of the two ansatz-dependent potentials. These two potentials give the statistical weight of two different regimes (or pure states) [24, 81] that have respectively a probability $\propto \exp(L\log(B)\phi_0(\tilde{E}))$ and $\propto \exp(L\log(B)\phi_1(\tilde{E}))$.

The belief propagation transition

The BP transition R_{BP} is defined as the rate until which AMP without spatial-coupling or power allocation is Bayes optimal, see the green curve on Fig. 9.9. It corresponds to the lowest rate for which there exist two maxima of the Bethe free entropy. So to find it we equate the free entropies of the two pure states (9.77), (9.78):

$$\phi_0(\tilde{E}) = \phi_1(\tilde{E}) \tag{9.80}$$

$$\Rightarrow R_c(\tilde{E}) = [(1/\text{snr} + \tilde{E})2\log(2)]^{-1} \tag{9.81}$$

$R_c(\tilde{E})$ is the critical line until which it exists only one maximum and thus one state, or equivalently where appears the second maximum. Above it $R > R_c$ (but before the static transition) there are two distincts maxima. As in the previous section, a particular role is played by the value $\tilde{E} = 1$ in which AMP is initialized on real reconstructions. So $R_c(\tilde{E} = 1)$ gives the asymptotic $B \to \infty$ BP transition:

$$R_{BP}^{\infty}(\text{snr}) := R_c(1) = [(1/\text{snr} + 1)2\log(2)]^{-1} \tag{9.82}$$

we find back (9.42). Above $R > R_{BP}^{\infty}(\text{snr})$, we are in the hard or impossible phase, below it is the easy regime. The formula (9.81) can be interpreted the other way around: we consider a practical situation where the rate is fixed above the critical one $R > R_{BP}^{\infty}$ (at fixed snr) and we are in the hard phase such that there are two maxima. From (9.81) we can define the critical $\tilde{E}_c(R) = \max\big([2R\log(2)]^{-1} - 1/\text{snr}, 0\big)$, where the free entropy expression changes for a given rate:

$$\tilde{E} < \tilde{E}_c \Rightarrow \phi_1(\tilde{E}) > \phi_0(\tilde{E}) \tag{9.83}$$

$$\tilde{E} > \tilde{E}_c \Rightarrow \phi_1(\tilde{E}) < \phi_0(\tilde{E}) \tag{9.84}$$

Thus we can write the potential at fixed rate as:

$$\phi(\tilde{E}) = \phi_0(\tilde{E})\mathbb{I}\big(\tilde{E} > \tilde{E}_c\big) + \phi_1(\tilde{E})\mathbb{I}\big(\tilde{E} < \tilde{E}_c\big) \tag{9.85}$$

From the fixed point equations of the potential:

$$\tilde{E} < \tilde{E}_c \Rightarrow \frac{\partial \phi_1(\tilde{E})}{\partial \tilde{E}} = 0 \Rightarrow \tilde{E} = 0 \tag{9.86}$$

$$\tilde{E} > \tilde{E}_c \Rightarrow \frac{\partial \phi_0(\tilde{E})}{\partial \tilde{E}} = 0 \Rightarrow \tilde{E} = 1 \tag{9.87}$$

we see that in the hard phase, it coexists asymptotically two maxima such that the first corresponds to a perfect reconstruction, the other to the metastable failure state: we can identify $\phi_1(\tilde{E})$ as the free entropy corresponding to the perfect reconstruction state, $\phi_0(\tilde{E})$ to the failure one. But when does the hard phase stop and the impossible one starts, i.e. where is the optimal transition?

The optimal transition

The Bayes optimal rate is defined as the rate where the two distinct maxima have same height, which means they have the same statistical weight:

$$\phi(\tilde{E} = 0) = \phi(\tilde{E} = 1) \tag{9.88}$$

$$\Rightarrow \phi_1(\tilde{E} = 0) = \phi_0(\tilde{E} = 1) \tag{9.89}$$

$$\Rightarrow \log(1/\text{snr}) = \log(1/\text{snr} + 1) - 2R_{opt}\log(2) \tag{9.90}$$

$$\Rightarrow R_{opt} = \frac{1}{2}\log_2(1 + \text{snr}) = C \tag{9.91}$$

where we recognize the Shannon capacity (3.91). The optimal transition of the superposition codes scheme is thus asymptotically given by the capacity and between R_{BP}^{∞} and C is the hard phase.

9.6.3 Results from the replica analysis

From this analysis, we can extract the phase diagram of the superposition codes scheme. Fig. 9.10 presents phase diagrams for different snr values. The blue curve is the BP transition extracted from the potential (9.21) which marks the end of optimality of the AMP decoder without spatial coupling or proper power allocation, while the red curve is the optimal transition: the highest rate until decoding is theoretically possible. The black dashed curve is the asymptotic BP transition (9.42).

A first observation is that the BP transition is converging quite slowly to its asymptotic value compared to the optimal one that converges faster to the capacity, which is good. We also note that the section size where start the transitions (and thus marks the appearance of the hard regime where the AMP decoder without spatial coupling is not Bayes optimal anymore) gets larger as the snr decreases. When the snr is not too large, we see that the optimal and BP transitions almost coincide at small B values, such as for $B = 16$ at snr = 7 and $B = 4$ for

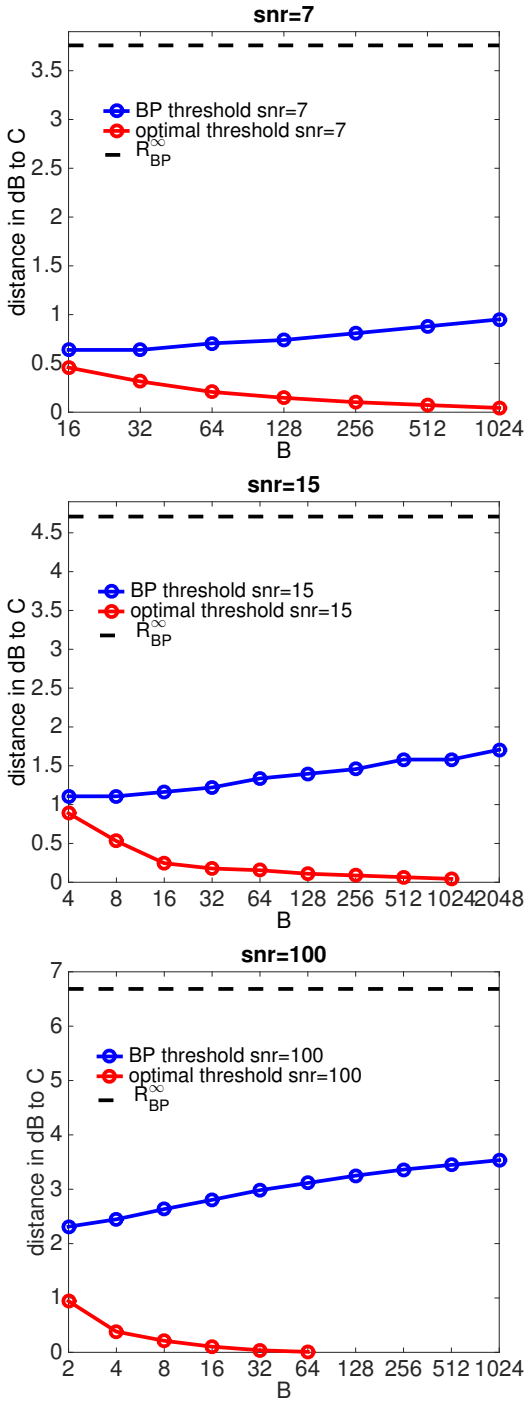

Figure 9.10 – All the points are computed from the Bethe free entropy (9.21) where the integral is computed by monte carlo. These are the phase diagrams of the superposition codes for different snr, where the x axis is the section size B, the y axis is the distance to the capacity C in dB. The blue and red curves are respectively the BP and optimal transitions. The black dashed line is the asymptotic value $B \to \infty$ of the BP threshold $R^{\infty}_{BP}(\text{snr})$ (9.42).

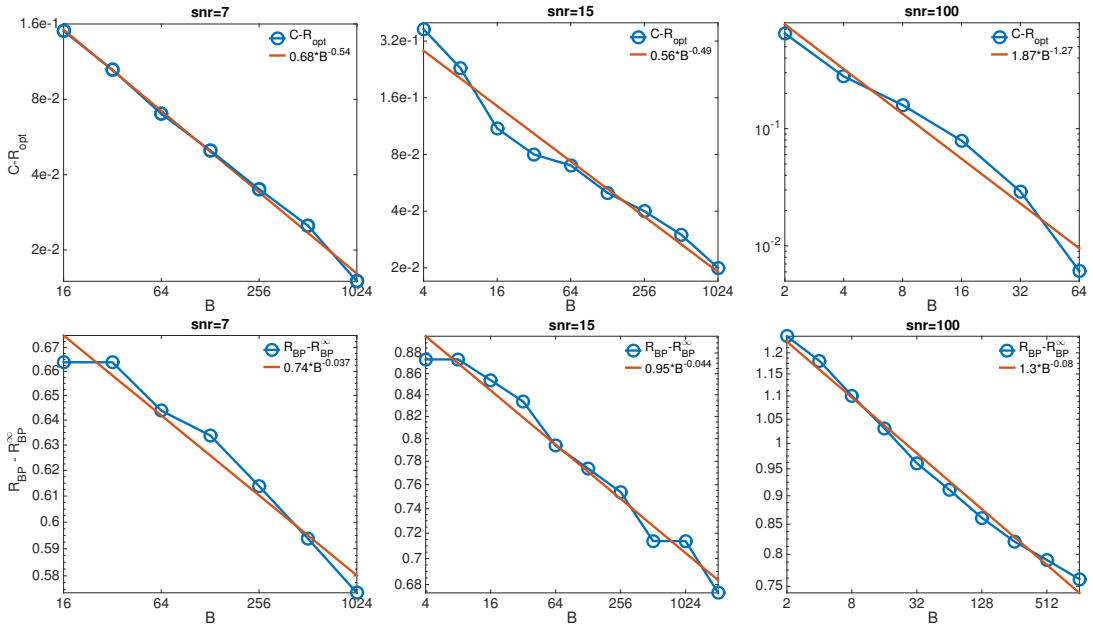

Figure 9.11 – All the blue points are computed by replica method from the potential (9.21) where the integral is computed by monte carlo. **Upper plots :** These plots show how fast the optimal transition $R_{opt}(B|\text{snr})$ is approaching the capacity. We plot the difference $C(\text{snr}) - R_{opt}(B|\text{snr})$ as a function of B in double logarithmic scale. The lines are guides for the eyes, and should not be taken as serious fits. They strongly suggest, however, a power law behavior. **Lower plots :** We did the same for $R_{BP}(B|\text{snr})$. We plot the difference $R_{BP}(B|\text{snr}) - R_{BP}^{\infty}(\text{snr})$ as a function of B in double logarithmic scale, where $R_{BP}^{\infty}(\text{snr})$ is the asymptotic BP transition computed by replica method, see (9.42). In every cases, we observe a behavior quite well predicted by a power law as well. The lines are again guides for the eyes, and the very low values of the exponents suggest a very slow logarithmic behavior.

snr = 15. Below this section size value, there are no more sharp phase transitions as only one maximum exists in the potential (9.21) and the dedoder, even without spatial coupling, is optimal at any rate: the SER increases continuously with the rate. As the snr increases, the curves split sooner until they remain different $\forall$ B such as in the snr = 100 case. See Fig. 9.12 and Fig. 9.13 for more details on the achievable values of the SER.

Fig. 9.11 gives details on the scalings of the convergence of the first order transitions to their asymptotic values. It seems that the rate of convergence of both transitions can be well approximated by power laws. On Fig. 9.11 we present the differences between the points of Fig. 9.10 and their asymptotic $B \to \infty$ values which are the capacity C for the optimal transition (as shown in the two previous sections) and B_{BP}^{∞} (9.42) for the BP transition. It appears that the scaling exponents amplitude tend to increase with the snr.

Fig. 9.12 represents how the optimal SER evolves at fixed rate $R = 1.3$ and snr = 15 as a function of the section size B (upper plot) and at fixed rate $R = 1.3$ and $B = 2$ as a function of the snr (lower plot). The observations are similar: in both cases, the curves seems to be well

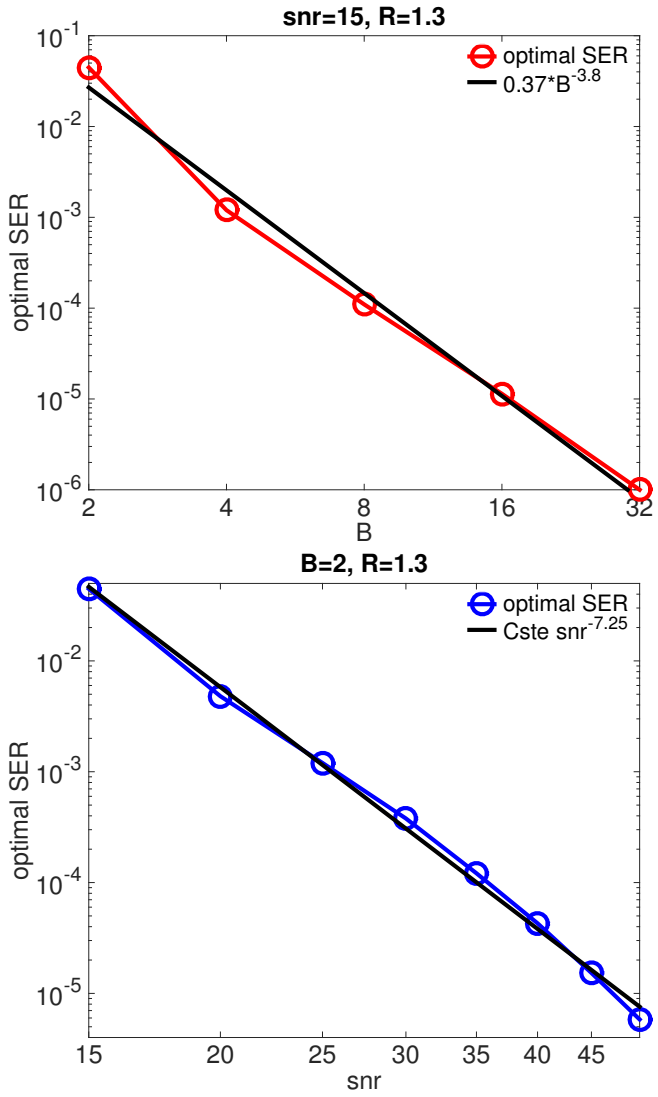

Figure 9.12 – On these plots we show how the optimal SER changes when B or the snr increase according to the replica theory (9.21). Both curves are plots at fixed rate $R = 1.3$ and are in double logarithmic scale. The red curve is function of B at fixed snr = 15, the blue one at fixed $B = 2$ as a function of the snr. The best linear fit is on top of the curves (Cste is a constant).

approximated by power laws with exponents given on the plots. The points are extracted from the replica potential (9.21).

Fig. 9.13 quantifies the optimal performances asymptotically attainable by the decoder, obtained from the state evolution analysis by initializing the recursion (9.11) at $E^{t=0} = 0$ and using (9.14). We plot the base 10 logarithm of the SER corresponding to the lower SER maximum of the potential (9.21) as a function of the rate and the section size B (again the state evolution and replica analyzes are equivalent to determine the fixed points as shown in 5.4). In high noise regimes, the depicted SER is always reachable by AMP without the need of spatial

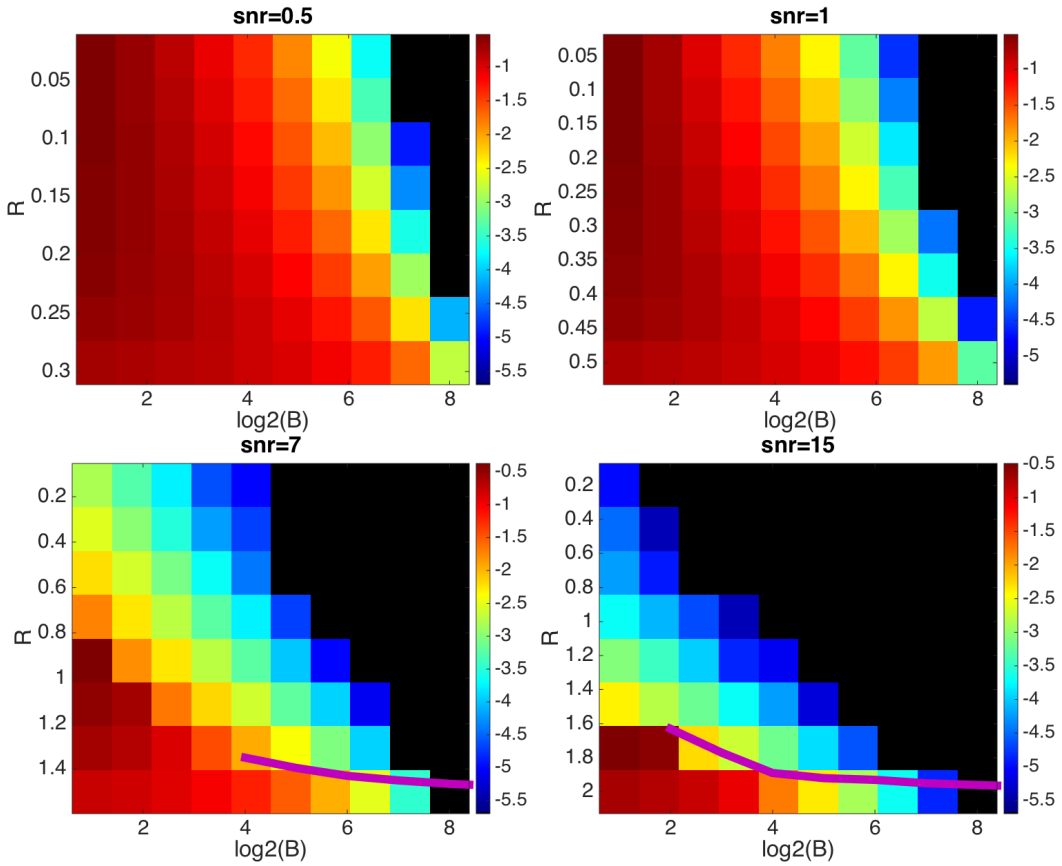

Figure 9.13 – On this plot, we show the logarithm in base 10 of the section error rate corresponding to the lowest SER maximum of the replica potential (9.21) in the (R,B) plane for different snr values. The values are obtained from the state evolution recursion (9.14) starting from the solution (i.e. with an initial error equal to 0). The recursions (9.11), (9.14) are computed by monte carlo with a sample size of $5B \times 10^5$. The black squares correspond to points where the computed value is $SER = 0$ which actually means a value that is lower to $(5 \times 10^5)^{-1}$ with high probability. The solid pink curve on the two lower plots correspond to the optimal rates $R_{opt}(B, \text{snr})$ as in Fig. 9.10. In the two upper plots that correspond to high noise regimes, there is no transition at all (the AMP decoder is thus always Bayes optimal, at least for these manageable section sizes B) and the optimal SER is a smooth increasing function of the rate R at fixed B and decreasing function of B at fixed R. The SER values in the two lower plots corresponding to low noise regimes matche the optimal SER as long as it is for a rate $R < R_{opt}(B, \text{snr})$ lower than the optimal one. For higher rates, the maximum of the potential corresponding to the plotted SER values is not the global maximum and thus cannot be reached, even with spatial coupling that works asymptotically until the optimal rate. For B smaller than 4 (resp. 2) on the snr = 7 (resp. snr = 15) plot, there is no sharp transition and the represented SER value is the optimal one and can be reached by AMP without spatial coupling, as in the high noise regime.

coupling as there are no sharp transitions. For lower noise regimes, the plotted SER matches the optimal one as long as $R < R_{opt}(B, \text{snr})$ where the optimal transitions correspond to the

solid pink curves. When there is no optimal transition (for a B before that the pink line starts), the SER is the optimal one and AMP can always reach it. The upper plot of Fig. 9.12 is a cut in the snr = 15 plot.

9.7 Optimality of the approximate message-passing decoder with a proper power allocation

In this section, we shall discuss a particular power allocation that allows AMP to be capacity achieving in the large size $L \gg 1$ limit, without the need for spatial coupling. We shall work again in the large section size $B \gg 1$ limit as well.

We first divide the system into G groups, see Fig. 9.8. For our analysis, each of these groups has to be large enough and must contain many sections, each of these sections being itself large so that $1 \ll B \ll L_G$, $1 \ll G \ll L$ where $L_G := L/G$ is the number of sections per group. Now, in each of these groups, we use a different power allocation: the non zero values of the sections inside the group g are all equal to c_g. This is precisly the case which we have studied in Sec. 9.5.1, so we can apply the corresponding state evolution in a straightforward manner.

Our claim is that we can use the following power allocation:

$$c_g = \frac{2^{-\frac{Cg}{G}}}{Z} \ \forall \ g \in \{1, \dots, G\} \tag{9.92}$$

where $C = \frac{1}{2}\log_2(1+\text{snr})$ is the Shannon capacity. We choose Z such that the power of the signal equals one:

$$\frac{1}{G}\sum_{g}^{G} c_g^2 = 1 \tag{9.93}$$

With this definition, we have:

$$Z^2 = \frac{2^{-\frac{2C}{G}}\left(1-2^{-2C}\right)}{G(1-2^{-\frac{2C}{G}})} \tag{9.94}$$

It will be useful to know the following identity:

$$\frac{1}{G}\sum_{g}^{\tilde{g}} c_g^2 = \frac{1-2^{-\frac{2C\tilde{g}}{G}}}{1-2^{-2C}} \tag{9.95}$$

Now, we want to show that, if we have decoded all the sections before the section $\tilde{g}$ at time t, then we will be able to decode section $\tilde{g}$ as well. If we can show this, then starting from $\tilde{g} = 1$ we will have a succession of decoding until all is decoded, and we would have shown that this power allocation works. In this situation, using (9.20) and the expression of the rate R (9.3), we

have for the section $\tilde{g}$:

$$(\tilde{\Sigma}_{\tilde{g}}^{t+1})^2 = R\log(2)\left[\frac{1/\mathrm{snr}+\mathscr{E}_{\tilde{g}-1}}{c_{\tilde{g}}^2}\right] \tag{9.96}$$

with:

$$\mathscr{E}_{\tilde{g}} := 1-\frac{1}{G}\sum_{g}^{\tilde{g}} c_g^2 \tag{9.97}$$

where we have used (9.93) with our assumption of having already decoded until $\tilde{g}-1$ included at time t: $\tilde{E}_g^t = BE_g^t = \mathbb{I}(g \geq \tilde{g})$. (9.97) is the average (rescaled by B) MSE if all has been decoded until $\tilde{g}$ included: it is given by the initial total rescaled MSE $\tilde{E}^{t=0} = 1$ from which we have to remove what has been already decoded. We now ask if the group $\tilde{g}$ can be decoded as well. The evolution of the error in this group is given by (9.19) and we have seen in sec. 9.6.1, that the condition for a perfect decoding in the large B limit is simply that $\tilde{\Sigma}_{\tilde{g}}^2 < 1/2$ which remains true per group as the only coupling with the other groups in the state evolution (9.19) is through the "temperature" $\tilde{\Sigma}_{\tilde{g}}^2$. We thus need the following to be satisfied (as long as $R < C$):

$$R\log(2)\left[\frac{1/\mathrm{snr}+\mathscr{E}_{\tilde{g}-1}}{c_{\tilde{g}}^2}\right] < \frac{1}{2} \tag{9.98}$$

If this condition is satisfied, there is no local BP transition to block the AMP reconstruction in the group $\tilde{g}$, then the decoder will move to the next group, etc. We thus need this condition to be correct $\forall\ \tilde{g} \in \{1,\dots,G\}$. Let us perform the large G limit (remembering that g/G stays however finite). Using (9.94) we have:

$$c_g^2 = \frac{2^{-\frac{2Cg}{G}}}{Z^2} \tag{9.99}$$

$$= \frac{G(1-2^{-\frac{2C}{G}})}{2^{-\frac{2C}{G}}\left(1-2^{-2C}\right)}2^{-\frac{2Cg}{G}} \tag{9.100}$$

$$= \frac{G(1-2^{-\frac{2C}{G}})}{\left(1-2^{-2C}\right)}2^{-\frac{2C(g-1)}{G}} \tag{9.101}$$

$$\approx G\frac{2^{-\frac{2C(g-1)}{G}}}{\left(1-2^{-2C}\right)}\left(\log(2)2C/G+O(1/G^2)\right) \tag{9.102}$$

$$\approx \frac{2C\log(2)\,2^{-\frac{2C(g-1)}{G}}}{1-2^{-2C}}+O(1/G) \tag{9.103}$$

Now, we note from (3.91) that the snr can be written as $\mathrm{snr} = 2^{2C}-1 = \frac{1-2^{-2C}}{2^{-2C}}$ so plugging (9.95)

inside (9.97) we have:

$$1/\mathrm{snr} + \mathscr{E}_{\tilde{g}-1} = \frac{2^{-2C}}{1-2^{-2C}} + 1 - \frac{1-2^{-\frac{2C(\tilde{g}-1)}{G}}}{1-2^{-2C}} \tag{9.104}$$

$$= \frac{2^{-\frac{2C(\tilde{g}-1)}{G}}}{1-2^{-2C}} \tag{9.105}$$

Therefore to leading order, we have using (9.103) that:

$$\frac{1/\mathrm{snr} + \mathscr{E}_{\tilde{g}-1}}{c_{\tilde{g}}^2} \approx \frac{1}{2C\log(2)} \tag{9.106}$$

so that the condition (9.98) becomes for large G:

$$\frac{R\log(2)}{2C\log(2)} = \frac{R}{2C} < \frac{1}{2} \tag{9.107}$$

or equivalently, $R < C$. This shows that, with proper power allocation and as long as $R < C$, a local minimum asymptotically cannot exist in the potential, or equivalently, that the AMP decoder cannot be stuck in such a spurious minimum: it will reach the solution with perfect reconstruction $SER = 0$.

9.8 Numerical experiments for finite size signals

We now present a number of numerical experiments testing the performance and behavior of the AMP decoder in different practical scenarios with finite size signals. The first experiment Fig. 9.14 quantifies the influence of the finite size effects over the superposition codes scheme with spatially-coupled Hadamard-based operators, decoded by AMP. For each plot, we fix the snr and the alphabet size B and repeat 10^4 decoding experiments per point with each time a different signal with constant power allocation and operator drawn from the ensemble $(L_c = 16, L_r = 17, w = 2, \sqrt{J} = 0.4, R, \beta_{seed} = 1.8)$. The curves present the empirical block error rate (blue and yellow curves) which is the fraction of instances that have not been perfectly decoded, i.e. such that the final $SER > 0$, and the average SER (red and purple curves). This is done for two different sizes $L = 2^8$ and $L = 2^{11}$. When the curves stop, it means that the empirical block error rate (and thus the section error rate as well) is actually 0. In reality it should reach a noise floor $< 10^{-4}$ but does not because of the same reasons explained in sec. 9.4. The dashed lines are the BP transition $R_{BP}(\mathrm{snr}, B)$ and optimal transition $R_{opt}(\mathrm{snr}, B)$ extracted respectively from the state evolution analysis and potential (9.21) and the solid black line is the capacity $C(\mathrm{snr})$. Thanks to the fact that at large enough section size B, the gap between the BP transition and capacity is consequent, it leaves room for the spatially-coupled AMP decoder to beat the transition, allowing to decode at $R > R_{BP}$ as in LDPC codes. For small section size B, the gap is too small to get real improvement over the full operators. We also note the previsible fact that as the signal size L increases, the results are improving: one can decode closer to the asymptotic transitions and reach a lower error floor. For $B = 256$, the

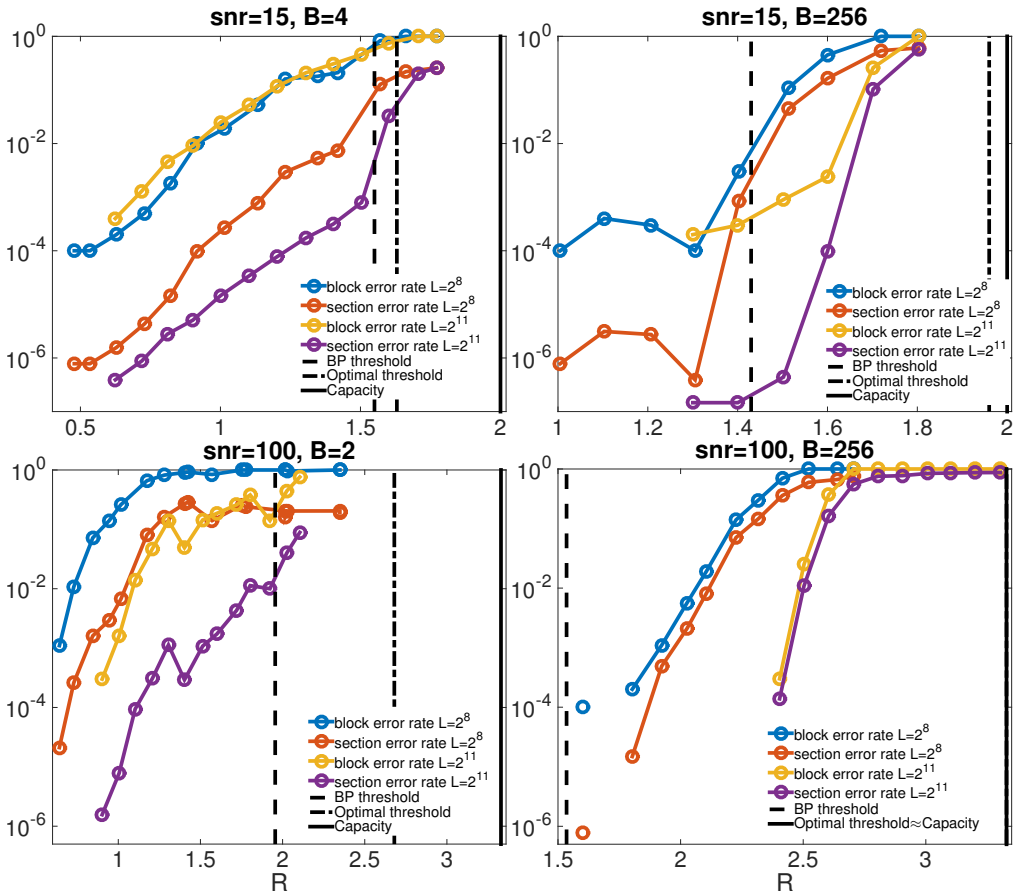

Figure 9.14 – On this plot, we show the empirical block error rate and average section error rate of the superposition codes using the AMP decoder combined with spatially-coupled Hadamard-based operators for two different snr, signal sizes L and section sizes B. The block error rate is the fraction of the 10^4 random instances we ran for each point that have not been perfectly reconstructed, i.e. in these instances at least one section has not been well recontructed and the final $SER > 0$. The SER has been averaged over the 10^4 random instances. The convergence criterion is that the mean change in the variables estimates between two consecutive iterations $\delta < 10^{-8}$ and the maximum number of iterations is $t_{max} = 3000$. The upper plots are for snr = 100, the lower for snr = 15 (notice the different x axes). The first dashed black line is the BP transition obtained by state evolution analysis, the second one is the optimal transition obtained by the replica method from the Bethe free entropy (9.21) and the solid black line is the capacity. In the (snr = 100, $B = 256$) case, the optimal transition is so close to the capacity that we plot a single line. For such sizes, the block error rate is 0 for rates lower than the lowest represented one. The spatially-coupled operators used for the experiments are drawn from the ensemble ($L_c = 16, L_r = 17, w = 2, \sqrt{J} = 0.4, R, \beta_{seed} = 1.8$).

sharp phase transition between the phases where decoding is possible and impossible by AMP with spatial coupling is clear and gets sharper as L increases.

The next experiment Fig. 9.15 is the phase diagram for superposition codes at fixed snr = 15

like on Fig. 9.10 but where we added on top finite size results. The asymptotic rates that can be reached are shown as a function of B (blue line for the BP transition, red one for the optimal rate). The solid black line is the capacity. Comparing the black and yellow curves, it is clear that even without spatial coupling or power allocation, AMP outperforms the iterative successive decoder of [142] for practical B values. With the Hadamard-based spatially-coupled AMP algorithm, this is true for any B and is even more pronounced (brown curve). The green (pink) curve shows that the homogeneous (spatially-coupled) Hadamard-based operator has very good performances for reasonably large signals, corresponding here to a blocklength $M < 64000$ (the blocklength is the size of the transmitted codeword $\tilde{\mathbf{y}}$).

Finally, the last experiment Fig. 9.16 is a comparison of the efficiency of the AMP decoder combined with spatial coupling or an optimized power allocation coming from [148]. We repeated their experiments and also compared the results to a spatial coupling strategy. Comparing the results with Hadamard-based operators, given by the red and yellow curves for power allocation and spatial coupling respectively, it is clear that spatial coupling (despite being not optimized at each rate as it is done for the power allocation) greatly outperforms an optimized power allocation scheme.

In addition, we see that our red curve corresponding to the optimized power allocation homogeneously outperforms the results of [148] with exactly the same parameters, given by the blue curve. As we have numerically shown that Hadamard-based operators gets same final performances as random ones as used in [148] (see chap. 7 and Fig. 9.4), the difference in performances must come from the AMP implementation: in our decoder that we denote by on-line decoder, there is no need of pre-processing computations but in the decoder of [148] denoted by off-line, quantities need to be computed in advance.

The advantage of spatial coupling over power allocation is independent of the decoder and the fact that we use Hadamard-based operators, as it outperforms the red curve as well which is also obtained with our on-line decoder and Hadamard-based operators. This is true at any rate except at very high values where spatial coupling does not decode at all, meanwhile the very first components of the signal are found by power allocation strategy as their power is very large. But it is not a really useful regime as only a small part of the signal is decoded anyway. The green points show that a mixed strategy of spatial coupling with optimized power allocation does not perform well compared to individual strategies. This is easily understood from the Fig. 9.8: a power allocation modifies the spatial coupling and worsen its original design. In addition we notice that at low rates, a power allocation strategy performs worst than constant power allocation as the very last components with very low power are rarely decoded.

9.9 Concluding remarks

We have fully derived and studied the approximate message-passing decoder, combined with spatial coupling or power allocation, for the sparse superposition error correction scheme

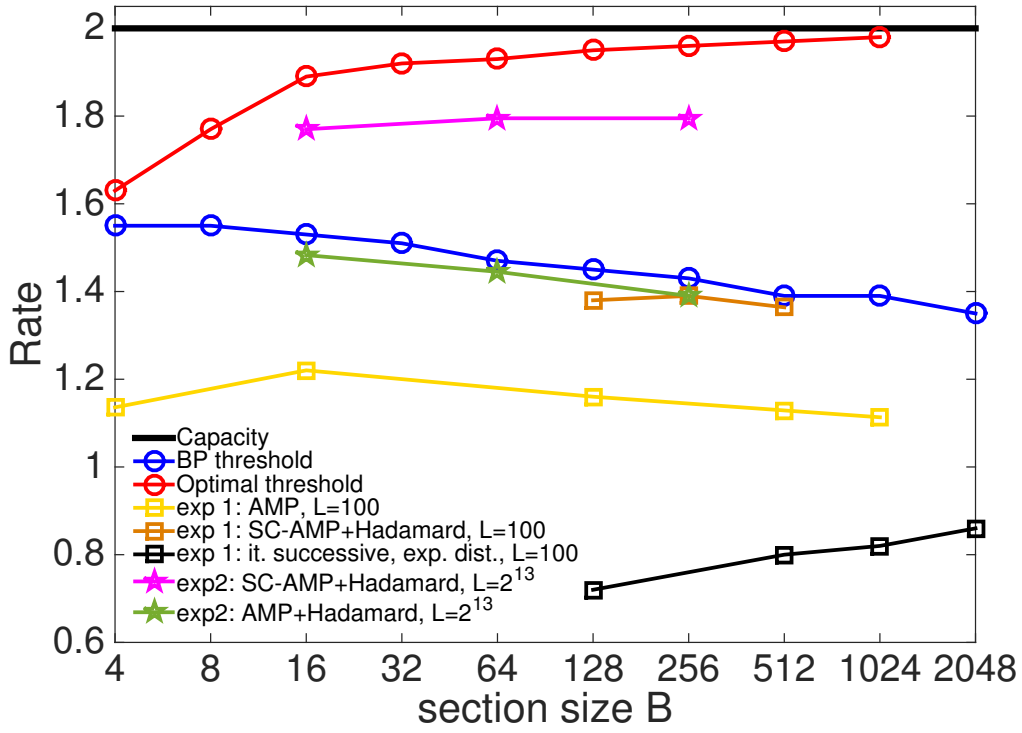

Figure 9.15 – Phase diagram and experimental results for superposition codes at finite size L for snr = 15 compared to the asymptotic results. The solid black line is the capacity which bounds the performance of any reconstruction algorithm for this snr, the blue line is the BP transition $R_{BP}(\text{snr}=15,B)$ obtained by state evolution analysis and the red line is the Bayesian optimal transition $R_{opt}(\text{snr}=15,B)$ obtained by from the potential (9.21). The yellow, black and brown curves are results of the following experiment (exp 1): decode 10^4 random instances and identify the empirical transition curve between a phase where the empirical probability $P\left(SER>10^{-1}\right)<10^{-3}$ (below the line) from a phase where $P\left(SER>10^{-1}\right)\geq 10^{-3}$ (more than 9 instances have failed over the 10^4 ones). The green and pink curves are the result of the second protocol (exp 2) which is a relaxed version of exp 1 with 10^2 random instances and $P\left(SER>10^{-1}\right)<10^{-1}$ below the line, $P\left(SER>10^{-1}\right)\geq 10^{-1}$ above. Note that in our experiments $SER<10^{-1}$ essentially means $SER=0$ at these sizes. The yellow curve compares our results with the iterative successive decoder (black curve) of [142, 143] where the number of sections $L=100$. Note that these data, taken from [142, 143], have been generated with an exponential power allocation rather than the constant one we used. Compared with the yellow curve (AMP with the same value of L) the better quality of AMP reconstruction is clear. The green and pink curves are here to show the efficiency of the Hadamard-based operators with AMP with (pink curve) or without (green curve) spatial coupling. For the experimental results, the maximum number of iterations of the algorithm is arbitrarily fixed to $t_{max}=500$. The parameters used for the spatially-coupled operators are $(L_r=16, L_c=17, w=2, \sqrt{J}=0.3, R, \beta_{seed}=1.2)$.

over the additive white Gaussian noise channel. Links have been established between the present problem and compressed sensing with structured sparsity.

On the theoretical side, we have computed the potential of the scheme thanks to the heuristic replica method and have shown that the scheme is capacity achieving in a proper limit. The analysis shows that there exist a sharp phase transition blocking the decoding by message-passing before the capacity but that the optimal Bayesian decoder obtained by combining message-passing to spatial coupling or power allocation can reach the capacity as the section size of the signal increases. We have also derived the state evolution recursions associated to the message-passing decoder, with or without spatial coupling and power allocation. The optimal performances have been studied and it appeared that the error decrease and the rates of convergence of the various transitions to their asymptotic values follow power laws.

On the more practical and experimental side, we have presented an efficient and capacity achieving solver based on spatially-coupled fast Hadamard-based operators. It allows to deal with very large instances and performs as well as random coding operators. Intensive numerical experiments have shown that a well designed spatial coupling performs way better than an optimized power allocation of the signal, both in terms of reconstruction error and robustness to noise. Finite size performances of the decoder under spatial coupling have been studied and it appeared that even for small signals, spatial coupling allows to obtain very good perfomances. In addition, we have shown that the message-passing decoder without spatial coupling beats the iterative successive decoder of Barron and Joseph for any manageable size and that its performances with spatial coupling are way better for any section size.

The scheme should be now compared in a more systematic way to other state-of-the-art error correction schemes over the additive white Gaussian noise channel. On the application side, from the structure of the reconstructed signal itself in superpostion codes, we can also interpret the problem as a structured group testing problem where one is looking for the only individual that has some property (for example infected) in each group, the sections of the signal.

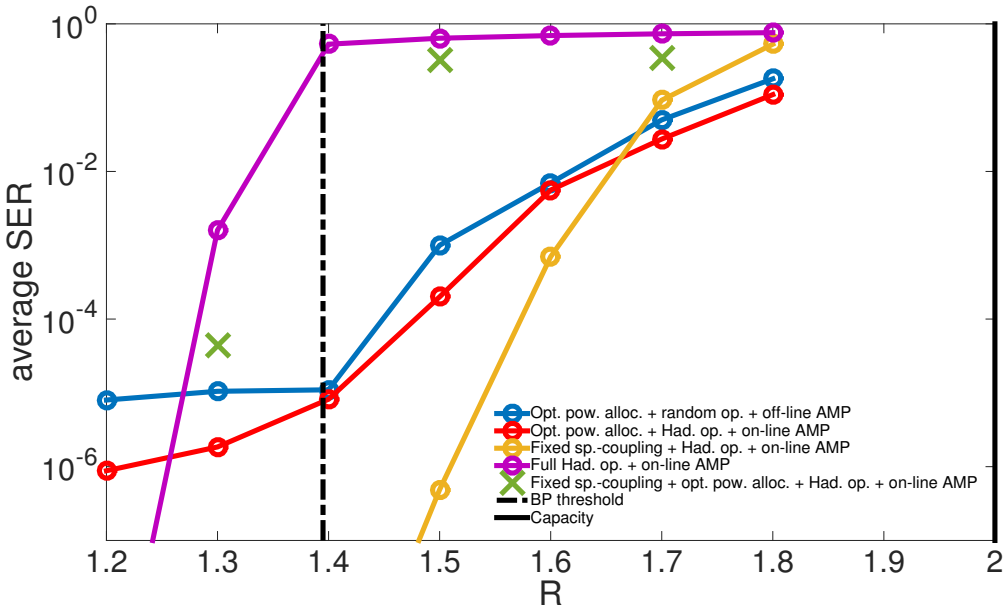

Figure 9.16 – The average section error rate SER in logarithmic scale as a function of the rate R for different settings, all at fixed (snr $= 15, B = 512, L = 1024$). The black dashed curve identifies the BP transition, the highest rate until which AMP can asymptotically perform well *without* spatial coupling or non constant power allocation, the black solid line is the Shannon capacity. **Blue curve :** It corresponds to the results of Fig. 3 of [148]: the points are averaged over 10^3 instances of random experiments using i.i.d Gaussian matrices and an optimized power allocation scheme where the parameters defining the power allocation are optimized for each rates. The values of the parameters and the associated power allocation scheme can be found in [148]. The denomination off-line AMP refers to the AMP decoder update rules of [148] that are different than ours and require an off-line pre-processing part as opposed to our procedure where all the quantities are computed on-line without any need of pre-processing. **Red curve :** We reproduced exactly the same experiment (with same power allocation scheme and parameters) as for the blue curve with two important differences: i) we used our on-line AMP decoder instead of their off-line implementation and ii) we used an Hadamard-based homogeneous operator instead of a random i.i.d Gaussian one. In addition, we runned 10^4 instances instead of 10^3 as we obtained an average SER equals to 0 for the two first points. **Purple curve :** This experiment with 10^4 instances per point is with an Hadamard-based homogeneous operator with on-line AMP decoding of constant power allocated signals. As it should, the decoder does not work anymore for $R > R_{BP}$. **Yellow curve :** The points of this experiment have been averaged over 10^4 instances. In this setting, we used our on-line AMP decoder and generated the signals with constant power allocation. We replaced the homogeneous operator by a spatially-coupled Hadamard-based operator, described in Fig. 7.1. The parameters defining the ensemble from which the operator is randomly generated are fixed once for all for the all experimental curve, as oposed to the power allocation curves where parameters have been optimized for each point. The ensemble is here given by $(L_c = 16, L_r = 17, w = 2, \sqrt{J} = 0.4, R, \beta_{seed} = 1.4)$. **Green crosses :** These points have been averaged over 10^4 instances. We used the same spatially-coupled Hadamard-based operator ensemble as for the yellow curve for decoding power allocated signals with same power allocation scheme as the blue and red curves. When the purple and yellow curves fall, it means that the points values are 0. The codeword size for all these curves is between 5×10^3 to 9×10^3.

10 Robust error correction for real-valued signals via message-passing decoding and spatial coupling

In the previous chapter we studied error correction over the additive white Gaussian noise channel. Some sparse discrete signal was encoded using a linear transform to get a real codeword sent through the channel. But let us imagine now that the noisy channel is even less reliable than the AWGN one as it adds gross Gaussian distributed errors in addition of the background AWGN. Is it possible anyway to send information reliably through such channel? The real transmitted signal in this new model can be interpreted as the codeword of another coding scheme for the AWGN. So if we manage to correct these gross errors, we come back to the original AWGN channel error correction problem and sparse superposition codes (or any other scheme for the AWGN) can be used afterwards.

As for the sparse superposition codes scheme, we will use the approximate message-passing algorithm as an efficient decoder. We will show that the error correction and its robustness towards noise can be enhanced considerably thanks to spatially-coupled coding matrices. We discuss the performances in the large signal limit using previous results on state evolution, as well as for finite size signals through numerical simulations. Even for relatively small sizes, the approach proposed here outperforms convex-relaxation-based decoders.

10.1 Introduction

Although information is discrete in the classical coding theory, there are situations of interest where one should consider real-valued signals, such as scrambling of discrete time analog signals for privacy [152], network [153, 154] or jointed source and channel coding [155], or in the impulse noise cancellation in orthogonal frequency division multiplexing systems [156]. We consider here a real channel model which adds gross errors on a fraction of elements and a small noise on all of them. The real signal could also be interpreted as the codeword of a previous error correction scheme, and the full error correction becomes the concatenation of two distinct schemes: the first for the background noise, on top of which we use the present

scheme to correct the gross errors.

To perform error correction for such real signals, a compressed sensing based scheme has been proposed by Donoho and Huo [157] and Candes and Tao [158]. Here we reconsider this problem taking full advantage of the approximate message-passing decoder and spatial coupling coding design.

The problem is easily stated. One is given a real-valued signal $\mathbf{s}$, and a channel that adds gross errors to a fraction of elements. Is there a way to encode the signal such that these errors can be corrected? Can this approach still be used when the channel is in addition adding a small AWGN to all elements (a situation arguably much closer to some real channels [157, 159, 160])? The method proposed in [157–159] is to first multiply the signal $\mathbf{s}$ by a random matrix in order to create a codeword of larger dimension, and then to use the classical compressed sensing approach, based on convex-relaxation decoding, in order to correct the errors of transmission.

In the present chapter, we replace the convex-relaxation decoding by the AMP decoder that uses the available prior information about the error statistical properties [34, 90] as opposed to convex optimization based solvers, see sec. 3.4.3. This provides a significant improvement in performances. Then we consider an approximately sparse channel where, in addition to the gross errors on a fraction of elements, there is a small AWGN over all components. We will show that the performances of the AMP decoder are stable under this additional noise, as already discussed in chap. 6. Finally we will use spatially-coupled measurement matrices in the decoding, which allow to further enhance the possibility for error correction (and up to its information theoretical limit in the case of strictly sparse noise).

10.2 Compressed sensing based error correction

Consider a real-valued vector of information $\mathbf{s} \in \mathbb{R}^N$, encode this vector by a full-rank real $M \times N$ matrix $\mathbf{A}$, with $\gamma := M/N > 1$ being the encoding rate (the redundancy or "over-sampling" introduced in the code), so that the codeword is $\mathbf{y} = \mathbf{A}\mathbf{s} \in \mathbb{R}^M$. The aim is to recover $\mathbf{s}$ lowering as much as possible the encoding rate. Since $\mathbf{A}$ is full rank, one can recover the original signal $\mathbf{s}$ from $\mathbf{y}$ multiplying it by the pseudo-inverse of $\mathbf{A}$. The codeword is then sent through a noisy channel and gives rise to the corrupted codeword $\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{e}$ where $\mathbf{e}$ is i.i.d with a distribution:

$$P(\mathbf{e}) = \prod_i^M \left[\rho \mathcal{N}(e_i|0, 1+\epsilon) + (1-\rho)\mathcal{N}(e_i|0,\epsilon)\right] \tag{10.1}$$

where $0 < \rho < 1$. So the noise distribution is of the approximate sparsity form studied in chap. 6. We thus have a fraction ρ of elements with gross (variance $1+\epsilon$) errors, the rest having small (variance ϵ) amplitudes. One then considers a full rank "parity-check"-like matrix $\mathbf{F}$ such that $\mathbf{F}\mathbf{A} = \mathbf{0}$. We construct such a pair of matrices by first choosing a $R \times M$ matrix $\mathbf{F}$ with i.i.d Gaussian distributed elements of zero mean and variance $1/M$ (or variance specified by the seeding matrix, see Fig. 5.4), the kernel of $\mathbf{F}$ is then the range of the encoding matrix $\mathbf{A}$. Note

that [158, 159] take **A** as the random matrix, but here we choose the opposite in order to be able to implement the spatially-coupled decoding. One must have $R \leq M - N$ for the couple $(\mathbf{F}, \mathbf{A})$ to exist, and in order to minimize the encoding rate γ, we take from now on $R := M - N$. The application of **F** to the corrupted codeword $\tilde{\mathbf{y}}$ results in the real-valued vector **h** given by:

$$\begin{aligned} \mathbf{h} &= \mathbf{F}(\mathbf{y} + \mathbf{e}) && (10.2)\\ &= \mathbf{F}(\mathbf{A}\mathbf{s} + \mathbf{e}) && (10.3)\\ &= \mathbf{F}\mathbf{e} && (10.4) \end{aligned}$$

where **h** has dimension R and **e** is an approximately sparse vector of dimension M. This is a compressed sensing problem (3.18) similar to what have been studied in chap. 6: reconstruct the approximately sparse M-d error vector **e** given R of its linear projections (measurements) **h**. In the context of compressed sensing **F** is the measurement matrix.

Let us first review the possibility of this error-correction scheme when the error **e** is exactly sparse, i.e. $\epsilon = 0$. Using an intractable ℓ_0 minimization, the gross error **e** in (10.4) can be found exactly as long as $R = M - N > M\rho$. So error correction in real-valued signals corrupted by strictly sparse gross noise is possible (but hard, see sec. 5.1.1) for encoding rates $\gamma > \gamma_{opt} = 1/(1-\rho)$. Popular tractable ℓ_1-minimization, as used in [158, 159], recovers the error **e** exactly when $M - N \geq \alpha_{DT} M$, where α_{DT} is the Donoho-Tanner measurement rate [161], see sec. 5.1.1, or equivalently when the encoding rate is larger than $\gamma \geq \gamma_{DT} = 1/(1-\alpha_{DT})$. These two transitions are depicted in Fig. 10.1 and one can see that γ_{DT} is considerably larger than γ_{opt}. A first step to improvement is to decode with an approximate message-passing approach.

10.2.1 Performance of the approximate message-passing decoder

As we fall exactly under the setting studied in the chap. 6, the posterior estimates of the noise components are directly given by the AMP decoder Fig. 5.5 with the operators definitions (5.116), (5.117), (5.118), (5.119) together with the denoisers (6.4), (6.5) where their parameters are given in the present case by:

$$\begin{aligned} w_1 &= \rho && (10.5)\\ \sigma_1^2 &= 1 + \epsilon && (10.6)\\ w_2 &= 1 - \rho && (10.7)\\ \sigma_2^2 &= \epsilon && (10.8) \end{aligned}$$

When parameters ρ, ϵ, γ are fixed whereas $M \to \infty$ and **F** is an homogeneous random i.i.d matrix, the evolution of the AMP algorithm can be described exactly using the state evolution given by (6.12) with initialization $E^{t=0} = \mathrm{Var}_{P_0}(e_i) = \rho + \epsilon$.

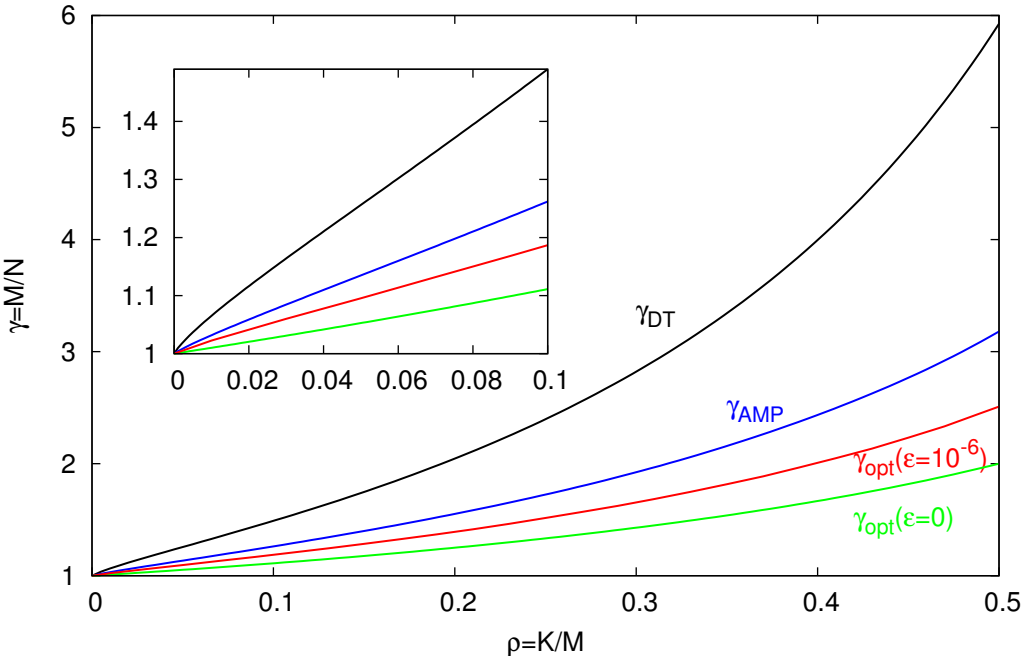

Figure 10.1 – Phase diagram showing the encoding rate $\gamma = M/N$ necessary to perform error correction over a channel with noise described by (10.1), plotted as a function of the noise sparsity ρ, zoom in the inset. The black (top) curve γ_{DT} depicts the limit of performance of the ℓ_1-minimization approach for $\epsilon = 0$, i.e. when the noise is strictly sparse. The blue (2nd from top) curve shows the limit of performance of the Bayesian approximate message-passing approach for $\epsilon = 0$. Note that up to about $\epsilon \lesssim 10^{-5}$ this curve does not change visibly, see Fig. 6.5. The green (bottom) curve, given by $\gamma_{opt} = 1/(1-\rho)$, depicts the lowest possible encoding rate for which exact decoding is possible for $\epsilon = 0$. Above the red (3rd from top) line, error correction with MSE comparable to $\epsilon = 10^{-6}$ is possible with the Bayes optimal estimation of the error vector. These two rates γ_{opt} can be reached in the limit of large signal size using the spatially-coupled Bayesian AMP approach.

State evolution for homogeneous matrices

The state evolution analysis of the Bayesian AMP for Gauss-Bernoulli noise **e** (that is, with $\epsilon = 0$) has been considered in great details in [34, 35]. In that case AMP reconstructs *perfectly* the solution in a region larger than the ℓ_1-minimization and up to the spinodal transition α_{BP}, see sec. 5.1.1. In the notation of the present problem, this leads exact decoding for considerably lower encoding rates: the resulting $\gamma_{BP} = 1/(1-\alpha_{BP})$ is shown in blue in Fig. 10.1 (where it is denoted γ_{AMP}). The advantage with respect to the ℓ_1-minimization decoding is clear. For a fraction of $\rho = 0.1$ of gross elements, for instance, the improvement goes from a necessary coding rate $\gamma_{DT} \approx 1.490$ for ℓ_1-minimization based decoders to $\gamma_{BP} \approx 1.262$ with AMP decoding with homogeneous matrices.

As already discussed in sec. 3.4.3, nevertheless, the ℓ_1 performance is independent of the distribution of the gross error, whereas the Bayesian AMP uses it. The properties of the channel are, however, often well known, in which case the improvement depicted if Fig. 10.1 is indeed achievable.

To assess how robust are these results towards approximately sparse noisy channels (nonzero value of ϵ in (10.1)) we use the state evolution analysis that was performed in chap. 6. It was shown that for about $\epsilon \lesssim 10^{-5}$ and $\alpha > \alpha_{BP}$ the AMP algorithm leads to reconstruction with MSE comparable to ϵ, see Fig. 6.4. This shows that the AMP approach is actually very robust to such noise.

State evolution for spatially-coupled matrices

Despite the advantage of the AMP-based decoding over the ℓ_1-minimization, it is still not asymptotically optimal since $\gamma_{BP} > \gamma_{opt}$, and one ideally aims to perform error correction with smallest possible encoding rates. In order to do so, we use a spatially-coupled measurement matrix **F**. In the present framework of error correction of real-valued signals, the spatial coupling can be implemented by first constructing the matrix **F** of the form Fig. 5.4, then determining the encoding matrix **A** as the null space of the matrix **F**.

The state evolution for these spatially-coupled matrices is given by (6.19), (6.20). The results are shown again in Fig. 10.1 in green (lower-most) curve for $\epsilon = 0$ and in red (3rd from top) curve for $\epsilon = 10^{-6}$. The conclusion is that with a properly spatially-coupled matrix **F**, one can perform error correction using AMP down to these very low encoding rates.

10.3 Numerical tests and finite size study

The asymptotic guarantees given in the last sections are encouraging, but evaluating analytically finite M corrections is intrinsically more difficult and hence we withdraw to numerical verifications of the achievable encoding rates for sizes relevant for practical applications.

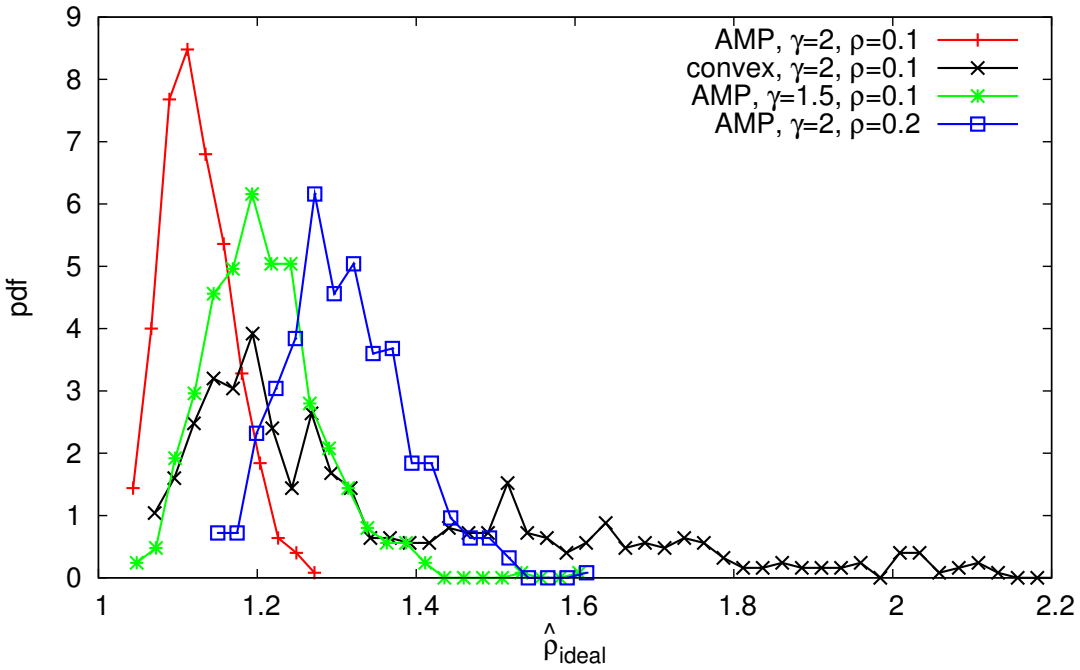

Figure 10.2 – Robustness to noise of the error correction of real signals of size $N = 256$. We compare the performances of the AMP-based and ℓ_1 decoders, using $\epsilon = 10^{-6}$ and homogeneous Gaussian i.i.d matrices $\mathbf{F}$. The figure shows the probability density (estimated over 500 instances) of the robustness ratio (10.9), called $\hat{\rho}_{ideal}$ in [159] at different values of the encoding rate γ and gross noise sparsity ρ. For coding rate $\gamma = 2$ and gross noise sparsity $\rho = 0.1$, both methods (AMP in red and ℓ_1 in black) are giving values close to one. However AMP is better: on average it gives 1.12 versus 1.36 for ℓ_1, which empirical distribution has larger tails than the one of AMP. Furthermore, AMP still performs very well when the fraction of gross errors is doubled (blue curve, with $\rho = 0.2$) or when the coding rate γ is lower (green curve, with $\gamma = 1.5$). In both cases, the ℓ_1-based reconstruction gives poor results, an average value $\hat{\rho}_{ideal} = 37.1$ for $(\rho = 0.2, \gamma = 2)$, versus 1.30 for AMP, and $\hat{\rho}_{ideal} = 18.3$ for $(\rho = 0.1, \gamma = 1.5)$ versus 1.20 for AMP.

For numerical verifications, we use a randomly generated N-d Gaussian signal $\mathbf{s}$ with zero mean and unit variance (the algorithm is *not* using this information) and a channel noise distributed according to (10.1), information that we know and use in the algorithm. We use the Bayesian AMP algorithm to estimate the error $\hat{\mathbf{e}}$. As already discussed, for exactly sparse channel $\epsilon = 0$, the exact reconstruction of $\mathbf{e}$ is possible and hence $\mathbf{s}$ can be recovered exactly. For an approximately sparse channel with $\epsilon > 0$, we use the AMP estimate of the error $\hat{\mathbf{e}}$ to compute the estimate of $\mathbf{A}\hat{\mathbf{x}}$ and finally use the pseudoinverse of $\mathbf{A}$ to estimate the signal $\hat{\mathbf{x}}$. We compare to the ℓ_1 decoding approach (including the performances-improving reprojection step) as developed in [158, 159].

The data for $N = 256$ are shown in Fig. 10.2 where the performance of the AMP decoding is compared to the ℓ_1-based decoding of [159]. Following [159] we introduce an estimator of the

robustness to noise called $\hat{\rho}_{ideal}$ as the ratio of the MSE of the reconstructed signal $\hat{\mathbf{x}}$ with the MSE of the "ideal" reconstruction $\hat{\mathbf{x}}_{ideal}$, where the pseudoinverse of $\mathbf{A}$ is applied to $\mathbf{y}$ that was corrupted only by the small AWGN without gross errors at all:

$$\hat{\rho}_{ideal} := \frac{||\hat{\mathbf{x}} - \mathbf{s}||_2}{||\hat{\mathbf{x}}_{ideal} - \mathbf{s}||_2} \tag{10.9}$$

Fig. 10.2 depicts the histogram of $\hat{\rho}_{ideal}$ over 500 random instances of the problem. We find that in all the cases we have tried with AMP (which were all in the favorable region of the asymptotic phase diagram, the easy phase above the BP transition), the robustness estimator $\hat{\rho}_{ideal}$ is very close to unity, even at these relatively small sizes. Moreover the robustness estimator of AMP was always on average closer to unity than the one based on ℓ_1 estimation and the distribution more peaked, thus demonstrating the advantage of the Bayesian AMP reconstruction in terms of performance, and noise robustness. Another important point is that the probability of an unsuccessful reconstruction is decaying exponentially fast when the system size increases. It is also decaying faster as γ increases (see Fig. 10.3, inset).

We also applied spatial coupling by choosing the matrix $\mathbf{F}$ as described in Fig. 5.4 drawn from the ensemble ($L_c = L_r = 10$, $w = 3$, $\sqrt{J} = \sqrt{0.2}$, $\alpha_{seed} = 0.22$) and varying α_{rest}. While such parameters are far from the limit $L \to \infty$, $w \to \infty$, $w/L \to 0$ in which the optimal performance is guaranteed, we still obtain a considerable improvement in the achievable encoding rate, as shown in Fig. 10.3. In Fig. 10.4 we give a more visual illustration of the performance of the spatially-coupled AMP decoder for $N = 4096$. For gross error sparsity $\rho = 0.1$ and small error variance $\epsilon = 10^{-6}$ we were able to perform reliable transmission at coding rate $\gamma = 1.256$. This has to be compared with the original approach of [159] which only allows, even for an infinite size system and in absence of small noise, an asymptotic $\gamma_{DT}(\rho = 0.1) \approx 1.5$.

10.4 Discussion

We have studied an error-correction scheme based on AMP for real channels that corrupt a fraction of the codeword by gross errors. It appeared that this scheme is highly robust to an additional AWGN. Spatial coupling allows to reach close to optimal results. Nevertheless, an important question remains: how to properly optimize the spatial coupling on finite size systems? Furthermore, the present scheme can be combined with the structured operators developed in chap. 7 in order to work with very large signals. It would be also interesting to study the combined use of the present strategy with sparse superposition codes for joint source-channel coding problems.

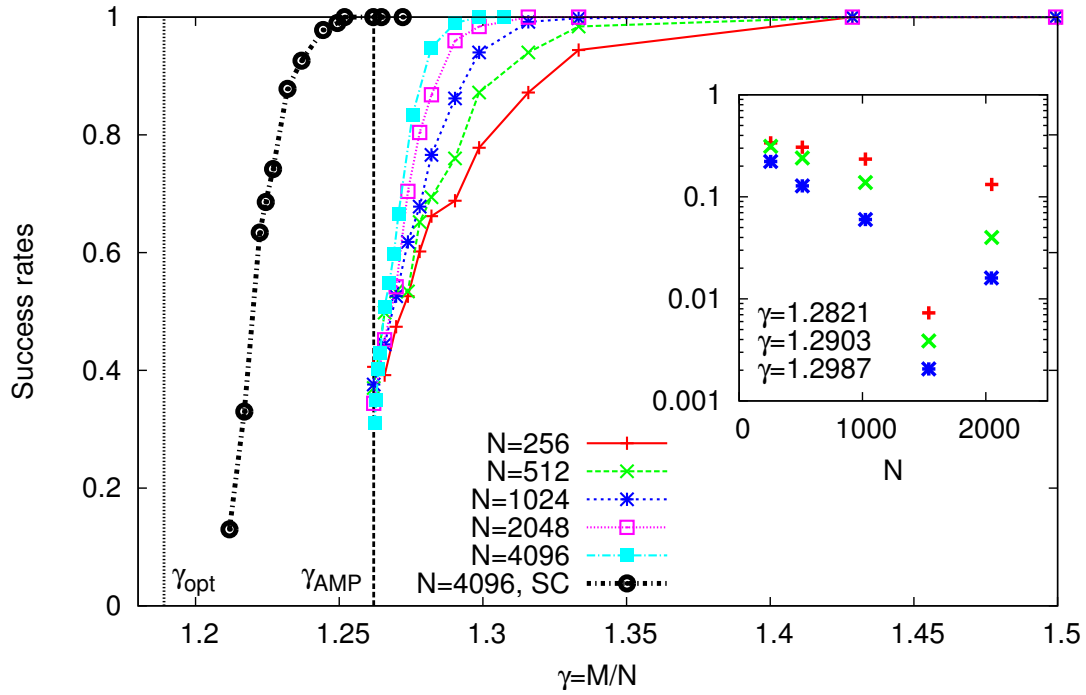

Figure 10.3 – Success rates of the decoding over 500 instances for different signal sizes N with noise parameters $\rho = 0.1$, $\epsilon = 10^{-6}$, both with spatially-coupled matrices $\mathbf{F}$ (SC) and homogeneous random Gaussian i.i.d ones, as a function of the coding rate γ. The vertical lines represent the limiting asymptotic coding rate for AMP with homogeneous (γ_{AMP}) and seeding matrices (γ_{opt}) respectively. The maximum number of iterations in these simulations is set to 1000. An instance is considered successful if the final mean square error of the reconstructed signal is less than 10^{-5}. The inner plot shows the empirical probability of failure over these instances for three different values of encoding rate γ. It decays exponentially with the signal size and the decay exponent-amplitude increases with the coding rate.

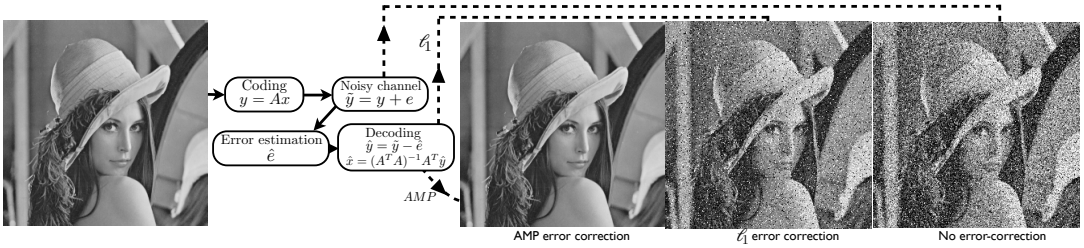

Figure 10.4 – Illustration of the error correction scheme with the spatially-coupled AMP approach and the ℓ_1-based method of [159], applied to the benchmark Lena picture. The original 256×256 image is decomposed in patches of size $N = 64^2$. The noisy channel is given by (10.1) using $\rho = 0.1$, $\epsilon = 10^{-6}$ and the coding rate is $\gamma = 1.256$ (to be compared with $\gamma_{opt}(\rho = 0.1, \epsilon = 10^{-6}) = 1.184$) is used together with a spatially-coupled matrix $\mathbf{F}$ with parameters ($L_c = L_r = 10$, $w = 3$, $\sqrt{J} = \sqrt{0.2}$, $\alpha_{seed} = 0.22$, $\alpha_{rest} = 0.1830$). While error correction is close to perfect with AMP, the results are as poor as no error correction at all with an ℓ_1 convex optimization solver.

Bibliography

[1] J. Barbier, F. Krzakala, L. Zdeborová, and P. Zhang, "The hard-core model on random graphs revisited," in *Journal of Physics: Conference Series*, vol. 473. IOP Publishing, 2013, p. 012021.

[2] J. Barbier, F. Krzakala, M. Mézard, and L. Zdeborová, "Compressed sensing of approximately-sparse signals: Phase transitions and optimal reconstruction," in *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[3] J. Barbier, C. Schülke, and F. Krzakala, "Approximate message-passing with spatially coupled structured operators, with applications to compressed sensing and sparse superposition codes," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015, no. 5, p. P05013, 2015. [Online]. Available: http://stacks.iop.org/1742-5468/2015/i=5/a=P05013

[4] J. Barbier, F. Krzakala, L. Zdeborova, and P. Zhang, "Robust error correction for real-valued signals via message-passing decoding and spatial coupling," in *Information Theory Workshop (ITW), 2013 IEEE*, Sept 2013, pp. 1–5.

[5] J. Barbier and F. Krzakala, "Replica analysis and approximate message passing decoder for superposition codes," in *2014 IEEE International Symposium on Information Theory*, 2014.

[6] ——, "Approximate message-passing decoder and capacity-achieving sparse superposition codes," *CoRR*, vol. abs/1503.08040, 2015. [Online]. Available: http://arxiv.org/abs/1503.08040

[7] D. J. MacKay, *Information theory, inference, and learning algorithms*. Citeseer, 2003, vol. 7.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, ser. Springer Texts in Statistics. Springer New York, 2013. [Online]. Available: http://books.google.fr/books?id=qcI_AAAAQBAJ

[9] E. W. Tramel, S. Kumar, A. Giurgiu, and A. Montanari, "Statistical estimation: From denoising to sparse regression and hidden cliques," *CoRR*, vol. abs/1409.5557, 2014. [Online]. Available: http://arxiv.org/abs/1409.5557

[10] L. Wasserman, *All of statistics : a concise course in statistical inference*. New York: Springer, 2010. [Online]. Available: http://www.amazon.de/All-Statistics-Statistical-Inference-Springer/dp/1441923225/ref=sr_1_2?ie=UTF8&qid=1356099149&sr=8-2

[11] C. Schülke, F. Caltagirone, and L. Zdeborová, "Blind sensor calibration using approximate message passing," *CoRR*, vol. abs/1406.5903, 2014. [Online]. Available: http://arxiv.org/abs/1406.5903

[12] J. Sakellariou, "Inverse inference in the asymmetric ising model," Ph.D. dissertation, Université Paris-Sud 11, 2013.

[13] A. Guggiola and G. Semerjian, "Minimal contagious sets in random regular graphs," *Journal of Statistical Physics*, vol. 158, no. 2, pp. 300–358, 2015.

[14] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption: clustering sparse networks," *CoRR*, vol. abs/1306.5550, 2013. [Online]. Available: http://arxiv.org/abs/1306.5550

[15] A. Saade, F. Krzakala, and L. Zdeborová, "Spectral Clustering of Graphs with the Bethe Hessian," *ArXiv e-prints*, Jun. 2014.

[16] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75 – 174, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0370157309002841

[17] A. Decelle, J. Hüttel, A. Saade, and C. Moore, "Computational Complexity, Phase Transitions, and Message-Passing for Community Detection," *ArXiv e-prints*, Sep. 2014.

[18] C. Moore and S. Mertens, *The nature of computation*. Oxford University Press, 2011.

[19] L. Wasserman, *All of Nonparametric Statistics (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[20] Y. Ma, D. Baron, and A. Beirami, "Mismatched Estimation in Large Linear Systems," *ArXiv e-prints*, May 2015.

[21] E. W. Tramel, A. Drémeau, and F. Krzakala, "Approximate message passing with restricted boltzmann machine priors," *CoRR*, vol. abs/1502.06470, 2015. [Online]. Available: http://arxiv.org/abs/1502.06470

[22] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Inference and Phase Transitions in the Detection of Modules in Sparse Networks," *Physical Review Letters*, vol. 107, no. 6, p. 065701, Aug. 2011.

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

[24] M. Mézard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[25] V. Sessak, "Inverse problems in spin models," Ph.D. dissertation, PhD Thesis, 2010, 2010.

[26] F. Ricci-Tersenghi, "The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 8, p. 15, Aug. 2012.

[27] F. Krzakala and L. Zdeborová, "Hiding quiet solutions in random constraint satisfaction problems," *Phys. Rev. Lett.*, vol. 102, p. 238701, 2009.

[28] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *CoRR*, vol. abs/0808.3572, 2008. [Online]. Available: http://arxiv.org/abs/0808.3572

[29] E. Candes, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *arXiv.org*, Mar. 2005.

[30] E. J. Candès and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" *IEEE Trans. Inform. Theory*, vol. 52, p. 5406, 2006.

[31] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, 2006.

[32] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, p. 1289, 2006.

[33] E. Candes and M. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, March 2008.

[34] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical physics-based reconstruction in compressed sensing," *Phys. Rev. X*, vol. 2, p. 021005, 2012.

[35] ——, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 8, p. P08009, August 2012.

[36] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," in *IEEE SIGNAL PROCESSING MAGAZINE*, 2007.

[37] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vandergheynst, "Compressed sensing imaging techniques for radio interferometry," *Monthly Notices of the Royal Astronomical Society*, vol. 395, no. 3, pp. 1733–1742, 2009. [Online]. Available: http://mnras.oxfordjournals.org/content/395/3/1733.abstract

[38] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 83–91, 2008.

[39] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, "Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium," *Scientific Reports*, vol. 4, p. 5552, Jul. 2014.

[40] A. Dremeau, A. Liutkus, D. Martina, O. Katz, C. Schulke, F. Krzakala, S. Gigan, and L. Daudet, "Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques," *ArXiv e-prints*, Feb. 2015.

[41] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval, "Near-field acoustic holography using sparse regularization and compressive sampling principles," *Acoustical Society of America Journal*, vol. 132, p. 1521, 2012.

[42] R. Mignot, L. Daudet, and F. Ollivier, "Compressed sensing for acoustic response reconstruction: Interpolation of the early part," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 225–228.

[43] P. Zhang, F. Krzakala, M. Mézard, and L. Zdeborová, "Non-adaptive pooling strategies for detection of rare faulty items," *CoRR*, vol. abs/1302.0189, 2013. [Online]. Available: http://arxiv.org/abs/1302.0189

[44] J. N. Sanders, X. Andrade, and A. Aspuru-Guzik, "Compressed sensing for the fast computation of matrices: Application to molecular vibrations," *ACS Central Science*, vol. 1, no. 1, pp. 24–32, 2015. [Online]. Available: http://dx.doi.org/10.1021/oc5000404

[45] C. Papadimitriou, *Computational Complexity*, ser. Theoretical computer science. Addison-Wesley, 1994. [Online]. Available: http://books.google.fr/books?id=JogZAQAAIAAJ

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[47] E. van den Berg, "Convex optimization for generalized sparse recovery," Ph.D. dissertation, The University of British Columbia (Vancouver, 2009.

[48] A. Y. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma, "A review of fast l1-minimization algorithms for robust face recognition," *CoRR*, vol. abs/1007.3753, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1007.html#abs-1007-3753

[49] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Royal Society of London Philosophical Transactions Series A*, vol. 367, pp. 4273–4293, Nov. 2009.

[50] E. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, Dec 2005.

[51] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," 2006.

[52] S. Becker, J. Bobin, and E. J. Candès, "Nesta: A fast and accurate first-order method for sparse recovery." *SIAM J. Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/siamis/siamis4.html#BeckerBC11

[53] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing*. Oxford: Oxford University Press, 2001.

[54] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[55] R. Gribonval, G. Chardon, and L. Daudet, "Blind calibration for compressed sensing by convex optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2713 –2716.

[56] F. Krzakala, M. Mézard, and L. Zdeborová, "Compressed sensing under matrix uncertainty: Optimum thresholds and robust approximate message passing," 2012, arXiv:1301.0901 [cs.IT], ICASSP 2013.

[57] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata., and L. Zdeborová, "Phase transitions and sample complexity in bayes-optimal matrix factorization," *CoRR*, vol. abs/1402.1298, 2014. [Online]. Available: http://arxiv.org/abs/1402.1298

[58] C. Schülke, F. Caltagirone, F. Krzakala, and L. Zdeborová, "Blind calibration in compressed sensing using message passing algorithms," *CoRR*, vol. abs/1306.4355, 2013. [Online]. Available: http://arxiv.org/abs/1306.4355

[59] F. Krzakala, M. Mézard, and L. Zdeborová, "Phase diagram and approximate message passing for blind calibration and dictionary learning," *CoRR*, vol. abs/1301.5898, 2013. [Online]. Available: http://arxiv.org/abs/1301.5898

[60] C. Bilen, G. Puy, R. Gribonval, and L. Daudet, "Convex optimization approaches for blind sensor calibration using sparsity," *CoRR*, vol. abs/1308.5354, 2013. [Online]. Available: http://arxiv.org/abs/1308.5354

[61] M. Figueiredo, "Lecture notes on bayesian estimation and classification." [Online]. Available: http://www.lx.it.pt/~mtf/learning/Bayes_lecture_notes.pdf

[62] F. Krzakala, L. Zdeborova, M. C. Angelini, and F. Caltagirone, "Statistical physics of inference and bayesian estimation." [Online]. Available: http://indico.ictp.it/event/a14244/material/10/0.pdf

[63] A. Montanari and R. L. Urbanke, "Modern coding theory: The statistical mechanics and computer science point of view," *CoRR*, vol. abs/0704.2857, 2007. [Online]. Available: http://arxiv.org/abs/0704.2857

[64] M. M. Zecchina R. and P. G., "Analytic and algorithmic solution of random satisfiability problems," *Science*, vol. 297, no. 812, 2002.

[65] D. Saad, Y. Kabashima, and T. Murayama, "Public key cryptography and error correcting codes as Ising models," in *American Institute of Physics Conference Series*, ser. American Institute of Physics Conference Series, vol. 553, Feb. 2001, pp. 89–94.

[66] S. Franz, M. Leone, A. Montanari, and F. Ricci-Tersenghi, "Dynamic phase transition for decoding algorithms," vol. 66, no. 4, p. 046120, Oct. 2002.

[67] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond*. Singapore: World Scientific, 1987, vol. 9.

[68] N. Sourlas, "Statistical mechanics and capacity-approaching error-correcting codes," *Physica A Statistical Mechanics and its Applications*, vol. 302, pp. 14–21, Dec. 2001.

[69] N. Sourlas, "Statistical mechanics and error-correcting codes," in *From Statistical Physics to Statistical Inference and Back*, ser. NATO ASI Series, P. Grassberger and J.-P. Nadal, Eds. Springer Netherlands, 1994, vol. 428, pp. 195–204. [Online]. Available: http://dx.doi.org/10.1007/978-94-011-1068-6_12

[70] Y. Kabashima and D. Saad, "Statistical mechanics of error-correcting codes," *EPL (Europhysics Letters)*, vol. 45, no. 1, p. 97, 1999. [Online]. Available: http://stacks.iop.org/0295-5075/45/i=1/a=097

[71] H. Nishimori and K. Y. M. Wong, "Statistical mechanics of image restoration and error-correcting codes," *Phys. Rev. E*, vol. 60, pp. 132–144, Jul 1999. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.60.132

[72] G. Migliorini and D. Saad, "Finite-connectivity spin-glass phase diagrams and low-density parity check codes," vol. 73, no. 2, p. 026122, Feb. 2006.

[73] R. Vicente, D. Saad, and Y. Kabashima, "Finite-connectivity systems as error-correcting codes," vol. 60, pp. 5352–5366, Nov. 1999.

[74] R. C. Alamino and D. Saad, "Statistical mechanics analysis of LDPC coding in MIMO Gaussian channels," *Journal of Physics A Mathematical General*, vol. 40, pp. 12 259–12 279, Oct. 2007.

[75] I. Kanter and D. Saad, "Finite-size effects and error-free communication in Gaussian channels," *Journal of Physics A Mathematical General*, vol. 33, pp. 1675–1681, Mar. 2000.

[76] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Sparse estimation with the swept approximated message-passing algorithm," *CoRR*, vol. abs/1406.4311, 2014. [Online]. Available: http://arxiv.org/abs/1406.4311

[77] F. Caltagirone, F. Krzakala, and L. Zdeborová, "On convergence of approximate message passing," *CoRR*, vol. abs/1401.6384, 2014. [Online]. Available: http://arxiv.org/abs/1401.6384

[78] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborová, "Variational free energies for compressed sensing," *CoRR*, vol. abs/1402.1384, 2014. [Online]. Available: http://arxiv.org/abs/1402.1384

[79] M. Opper and D. Saad, *Advanced Mean Field Methods : Theory and Practice*. MIT press, 2001.

[80] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *CoRR*, vol. abs/0805.0510, 2008. [Online]. Available: http://arxiv.org/abs/0805.0510

[81] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, "Gibbs states and the set of solutions of random constraint satisfaction problems," *CoRR*, vol. abs/cond-mat/0612365, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr0612.html#abs-cond-mat-0612365

[82] L. Zdeborová, "Statistical physics of hard optimization problems," *CoRR*, vol. abs/0806.4112, 2008. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr0806.html#abs-0806-4112

[83] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theor.*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1109/TIT.2005.850085

[84] H.-J. Zhou and W.-M. Zheng, "Loop-corrected belief propagation for lattice spin models," *ArXiv e-prints*, May 2015.

[85] M. Mézard, G. Parisi, and R. Zecchina, "Analytic and algorithmic solution of random satisfiability problems," *Science*, vol. 297, pp. 812–815, 2002.

[86] A. Braunstein, M. Mézard, and R. Zecchina, "Survey propagation: an algorithm for satisfiability," *CoRR*, vol. cs.CC/0212002, 2002. [Online]. Available: http://arxiv.org/abs/cs.CC/0212002

[87] A. Braunstein, M. Mezard, M. Weigt, and R. Zecchina, "Constraint Satisfaction by Survey Propagation," *eprint arXiv:cond-mat/0212451*, Dec. 2002.

[88] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.

[89] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. of the IEEE Int. Symp. on Inform. Theory (ISIT)*, 2011, pp. 2168 –2172.

[90] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *IEEE Information Theory Workshop (ITW)*, 2010, pp. 1 –5.

[91] D. Baron, S. Sarvotham, and R. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 269 – 280, 2010.

[92] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of 'solvable model of a spin-glass'," *Phil. Mag.*, vol. 35, pp. 593–601, 1977.

[93] J. P. Neirotti and D. Saad, "Improved message passing for inference in densely connected systems," *CoRR*, vol. abs/cs/0503070, 2005. [Online]. Available: http://arxiv.org/abs/cs/0503070

[94] D. Guo and C.-C. Wang, "Asymptotic mean-square optimality of belief propagation for sparse linear systems," *Information Theory Workshop, 2006. ITW '06 Chengdu.*, pp. 194–198, 2006.

[95] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, 2010, pp. 1 –6.

[96] ——, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*. IEEE, 2010, pp. 1–6.

[97] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[98] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," *Communications of the ACM*, vol. 53, no. 10, pp. 95–103, 2010.

[99] J. P. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," *CoRR*, vol. abs/1412.2005, 2014. [Online]. Available: http://arxiv.org/abs/1412.2005

[100] Q. Guo and J. Xi, "Approximate message passing with unitary transformation," *CoRR*, vol. abs/1504.04799, 2015. [Online]. Available: http://arxiv.org/abs/1504.04799

[101] Y. Wu and S. Verdu, "Optimal phase transitions in compressed sensing," 2011, arXiv:1111.6822v1 [cs.IT].

[102] S. Rangan, A. Fletcher, and V. Goyal, "Asymptotic analysis of map estimation via the replica method and applications to compressed sensing," *arXiv:0906.3234v2*, 2009.

[103] D. Guo, D. Baron, and S. Shamai, "A single-letter characterization of optimal noisy compressed sensing," in *47th Annual Allerton Conference on Communication, Control, and Computing*, 2009, pp. 52 – 59.

[104] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764 –785, 2011.

[105] D. Guo and C.-C. Wang, "Random sparse linear system observed via arbitrary channels: A decoupling principle," *Proc. IEEE Int. Symp. Inform. Th., Nice, France*, pp. 946–950, 2007.

[106] C. Wen and K. Wong, "Analysis of compressed sensing with spatially-coupled orthogonal matrices," *CoRR*, vol. abs/1402.3215, 2014. [Online]. Available: http://arxiv.org/abs/1402.3215

[107] A. Jimenez Felstrom and K. Zigangirov, "Time-varying periodic convolutional codes with low-density parity-check matrix," *Information Theory, IEEE Transactions on*, vol. 45, no. 6, pp. 2181 –2191, 1999.

[108] S. Kudekar, T. Richardson, and R. Urbanke, "Threshold saturation via spatial coupling: Why convolutional ldpc ensembles perform so well over the bec," in *Proc. of the IEEE Int. Symposium on Information Theory (ISIT),*, 2010, pp. 684–688.

[109] ——, "Spatially coupled ensembles universally achieve capacity under belief propagation," 2012, arXiv:1201.2999v1 [cs.IT].

[110] D. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," in *Proc. of the IEEE Int. Symposium on Information Theory (ISIT)*, 2012, pp. 1231–1235.

[111] A. Giurgiu, N. Macris, and R. L. Urbanke, "Spatial coupling as a proof technique," *CoRR*, vol. abs/1301.5676, 2013. [Online]. Available: http://arxiv.org/abs/1301.5676

[112] S. H. Hassani, N. Macris, and R. L. Urbanke, "Coupled graphical models and their thresholds," *CoRR*, vol. abs/1105.0785, 2011. [Online]. Available: http://arxiv.org/abs/1105.0785

[113] D. Achlioptas, S. Hamed Hassani, N. Macris, and R. Urbanke, "New Bounds for Random Constraint Satisfaction Problems via Spatial Coupling," Tech. Rep., 2013.

[114] M. C. Angelini, F. Ricci-Tersenghi, and Y. Kabashima, "Compressed sensing with sparse, structured matrices," *CoRR*, vol. abs/1207.2853, 2012. [Online]. Available: http://arxiv.org/abs/1207.2853

[115] F. Caltagirone and L. Zdeborová, "Properties of spatial coupling in compressed sensing," *arXiv preprint arXiv:1401.6380*, 2014.

[116] F. Caltagirone, S. Franz, R. Morris, and L. Zdeborová, "Dynamics and termination cost of spatially coupled mean-field models." *CoRR*, vol. abs/1310.2121, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1310.html#CaltagironeFMZ13

[117] A. Montanari, "Graphical models concepts in compressed sensing," *Compressed Sensing: Theory and Applications*, pp. 394–438, 2012.

[118] J. P. Vila and P. Schniter, "Expectation-maximization bernoulli-gaussian approximate message passing," in *Proc. Asilomar Conf. on Signals, Systems, and Computers (Pacific Grove, CA)*, 2011.

[119] S. Kudekar and H. Pfister, "The effect of spatial coupling on compressive sensing," in *Communication, Control, and Computing (Allerton)*, 2010, pp. 347–353.

[120] A. Yedla, Y. Jian, P. Nguyen, and H. Pfister, "A simple proof of threshold saturation for coupled scalar recursions," 2012, arXiv:1204.5703v1 [cs.IT].

[121] A. Amraoui, A. Montanari, and R. Urbanke, "How to find good finite-length codes: from art towards science," *European Transactions on Telecommunications*, vol. 18, no. 5, pp. 491–508, 2007. [Online]. Available: http://dx.doi.org/10.1002/ett.1182

[122] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a markov-tree prior," *CoRR*, vol. abs/1108.2632, 2011. [Online]. Available: http://arxiv.org/abs/1108.2632

[123] J. P. Vila, P. Schniter, and J. Meola, "Hyperspectral unmixing via turbo bilinear approximate message passing," *CoRR*, vol. abs/1502.06435, 2015. [Online]. Available: http://arxiv.org/abs/1502.06435

[124] W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang, "Compressive sensing via nonlocal low-rank regularization," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3618–3632, 2014. [Online]. Available: http://dx.doi.org/10.1109/TIP.2014.2329449

[125] J. Tan, Y. Ma, and D. Baron, "Compressive imaging via approximate message passing with image denoising," *IEEE Transactions on Signal Processing*, vol. 63, no. 8, pp. 2085–2092, 2015. [Online]. Available: http://dx.doi.org/10.1109/TSP.2015.2408558

[126] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *CoRR*, vol. abs/1406.4175, 2014. [Online]. Available: http://arxiv.org/abs/1406.4175

[127] T. T. Do, T. D. Tran, and L. Gan, "Fast compressive sampling with structurally random matrices," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3369–3372.

[128] A. Javanmard and A. Montanari, "Subsampling at information theoretically optimal rates," 2012, arXiv:1202.2525v1 [cs.IT].

[129] U. S. Kamilov, A. Bourquard, and M. Unser, "Sparse image deconvolution with message passing," sPARSE 2013.

[130] M. Vehkapera, Y. Kabashima, and S. Chatterjee, "Analysis of regularized ls reconstruction and random matrix ensembles in compressed sensing," *arXiv preprint arXiv:1312.0256*, 2013.

[131] C. Wen and K. Wong, "Analysis of compressed sensing with spatially-coupled orthogonal matrices," *arXiv preprint arXiv:1402.3215*, 2014.

[132] A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk, "Asymptotic analysis of complex lasso via complex approximate message passing (camp)," 2012.

[133] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 815–822.

[134] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, "Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)," *CoRR*, vol. abs/1108.0477, 2011. [Online]. Available: http://arxiv.org/abs/1108.0477

[135] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *CoRR*, vol. abs/1405.5618, 2014. [Online]. Available: http://arxiv.org/abs/1405.5618

[136] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *arXiv preprint arXiv:1207.7321*, 2012.

[137] C. Li, "An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing," Master's thesis, Rice University, Sep. 2009.

[138] D. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *arXiv Preprint*, no. 1111.1041, January 2013.

[139] M. A. Borgerding and P. Schniter, "Generalized approximate message passing for the cosparse analysis model," *arXiv Preprint*, no. 1312.3968, December 2013.

[140] J. Kang, H. Jung, H.-N. Lee, and K. Kim, "Spike-and-slab approximate message-passing for high-dimensional picewise-constant recovery," *arXiv Preprint*, no. 1408.3930, August 2014.

[141] V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candes, and M. Dahan, "Compressive fluorescence microscopy for biological and hyperspectral imaging," *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. E1679–E1687, 2012. [Online]. Available: http://www.pnas.org/content/109/26/E1679.abstract

[142] A. R. Barron and A. Joseph, "Sparse superposition codes: Fast and reliable at rates approaching capacity with gaussian noise," *Manuscript. Available at ‚Äúhttp://www. stat. yale. edu/ arb4*, 2010.

[143] ——, "Analysis of fast sparse superposition codes," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1772–1776.

[144] A. Joseph and A. R. Barron, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *Information Theory, IEEE Transactions on*, vol. 58, no. 5, pp. 2541–2557, 2012.

[145] A. R. Barron and S. Cho, "High-rate sparse superposition codes with iteratively optimal estimates," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 120–124.

[146] S. Cho and A. Barron, "Approximate iterative bayes optimal estimates for high-rate sparse superposition codes."

[147] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[148] C. Rush, A. Greig, and R. Venkataramanan, "Capacity-achieving sparse superposition codes via approximate message passing decoding," *arXiv preprint arXiv:1501.05892*, 2015.

[149] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference*, p. iat004, 2013.

[150] B. Derrida, "Random-energy model: Limit of a family of disordered models," *Physical Review Letters*, vol. 45, no. 2, pp. 79–82, 1980.

[151] G. B. Arous, L. V. Bogachev, and S. A. Molchanov, "Limit theorems for sums of random exponentials," *Probability theory and related fields*, vol. 132, no. 4, pp. 579–612, 2005.

[152] A. Wyner, "An analog scrambling scheme which does not expand bandwidth, part i: Discrete time," *Information Theory, IEEE Transactions on*, vol. 25, no. 3, pp. 261–274, 1979.

[153] S. Feizi and M. Medard, "A power efficient sensing/communication scheme: Joint source-channel-network coding by using compressive sensing," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, 2011, pp. 1048–1054.

[154] S. Shintre, S. Katti, S. Jaggi, B. Dey, D. Katabi, and M. Medard, "Real and complex network codes: Promises and challenges," 2008.

[155] M. Grangetto, P. Cosman, and G. Olmo, "Joint source/channel coding and map decoding of arithmetic codes," *Communications, IEEE Transactions on*, vol. 53, no. 6, pp. 1007–1016, 2005.

[156] G. Caire, T. Al-Naffouri, and A. Narayanan, "Impulse noise cancellation in ofdm: an application of compressed sensing," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, 2008, pp. 1293–1297.

[157] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *Information Theory, IEEE Transactions on*, vol. 47, no. 7, pp. 2845–2862, 2001.

[158] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, p. 4203, 2005.

[159] E. J. Candes and P. A. Randall, "Highly robust error correction byconvex programming," *IEEE Trans. Inf. Theor.*, vol. 54, no. 7, pp. 2829–2840, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1109/TIT.2008.924688

[160] L. Lampe, "Bursty impulse noise detection by compressed sensing," in *Power Line Communications and Its Applications (ISPLC), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 29–34.

[161] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proc. Natl. Acad. Sci.*, vol. 102, no. 27, pp. 9446–9451, 2005.