Graphical Models Concepts in Compressed Sensing

Andrea Montanari*

Abstract

This paper surveys recent work in applying ideas from graphical models and message passing algorithms to solve large scale regularized regression problems. In particular, the focus is on compressed sensing reconstruction via ℓ_1 penalized least-squares (known as LASSO or BPDN). We discuss how to derive fast approximate message passing algorithms to solve this problem. Surprisingly, the analysis of such algorithms allows to prove exact high-dimensional limit results for the LASSO risk.

This paper will appear as a chapter in a book on 'Compressed Sensing' edited by Yonina Eldar and Gitta Kutynok.

1 Introduction

The problem of recostructing a high-dimensional vector $x \in \mathbb{R}^n$ from a collection of observations $y \in \mathbb{R}^m$ arises in a number of contexts, ranging from statistical learning to signal processing. It is often assumed that the measurement process is approximately linear, i.e. that

$$y = Ax + w\,, \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$ is a known measurement matrix, and w is a noise vector.

The graphical models approach to such reconstruction problem postulates a joint probability distribution on (x, y) which takes, without loss of generality, the form

$$p(\mathrm{d}x,\, \mathrm{d}y) = p(\mathrm{d}y|x)\, p(\mathrm{d}x)\,. \tag{1.2}$$

The conditional distribution $p(\mathrm{d}y|x)$ models the noise process, while the prior $p(\mathrm{d}x)$ encodes information on the vector x. In particular, within compressed sensing, it can describe its sparsity properties. In particular, within a *graphical models* approach, either of these distributions (or both) factorizes according to a specific graph structure. The resulting posterior distribution $p(\mathrm{d}x|y)$ is used for inferring x given y.

There are many reasons to be skeptical about the idea that the joint probability distribution $p(\mathrm{d}x,\, \mathrm{d}y)$ can be determined, and hence used for reconstructing x. One might be tempted to drop the whole approach as a consequence. We argue that sticking to this point of view is instead fruitful for several reasons:

*Department of Electrical Engineering and Department of Statistics, Stanford University

1. *Algorithmic.* Most of existing reconstruction methods can be derived as Bayesian estimators (e.g. maximum a posteriori probability) for specific forms of $p(\mathrm{d}x)$ and $p(\mathrm{d}y|x)$. The connection is useful both in interpreting/comparing different methods, and in adapting known algorithms for Bayes estimation (e.g. graphical models inference algorithms).

2. *Minimax.* When the prior $p(\mathrm{d}x)$ or the noise distributions, and therefore the conditional distribution $p(\mathrm{d}y|x)$, 'exist' but are unknown, it is reasonable to assume that they belong to specific structure classes. For instace, within compressed sensing one often assumes that x has at most k non-zero entries. One can then take $p(\mathrm{d}x)$ to be a distribution supported on k-sparse vectors $x \in \mathbb{R}^n$. If $\mathcal{F}_{n,k}$ denotes the class of such distributions, the minimax criterion approach strives to achieve the best uniform guarantee over $\mathcal{F}_{n,k}$. In other words, the minimax estimator achieves the *smallest* expected error (e.g. mean square error) for the worst distribution in $\mathcal{F}_{n,k}$.

   It is a remarkable fact in statistical decision theory that the minimax estimator coincides with the Bayes estimator for a specific (worst case) $p \in \mathcal{F}_{n,k}$.

3. *Modeling.* In some applications it is is possible to construct fairly accurate models both of the prior distribution $p(\mathrm{d}x)$ and of the measurement process $p(\mathrm{d}y|x)$. This is the case for instance in some communications problems, whereby x is the signal produced by a transmitter (and generated uniformly at random according to a known codebook), and w is the noise produced by a well-defined physical process (e.g. thermal noise in the receiver circuitry).

The rest of this chapter is organized as follows. Section 2 describes a graphical model naturally associated to the compressed sensing reconstruction problem. Section 3 provides important background on the one-dimensional case. Section 4 describes a standard message passing algorithm –the min-sum algorithm– and how it can be simplified to solve the LASSO optimization problem. The algorithm is further simplified in Section 5 yielding the AMP algoritm. The analysis of this algorithm is outlined in Section 6. As a consequence of this analysis, it is possible to compute exact high-dimensional limits for the behavior of the LASSO estimator. Finally in Section 7 we discuss a few examples of how the approach developed here can be used to address reconstruction problems in which a richer structural information is available.

2 The basic model and its graph structure

Specifying the conditional distribution of y given x is equivalent to specifying the distribution of the noise vector w. In the rest of this chapter we shall take $p(w)$ to be a gaussian distribution of mean 0 and variance $\beta^{-1}\mathbf{I}$, whence

$$p(\mathrm{d}y|x) = \Big(\frac{\beta}{2\pi}\Big)^{n/2} \exp\Big\{-\frac{\beta}{2}\|y - Ax\|^2\Big\}\,. \tag{2.1}$$

The simplest choice for the prior consists in taking $p(\mathrm{d}x)$ to be a product distribution with identical components. We thus obtain the joint distribution

$$p(\mathrm{d}x,\, \mathrm{d}y) = \Big(\frac{\beta}{2\pi}\Big)^{n/2} \exp\Big\{-\frac{\beta}{2}\|y - Ax\|^2\Big\}\,\mathrm{d}y \prod_{i=1}^{n} p(\mathrm{d}x_i)\,. \tag{2.2}$$

It is clear at the outset that generalizations of this basic model can be easily defined, in such a way to incorporating further information on the vector x or on the measurement process. As an example,

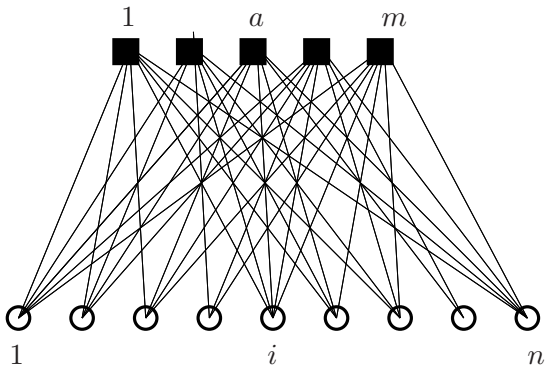

Figure 1: Factor graph associated to the probability distribution (2.5). Empty circles correspond to variables x_i, $i \in [n]$ and squares correspond to measurements y_a, $a \in [m]$.

consider the case of block-sparse signals: The index set $[n]$ is partitioned into blocks $B(1)$, $B(2)$, $\dots B(\ell)$ of equal length n/ℓ, and only a small fraction of blocks is non-vanishing. This situation can be captured by assuming that the prior $p(\mathrm{d}x)$ factors over blocks. One thus obtain the joint distribution

$$p(\mathrm{d}x,\, \mathrm{d}y) = \Big(\frac{\beta}{2\pi}\Big)^{n/2} \exp\Big\{ -\frac{1}{2\sigma_0^2}\|y - Ax\|^2\Big\}\, \mathrm{d}y \prod_{j=1}^{\ell} p(\mathrm{d}x_{B(j)})\,, \tag{2.3}$$

where $x_{B(j)} \equiv (x_i : i \in B(j)) \in \mathbb{R}^{n/\ell}$. Other examples of structured priors will be discussed in Section 7.

The posterior distribution of x given observations y is easily computed from Eq. (2.2):

$$p(\mathrm{d}x|\, y) = \frac{1}{Z(y)} \exp\Big\{ -\frac{\beta}{2}\|y - Ax\|^2\Big\} \prod_{i=1}^{n} p(\mathrm{d}x_i)\,, \tag{2.4}$$

where $Z(y) = (2\pi/\beta)^{n/2} p(y)$ ensures the normalization $\int p(\mathrm{d}x|y) = 1$. Finally, the square residuals $\|y - Ax\|^2$ decompose in a sum of m terms yielding

$$p(\mathrm{d}x|\, y) = \frac{1}{Z(y)} \prod_{a=1}^{m} \exp\Big\{ -\frac{\beta}{2}\big(y_a - A_a^T x\big)^2\Big\} \prod_{i=1}^{n} p(\mathrm{d}x_i)\,, \tag{2.5}$$

where A_a is the a-th row of the matrix a. This factorized structure is conveniently described by a *factor graph*, i.e. a bipartite graph including a 'variable node' $i \in [n]$ for each variable x_i, and a 'factor node' $a \in [m]$ for each term $\psi_a(x) = \exp\{-\beta(y_a - A_a^T x)^2/2\}$. Variable i and factor a are connected by an edge if and only if $\psi_a(x)$ depends non-trivially on x_i, i.e. if $A_{ai} \neq 0$. One such factor graphs is reproduced in Fig. 1.

An estimate of the signal can be extracted from the posterior distribution (2.5) in various ways. One possibility is to use conditional expectation

$$\widehat{x}_\beta(y; p) \equiv \int_{\mathbb{R}^n} x\; p(\mathrm{d}x|y)\,. \tag{2.6}$$

Classically, this is justified by the fact that it achieves the minimal mean square provided the $p(\mathrm{d}x, \mathrm{d}y)$ is the *actual* joint distribtion of (x, y). In the present context, the best justification is that a broad class of estimators can be written in the form (2.6).

An important problem with the estimator (2.6) is that it is in general hard to compute. In order to obtain a tractable estimator, we assume that $p(\mathrm{d}x_i) = c\, p_{\beta h}(x_i)\mathrm{d}x_i$ for $p_{\beta h}(x_i) = e^{-\beta h(x_i)}$ an unnormalized probability density function. One can then replace the integral in $\mathrm{d}x$ with a maximization over x and define

$$\widehat{x}(y; h) \equiv \mathrm{argmin}_{z\in\mathbb{R}^n} \mathcal{C}_{A,y}(z; \lambda)\,, \qquad (2.7)$$
$$\mathcal{C}_{A,y}(z; \theta) \equiv \frac{1}{2}\|y - Az\|^2 + \sum_{i=1}^{n} h(z_i)\,,$$

where we assumed for simplicity that $\mathcal{C}_{A,y}(z; \theta)$ has a unique minimum.

The estimator $\widehat{x}(y; \theta)$ can be thought of as the $\beta \to \infty$ limit of the general estimator (2.6). Indeed, it is easy to check that, provided $x_i \mapsto h(x_i)$ is upper semicontinuous, we have

$$\lim_{\beta\to\infty} \widehat{x}_\beta(y; p_{\beta h}) = \widehat{x}(y; h)\,.$$

Further, $\widehat{x}(y; h)$ takes the familiar form of a regression estimator with separable regularization. If $h(\,\cdot\,)$ is convex, the computation of $\widehat{x}$ is tractable. Important special cases include $h(x_i) = \lambda x_i^2$, which corresponds to ridge regression, and $h(x_i) = \lambda|x_i|$ which corresponds to the LASSO [Tib96] or basis pursuit denoising (BPDN) [CD95]. Due to the special role it plays in compressed sensing, we will devote special attention to this case, that we rewrite explicitely below with a slight abuse of notation

$$\widehat{x}(y) \equiv \mathrm{argmin}_{z\in\mathbb{R}^n} \mathcal{C}_{A,y}(z)\,, \qquad (2.8)$$
$$\mathcal{C}_{A,y}(z) \equiv \frac{1}{2}\|y - Az\|^2 + \lambda\|z\|_1\,.$$

3 Revisiting the scalar case

Before proceeding further, it is convenient to pause for a moment and consider the special case of a single measurement of a scalar quantity, i.e. the case $m = n = 1$. We therefore have

$$y = x + w\,, \qquad (3.1)$$

and want to estimate x from y. Despite the apparent simplicity, there exists a copious literature on this problem with many open problems [DJHS92, DJ94b, DJ94a, Joh02]. Here we only want to clarify a few points that will come up again in what follows.

In order to compare various estimators we will assume that (x, y) are indeed random variables with some underlying probability distribution $p_0(\mathrm{d}x, \mathrm{d}y) = p_0(\mathrm{d}x)p_0(\mathrm{d}y|x)$. It is important to stress that this distribution is conceptually distinct from the one used in inference, cf. Eq. (2.6), and that generally the two do not coincide.

For the sake of simplicity we also consider gaussian noise $w \sim \mathsf{N}(0, \sigma^2)$ with known noise level σ^2. Various estimator will be comparerd with resect to the resulting mean square error

$$\mathrm{MSE} = \mathbb{E}\{|\widehat{x}(y) - x|^2\}\,.$$

We shall then consider two cases

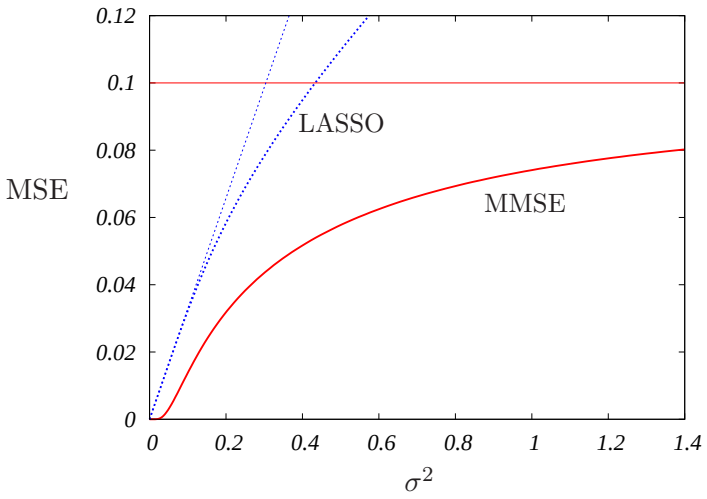

Figure 2: Mean square error for estimating a three points random variable, with probability of non-zero $\varepsilon = 0.1$, in gaussian noise. Red line: Minimal mean square error achieved by conditional expectation (thick) and its large noise asymptote (thin). Blue line: Mean square error for LASSO or equivalently for soft thresholding (thick) and its small noise asymptote (thin).

I. The signal distribution $p_0(x)$ is known as well. This can be regarded as an 'oracle' case. To make contact with compressed sensing, we shall consider distributions that generate sparse signals, i.e. that put mass at least $1-\varepsilon$ on $x = 0$. In formulae $p_0(\{0\}) \geq 1-\varepsilon$.

II. The signal distribution is unknown but it is known that it is 'sparse', namely that it belongs to the class

$$\mathcal{F}_\varepsilon \equiv \left\{ p_0 \, : \;\; p_0(\{0\}) \geq 1-\varepsilon \right\}. \tag{3.2}$$

In the first case, it is known that the minimum mean square error is achieved by the conditional expectation

$$\widehat{x}^{\rm MMSE}(y) = \int_{\mathbb{R}} x \, p_0(\mathrm{d}x|y)$$

In Figure 2 we plot the resulting MSE for a 3 point distribution

$$p_0 = \frac{\varepsilon}{2}\,\delta_{+1} + (1-\varepsilon)\,\delta_0 + \frac{\varepsilon}{2}\,\delta_{-1}\,. \tag{3.3}$$

The MMSE is non-decreasing in σ^2, converges to 0 in the noiseless limit $\sigma \to 0$ (indeed the simple rule $\widehat{x}(y) = y$ achieves MSE equal to σ^2) and to ε in the large noise limit $\sigma \to \infty$ (MSE equal to ε is achieved by $\widehat{x} = 0$).

In the more realistic situaton II, we do not know the prior. An interesting exercise (indeed not a trivial one) is to consider the LASSO estimator (2.8), which in this case reduces to

$$\widehat{x}(y;\lambda) = \mathrm{argmin}_{z\in\mathbb{R}}\Big\{\frac{1}{2}(y-z)^2 + \lambda\,|z|\Big\}\,. \tag{3.4}$$

This one dimensional optimization admits an explicit solution in terms of the *soft thresholding function* $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ defined as follows

$$\eta(y;\theta) = \begin{cases} y-\theta & \text{if } x > \theta, \\ 0 & \text{if } -\theta \le y \le \theta, \\ y+\theta & \text{otherwise.} \end{cases} \tag{3.5}$$

The *threshold* value θ has to be chosen equal to the regularization parameter λ yielding the simple solution

$$\widehat{x}(y;\lambda) = \eta(y;\theta)\,, \quad \text{for } \lambda = \theta\,. \tag{3.6}$$

(We emphasize the identity of λ and θ in the scalar case, because it breaks down in the vector case.) In Fig. 2 we plot the resulting MSE when $\theta = \alpha\sigma$, with $\alpha \approx 1.1402$.

How should the parameter θ (or equivalently λ) be fixed? The rule is conceptually simple: θ should minimize the maximal mean square error for the class $\mathcal{F}_\varepsilon$. Remarkably this complex saddle point problem can be solved rather explicitely.

Let us outline this solution. First of all, it makes sense to scale λ as the noise standard deviation, because the estimator is supposed to filter out the noise. We then let $\theta = \alpha\sigma$. We then denote the LASSO/soft thresholding mean square error by $\mathsf{mse}(\sigma^2; p_0, \alpha)$ when the noise variance is σ^2, $x \sim p_0$, and the regularization parameter is $\theta = \alpha\sigma$. The worst case mean square error is given by $\sup_{p_0 \in \mathcal{F}_\varepsilon} \mathsf{mse}(\sigma^2; p_0, \alpha)$. Since the class $\mathcal{F}_\varepsilon$ is invariant by rescaling, this worst case MSE must be proportional to the only scale in the problem, i.e. σ^2. We get

$$\sup_{p_0 \in \mathcal{F}_\varepsilon} \mathsf{mse}(\sigma^2; p_0, \alpha) = M(\varepsilon, \alpha)\sigma^2\,. \tag{3.7}$$

The function M can be computed explictly yielding

$$M(\varepsilon, \alpha) = \varepsilon\,(1+\alpha^2) + (1-\varepsilon)[2(1+\alpha^2)\,\Phi(-\alpha) - 2\alpha\,\phi(\alpha)] \tag{3.8}$$

where $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ is the gaussian density and $\Phi(z) = \int_{-\infty}^{z} \phi(u)\,\mathrm{d}u$ is the gaussian distribution. It is also not hard to show that that $M(\varepsilon, \alpha)$ is the slope of the soft thresholding MSE at $\sigma^2 = 0$ in a plot like the one in Fig. 2.

Minimizing the above espression over α, we obtain the soft thresholding minimax risk, and the corresponding optimal threshold value

$$M^{\#}(\varepsilon) \equiv \min_{\alpha \in \mathbb{R}_+} M(\varepsilon, \alpha)\,, \qquad \alpha^{\#}(\varepsilon) \equiv \arg\min_{\alpha \in \mathbb{R}_+} M(\varepsilon, \alpha)\,. \tag{3.9}$$

The functions $M^{\#}(\varepsilon)$ and $\alpha^{\#}(\varepsilon)$ are plotted in Fig. 3. For comparison we also plot the analogous functions when the class $\mathcal{F}_\varepsilon$ is replaced by $\mathcal{F}_\varepsilon(a) = \{p_0 \in \mathcal{F}_\varepsilon : \int x^2\, p_0(\mathrm{d}x) \le a^2\}$ of sparse random variables with bounded second moment. Of particular interest is the behavior of these curves in the very sparse limit $\varepsilon \to 0$

$$M^{\#}(\varepsilon) = 2\varepsilon \log(1/\varepsilon)\,\cdot\,\big\{1 + o(1)\big\}\,, \qquad \alpha^{\#}(\varepsilon) = \sqrt{2\log(1/\varepsilon)}\,\cdot\,\big\{1 + o(1)\big\}\,. \tag{3.10}$$

Getting back to Fig. 2, the reader will notice that there is a significant gap between the minimal MSE and the MSE achieved by soft-thresholding. This is the price paid by using an estimator that

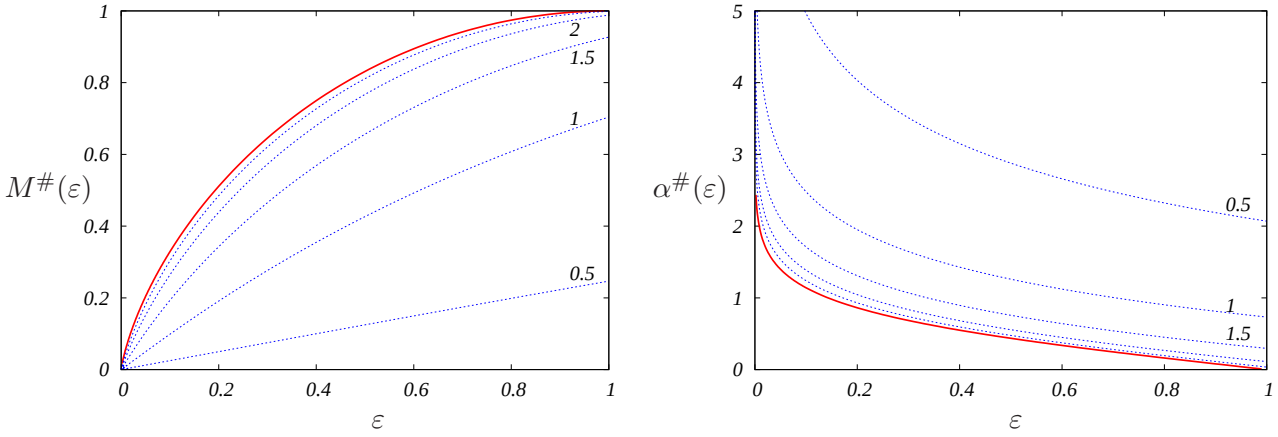

Figure 3: Left frame (red line): minimax mean square error under soft thresholding for estimation of ε-sparse random variable in gaussian noise. Blue lines corresponds to signals of bounded second moment (labels on the curves refer to the maximum allowed value of $[\int x^2\, p_0(\mathrm{d}x)]^{1/2}$). Right frame (red line): Optimal threshold level for the same estimation problem. Blue lines again refer to the case of bounded second moment.

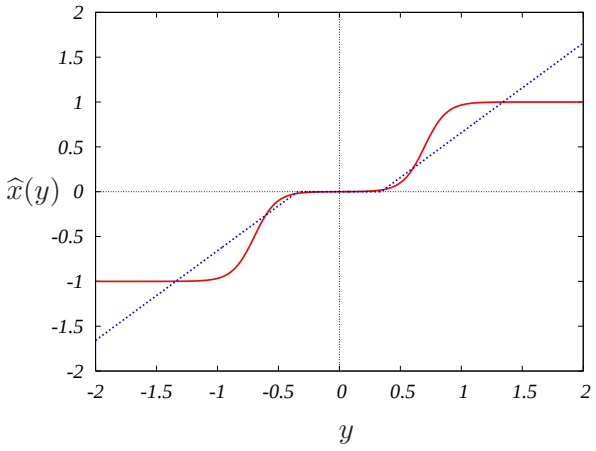

Figure 4: Red line: The MMSE estimator for the three-point distribution (3.3) with $\varepsilon = 0.1$, when the noise has standard deviation $\sigma = 0.3$. Blue line: the minimax soft threshold estimator for the same setting. The corresponding mean square errors are plotted in Fig. 2.

is *uniformly good* over the class $\mathcal{F}_\varepsilon$ instead of one that is tailored for the distribution p_0 at hand. Figure 4 compares the two estimators for $\sigma = 0.3$. One might wonder whether *all* this price has to be paid, i.e. whether we can reduce the gap by using a more sophisticate nonlinearity instead of the soft threshold $\eta(y;\theta)$. The answer is yes and no. On one hand, there exist provably superior –although more complex– minimax estimators over $\mathcal{F}_\varepsilon$. On the other, such estimators have the same minimax risk $M^\#(\varepsilon) = (2\log(1/\varepsilon))^{-1} \cdot \big\{1 + o(1)\big\}$ in the very sparse limit.

4 Inference via message passing

The task of extending the theory of the previous section to the vector case (1.1) might appear daunting. It turns hout that such extension in instead possible in specific high-dimensional limits. The key step consists in introducing an appropriate message passing algorithm to solve the optimization problem (2.8) and then analyzing its behavior.

4.1 The min-sum algorithm

We start by considering the min-sum algorithm. Min-sum is a popular optimization algorithm for graph-structured cost fuctions (see for instance [Pea88, MM09, MR07] and references therein). In order to introduce the algorithm, we consider a general cost function over $x = (x_1, \dots, x_n)$, that decomposes according to a factor graph as the one shown in Fig. 1:

$$\mathcal{C}(x) = \sum_{a\in F} \mathcal{C}_a(x_{\partial a}) + \sum_{i\in V} \mathcal{C}_i(x_i)\,. \tag{4.1}$$

Here F is the set of m *factor nodes* (squares in Fig. 1) and V is the set of n *variable nodes* (circles in the same figure). Further ∂a is the set of neighbors of node a and $x_{\partial a} = (x_i \,:\, i \in \partial a)$. The min-sum algorithm is an iterative algorithm of the belief-propagation type. Its basic variables are messages: a message is associated to each directed edge in the underlying factor graph. In the present case, messages are functions on the optimization variables, and we will denote them as $J^t_{i\to a}(x_i)$ (from variable to factor), $\widehat{J}^t_{a\to i}(x_i)$ (from factor to variable), with t indicating the iteration number. Messages are meaningful up to an additive constant, and therefore we will use the special symbol $\cong$ to denote identity up to an additive constant independent of the argument x_i. At t-th iteration they are updates as follows

$$J^{t+1}_{i\to a}(x_i) \;\cong\; \mathcal{C}_i(x_i) + \sum_{b\in\partial i\setminus a} \widehat{J}^t_{b\to i}(x_i)\,, \tag{4.2}$$

$$\widehat{J}^t_{a\to i}(x_i) \;\cong\; \min_{x_{\partial a\setminus i}} \Big\{\mathcal{C}_a(x_{\partial a}) + \sum_{j\in\partial a\setminus i} J^t_{j\to a}(x_j)\Big\}\,. \tag{4.3}$$

Eventually the optimum is approximated by

$$\widehat{x}^{t+1}_i \;=\; \arg\min_{x_i\in\mathbb{R}} J^{t+1}_i(x_i)\,, \tag{4.4}$$

$$J^{t+1}_i(x_i) \;\cong\; \mathcal{C}_i(x_i) + \sum_{b\in\partial i} \widehat{J}^t_{b\to i}(x_i) \tag{4.5}$$

There exists a vast literature justifying the use of algorithms of this type, applying them on concrete problems, and developing modifications of the basic iteration with better properties. Here we limit

ourself to recalling that the iteration (4.2), (4.3) can be regarded as a dynamic programming-like iteration that computes the minimum cost when the underlyng graph is a tree. Its application to loopy graphs is not generally guaranteed to converge.

At this point we notice that the LASSO cost function Eq. (2.8) can be decomposed as in Eq. (4.1)

$$\mathcal{C}_{A,y}(x) \equiv \tfrac{1}{2}\sum_{a\in F}(y_a - A_a^T x)^2 + \lambda \sum_{i\in V}|x_i|\,. \tag{4.6}$$

Since we will focus on dense measurement matrices, we can assume that the factor graph is the complete bipartite graph. The min-sum updates read

$$J_{i\to a}^{t+1}(x_i) \cong \lambda|x_i| + \sum_{b\in[m]\setminus a} \widehat{J}_{b\to i}^{t}(x_i)\,, \tag{4.7}$$

$$\widehat{J}_{a\to i}^{t}(x_i) \cong \min_{x_{\partial a\setminus i}} \Big\{\frac{1}{2}(y_a - A_a^T x)^2 + \sum_{j\in[n]\setminus i} J_{j\to a}^{t}(x_j)\Big\}\,. \tag{4.8}$$

4.2 Simplifying min-sum by quadratic approximation

Unfortunately, an exact implementation of the min-sum iteration appears extremely difficult because it requires to keep track of $2mn$ messages, each being a function on the real axis. A possible approach consists in developing *numerical* approximations to the messages. This line of research was initiated in [SBB10].

Here we will overview an alternative approach that consists in deriving *analytical* approximations [DMM09, DMM10a, DMM10b]. Its advantage is that it leads to a remarkably simple algorithm, which will be discussed in the next section. In order to justfy this algorithm we will first derive a simplified message passing algorithm, whose messages are simple real numbers (instead of functions), and then (in the next section) reduce the number of messages from $2mn$ to $m+n$.

Throughout the derivation we shall assume that the matrix A is normalized in such a way that its columns have zero mean and unit ℓ_2 norm. Explicitely, we have $\sum_{a=1}^n A_{ai} = 0$ and $\sum_{a=1}^n A_{ai}^2 = 1$. We also assume that its entries have roughly the same magnitude $O(1/\sqrt{n})$. These assumptions are verified by many examples of sensing matrices in compressed sensing, e.g. random matrices with i.i.d. entries or random Fourier sections. Modifications of the basic algorithm that cope with strong violations of these assumptions are discussed in Ref. [BM10a].

It is easy to see by induction that the messages $J_{i\to a}^t(x_i)$, $\widehat{J}_{a\to i}^t(x_i)$ remain, for any t, convex functions, provided they are initialized as convex functions at $t=0$. In order to simplify the min-sum equations, we will approximate them by quadratic functions. Our first step consists in noticing that, as a consequence of Eq. (4.8), the function $\widehat{J}_{a\to i}^t(x_i)$ depends on its argument only through the combination $A_{ai}x_i$. Since $A_{ai} \ll 1$, we can approximate this dependence through a Taylor expansion (without loss of generality setting $\widehat{J}_{a\to i}^t(0) = 0$):

$$\widehat{J}_{a\to i}^t(x_i) \cong -\alpha_{a\to i}^t(A_{ai}x_i) + \frac{1}{2}\beta_{a\to i}^t(A_{ai}x_i)^2 + O(A_{ai}^3 x_i^3)\,. \tag{4.9}$$

The reason for stopping this expanson at third order should become clear in a moment. Indeed substituting in Eq. (4.7) we get

$$J_{i\to a}^{t+1}(x_i) \cong \lambda|x_i| - \Big(\sum_{b\in\partial i\setminus a} A_{bi}\alpha_{b\to i}^t\Big)x_i + \frac{1}{2}\Big(\sum_{b\in\partial i\setminus a} A_{bi}^2\beta_{a\to i}^t\Big)x_i^2 + O(nA_{\cdot i}^3 x_i^3)\,. \tag{4.10}$$

Since $A_{ai} = O(1/\sqrt{n})$, the last term is negligible. At this point we want to approximate $J^t_{i\to a}$ by its second order Taylor expansion around its minimum. The reason for ths is that only this order of the expansion matters when plugging these messages in Eq. (4.8) to compute $\alpha^t_{a\to i}$, $\beta^t_{a\to i}$. We thus define the quantities $x^t_{i\to a}$, $\gamma^t_{i\to a}$ as parameters of this Taylor expansion:

$$J^t_{i\to a}(x_i) \cong \frac{1}{2\gamma^t_{i\to a}}(x_i - x^t_{i\to a})^2 + O((x_i - x^t_{i\to a})^3)\,. \tag{4.11}$$

Here we include also the case in which the minimum of $J^t_{i\to a}(x_i)$ is achieved at $x_i = 0$ (and hence the function is not differentiable at its minimum) by letting $\gamma^t_{i\to a} = 0$ in that case. Comparing Eqs. (4.10) and (4.11), and recalling the definition of $\eta(\,\cdot\,;\,\cdot\,)$, cf. Eq. (3.5), we get

$$x^{t+1}_{i\to a} = \eta(\mathsf{a}_1;\mathsf{a}_2)\,, \qquad \gamma^{t+1}_{i\to a} = \eta'(\mathsf{a}_1;\mathsf{a}_2)\,, \tag{4.12}$$

where $\eta'(\,\cdot\,;\,\cdot\,)$ denotes the derivative of η with respect to its first argument and we defined

$$\mathsf{a}_1 \equiv \frac{\sum_{b\in\partial i\setminus a} A_{bi}\alpha^t_{b\to i}}{\sum_{b\in\partial i\setminus a} A^2_{bi}\beta^t_{b\to i}}\,, \qquad \mathsf{a}_2 \equiv \frac{\lambda}{\sum_{b\in\partial i\setminus a} A^2_{bi}\beta^t_{b\to i}}\,, \tag{4.13}$$

Finally, by plugging the parametrization (4.11) in Eq. (4.8) and comparing with Eq. (4.9), we can compute the parameters $\alpha^t_{a\to i}$, $\beta^t_{a\to i}$. A long but straightforward calculation yields

$$\alpha^t_{a\to i} = \frac{1}{1+\sum_{j\in\partial a\setminus i} A^2_{aj}\gamma^t_{j\to a}}\Big\{y_a - \sum_{j\in\partial a\setminus i} A_{aj}x^t_{j\to a}\Big\}\,, \tag{4.14}$$

$$\beta^t_{a\to i} = \frac{1}{1+\sum_{j\in\partial a\setminus i} A^2_{aj}\gamma^t_{j\to a}}\,. \tag{4.15}$$

Equations (4.12) to (4.15) define a message passing algorithm that is considerably simpler than the original min-sum algorithm: each message consists of a pair of real numbers, namely $(x^t_{i\to a}, \gamma^t_{i\to a})$ for variable-to-factor messages and $(\alpha_{a\to i}, \beta_{a\to i})$ for factor-to-variable messages. In the next section we will simplify it further and construct an algorithm (AMP) with several interesting properties. Let us pause a moment for making two observations:

1. The soft-thresholding operator that played an important role in the scalar case, cf. Eq. (3), reappeared in Eq. (4.12). Notice however the threshold value that follows as a consequence of our derivation is not the naive one, namely the regularization parameter λ, but rather a renormalized one.

2. Our derivation leveraged on the assumption that the matrix entries A_{ai} are all of the same order, namely $O(1/\sqrt{n})$. It would be interesting to repeat the above derivation under different assumptions on the sensing matrix.

5 Approximate message passing

The algorithm derived above is still complex in that its memory requirements scale proportionally to the *product* of the number of dimensions of the signal and of the number of measurements. Further its complexity scales quadratically as well. In this section we will introduce a simpler algorithm, and subsequently discuss its derivation from the one in the previous section.

5.1 The AMP algorithm, some of its properties, ...

The AMP (for approximate message passing) algorithm is parameterized by two sequences of scalars: the thresholds $\{\theta_t\}_{t\geq 0}$ and the 'reaction terms' $\{\mathsf{b}_t\}_{t\geq 0}$. Starting with initial condition $x^0 = 0$, it constructs a sequence of estimates $x^t \in \mathbb{R}^N$, and residuals $r^t \in \mathbb{R}^n$, according to the following iteration

$$x^{t+1} = \eta(x^t + A^T r^t\,;\theta_t), \tag{5.1}$$

$$r^t = y - Ax^t + \mathsf{b}_t\, r^{t-1}\,, \tag{5.2}$$

for all $t \geq 0$. Here and below, given a scalar function $f : \mathbb{R} \to \mathbb{R}$, and a vector $u \in \mathbb{R}^\ell$, we adopt the convention of denoting by $f(u)$ the vector $(f(u_1), \dots, f(u_\ell))$.

The choice of parameters $\{\theta_t\}_{t\geq 0}$ and $\{\mathsf{b}_t\}_{t\geq 0}$ is tightly constrained by the connection with the min-sum algorithm, as it will be discussed below, but the connection with the LASSO is more general.

Proposition 5.1. *Let (x_*, z_*) be a fixed point of the iteration (5.1), (5.2) for $\theta_t = \theta$, $\mathsf{b}_t = \mathsf{b}$ fixed. Then (x_*, y_*) is a minimum of the* LASSO *cost function (2.8) for*

$$\lambda = \theta(1 - \mathsf{b})\,. \tag{5.3}$$

Proof. From Eq. (5.1) we get the fixed point condition

$$x + \theta v = x + A^T r\,, \tag{5.4}$$

for $v \in \mathbb{R}^n$ such that $v_i = \mathrm{sign}(x_i)$ if $x_i \neq 0$ and $v_i \in [-1, +1]$ otherwise. In other words, v is in the subgradient of the ℓ_1-norm at x, $\partial\|x\|_1$. Further from Eq. (5.2) we get $(1 - \mathsf{b})r = y - Ax$. Substituting in the above, we get

$$\theta(1 - \mathsf{b})v = A^T(y - Ax)\,,$$

which is just the stationarity condition for the LASSO cost function if $\lambda = \mathsf{b}(1 - \theta)$. □

As a consequence of this proposition, if we find sequences $\{\theta_t\}_{t\geq 0}$, $\{\mathsf{b}_t\}_{t\geq 0}$ that converge, and such that the estimates x^t converge as well, then we are guaranteed that the limit is a LASSO optimum. The connection with the message passing min-sum algorithm (see below) implies an unambiguous prescription for b_t:

$$\mathsf{b}_t = \frac{1}{m}\,\|x^t\|_0\,, \tag{5.5}$$

where $\|u\|_0$ denothes the 0 pseudo-norm of vector u, i.e. the number of its non-zero components. The choice of the sequence of thresholds $\{\theta_t\}_{t\geq 0}$ is somewhat more flexible. Recalling the discussion of the scalar case, it appears to be a good choice to use $\theta_t = \alpha\tau_t$ where $\alpha > 0$ and τ_t is the root mean square error of the un-thresholded estimate $(x^t + A^T r^t)$. It can be shown that the latter is (in an high-dimensional setting) well approximated by $(\|r^t\|^2/m)^{1/2}$. We thus obtain the prescription

$$\theta_t = \alpha\widehat{\tau}_t\,, \qquad \widehat{\tau}_t^2 = \frac{1}{m}\,\|r^t\|^2\,. \tag{5.6}$$

Alternative estimates can used to replace $\widehat{\tau}_t$, for instance using the median of $\{|z_i^t|\}_{i\in[m]}$. Explicitely, denoting by $|u|_{(\ell)}$ the ℓ-th largest magnitude among the entries of a vector u, we can use

$$\widehat{\tau}_t^2 = \frac{1}{\Phi^{-1}(3/4)}|r^t|_{(m/2)}\,, \tag{5.7}$$

with $\Phi^{-1}(3/4) \approx 0.6745$ the median of the absolute values of a gaussian random variable.

By Proposition 5.1, if the iteration converges to $(\widehat{x}, \widehat{r})$, then this is minimum of the LASSO cost function, with regularization parameter

$$\lambda = \alpha\,\frac{\|\widehat{r}\|^2}{m}\left(1 - \frac{\|\widehat{x}\|_0}{m}\right). \tag{5.8}$$

(in case the the threshold is chosen as per Eq. (5.6)). While the relation between α and λ is not fully explicit (it requires to find the optimum $\widehat{x}$), in practice α is as useful as a λ: both play the role of knobs to adjust the level of sparsity of the solution seeked.

We conclude by noting that the AMP algorithm (5.1), (5.2) is quite close to iterative soft thresholding (IST), a well known algorithm for the same problem that proceeds by

$$x^{t+1} = \eta(x^t + A^T r^t\,;\theta_t)\,, \tag{5.9}$$

$$r^t = y - Ax^t\,. \tag{5.10}$$

The only (but important) difference lies in the addition of the term $\mathsf{b}_t r^{t-1}$. This can be regarded as a momentum term with a very specific prescription on its size, cf. Eq. (5.5). A similar term –with motivations analogous to the one presented below– is popular under the name of 'Onsager term' in statistical physics.

5.2 ... and its derivation

In this section we present an heuristic derivation of the AMP iteration in Eqs. (5.1), (5.2) starting from the standard message passing formulation given by Eq. (4.12) to (4.15). Our objective is to develop an intuitive understanding of the AMP iteration, as well as of the prescription (5.5). Throughout our argument, we treat m as scaling linearly with n.

We start by noticing that the sums $\sum_{j\in\partial a\setminus i} A_{aj}^2\gamma_{j\to a}^t$ and $\sum_{b\in\partial i\setminus a} A_{bi}^2\beta_{b\to i}^t$ are sums of $\Theta(n)$ terms, each of order $1/n$ (because $A_{ai}^2 = O(1/n)$). It is reasonable to think that a law of large numbers applies and that therefore these sums can be replaced by quantities that do not depend on the instance or on the row/column index.

We then let $r_{a\to i}^t = \alpha_{a\to i}^t/\beta_{a\to i}^t$ and rewrite the message passing iteration as

$$z_{a\to i}^t = y_a - \sum_{j\in[n]\setminus i} A_{aj}x_{j\to a}^t\,, \tag{5.11}$$

$$x_{i\to a}^{t+1} = \eta\Big(\sum_{b\in[m]\setminus a} A_{bi}z_{b\to i}^t;\theta_t\Big)\,, \tag{5.12}$$

where $\theta_t \approx \lambda/\sum_{b\in\partial i\setminus a} A_{bi}^2\beta_{b\to i}^t$ is –as mentioned– treated as independent of b.

Notice that on the right-hand side of both equations above, the messages appears in sums over $\Theta(n)$ terms. Consider for instance the messages $\{z_{a\to i}^t\}_{i\in[n]}$ for a fixed node $a \in [m]$. These depend

on $i \in [n]$ only because the term excluded from the sum changes. It is therefore natural to guess that $r^t_{a\to i} = r^t_a + O(n^{-1/2})$ and $x^t_{i\to a} = x^t_i + O(m^{-1/2})$, where r^t_a only depends on the index a (and not on i), and x^t_i only depends on i (and not on a).

A naïve approximation would consist in neglecting the $O(n^{-1/2})$ correction but this approximation turns out to produce a non-vanishing error in the large-n limit. We instead set

$$z^t_{a\to i} = z^t_a + \delta z^t_{a\to i}\,, \qquad x^t_{i\to a} = x^t_i + \delta x^t_{i\to a}\,.$$

Substituting in Eq. (5.11), we get

$$\begin{aligned} z^t_a + \delta z^t_{a\to i} &= y_a - \sum_{j\in[n]} A_{aj}(x^t_j + \delta x^t_{j\to a}) + A_{ai}(x^t_i + \delta x^t_{i\to a})\,,\\ x^{t+1}_i + \delta x^{t+1}_{i\to a} &= \eta\Big(\sum_{b\in[m]} A_{bi}(z^t_b + \delta z^t_{b\to i}) - A_{ai}(z^t_a + \delta z^t_{a\to i});\,\theta_t\Big)\,. \end{aligned}$$

We will now drop the terms that are negligible without writing explicitly the error terms. First of all notice that single terms of the type $A_{ai}\delta z^t_{a\to i}$ are of order $1/n$ and can be safely neglected. Indeed $\delta z_{a\to i} = O(n^{-1/2})$ by our anzatz, and $A_{ai} = O(n^{-1/2})$ by definition. We get

$$\begin{aligned} z^t_a + \delta z^t_{a\to i} &= y_a - \sum_{j\in[n]} A_{aj}(x^t_j + \delta x^t_{j\to a}) + A_{ai}x^t_i\,,\\ x^{t+1}_i + \delta x^{t+1}_{i\to a} &= \eta\Big(\sum_{b\in[m]} A_{bi}(z^t_b + \delta z^t_{b\to i}) - A_{ai}z^t_a;\theta_t\Big)\,. \end{aligned}$$

We next expand the second equation to linear order in $\delta x^t_{i\to a}$ and $\delta z^t_{a\to i}$:

$$\begin{aligned} z^t_a + \delta z^t_{a\to i} &= y_a - \sum_{j\in[n]} A_{aj}(x^t_j + \delta x^t_{j\to a}) + A_{ai}x^t_i\,,\\ x^{t+1}_i + \delta x^{t+1}_{i\to a} &= \eta\Big(\sum_{b\in[m]} A_{bi}(z^t_b + \delta z^t_{b\to i});\theta_t\Big) - \eta'\Big(\sum_{b\in[m]} A_{bi}(z^t_b + \delta z^t_{b\to i});\theta_t\Big)A_{ai}z^t_a\,. \end{aligned}$$

Notice that the last term on the right hand side of the first equation is the only one dependent on i, and we can therefore identify this term with $\delta z^t_{a\to i}$. We obtain the decomposition

$$z^t_a = y_a - \sum_{j\in[n]} A_{aj}(x^t_j + \delta x^t_{j\to a})\,, \tag{5.13}$$

$$\delta z^t_{a\to i} = A_{ai}x^t_i\,. \tag{5.14}$$

Analogously for the second equation we get

$$x^{t+1}_i = \eta\Big(\sum_{b\in[m]} A_{bi}(z^t_b + \delta z^t_{b\to i});\theta_t\Big)\,, \tag{5.15}$$

$$\delta x^{t+1}_{i\to a} = -\eta'\Big(\sum_{b\in[n]} A_{bi}(z^t_b + \delta z^t_{b\to i});\theta_t\Big)A_{ai}z^t_a\,. \tag{5.16}$$

Substituting Eq. (5.14) in Eq. (5.15) to eliminate $\delta z^t_{b\to i}$ we get

$$x_i^{t+1} = \eta\Big(\sum_{b\in[n]} A_{bi} z_b^t + \sum_{b\in[n]} A_{bi}^2 x_i^t; \theta_t\Big)\,, \tag{5.17}$$

and using the normalization of A, we get $\sum_{b\in[m]} A_{bi}^2 \to 1$, whence

$$x^{t+1} = \eta(x^t + A^T r^t; \theta_t)\,. \tag{5.18}$$

Analogously substituting Eq. (5.16) in (5.13), we get

$$z_a^t = y_a - \sum_{j\in[n]} A_{aj} x_j^t + \sum_{j\in[n]} A_{aj}^2 \eta'(x_j^{t-1} + (A^T r^{t-1})_j; \theta_{t-1}) z_a^{t-1}\,. \tag{5.19}$$

Again, using the law of large numbers and the normalization of A, we get

$$\sum_{j\in[n]} A_{aj}^2 \eta'(x_j^{t-1} + (A^T r^{t-1})_j; \theta_{t-1}) \approx \frac{1}{m}\sum_{j\in[n]} \eta'(x_j^{t-1} + (A^T r^{t-1})_j; \theta_{t-1}) = \frac{1}{m}\|x^t\|_0\,, \tag{5.20}$$

whence substituting in (5.19), we obtain Eq. (5.2), withe the prescription (5.5) for the Onsager term. This finishes our derivation.

6 High-dimensional analysis

The AMP algorithm enjoys several unique properties. In particular it admits an *asymptotically exact* analysis along sequences of instances of diverging size. This is quite remarkable, since all analysis available for other algorithms that solve the LASSO hold only 'up to undetermined constants'.

In particular in the large system limit (and with the exception of a 'phase transition' line), AMP can be shown to converge exponentially fast to the LASSO optimum. Hence the analysis of AMP yields asymptotically exact predictions on the behavior of the LASSO, including in particular the asymptotic mean square error per variable.

How is this small miracle possible? Figure 5 illustrates the key point. It shows the distribution of un-thresholded estimates $(x^t + A^T r^t)_i$ for coordinates i such that the original signal had value $x_i = +1$. These estimates where were obtained using the AMP algorithm (5.1), (5.2) with choice (5.5) of b_t (plot on the left) and the iterative soft thresholding algorithm (5.9), (5.10) (plot on the right). The same instances (i.e. the same matrices A and measurement vectors y) were used in the two cases, but the resulting distributions are dramatically different. In the case of AMP, the distribution is close to gaussian, with mean on the correct value, $x_i = +1$. For iterative soft thresholding the estimates do not have the correct mean and are not gaussian.

As we will see in the next sections, these empirical observations can be confirmed rigorously in the limit of a large number of dimensions.

6.1 State evolution

We will consider sequences of instances of increasing sizes, along which the AMP algorithm behavior admits a non-trivial limit. While rigorous results have been proved so far only in the case in which the sensing matrices A have i.i.d. gaussian entries, it is nevertheless useful to collect a few basic properties that the sequence needs to satisfy.

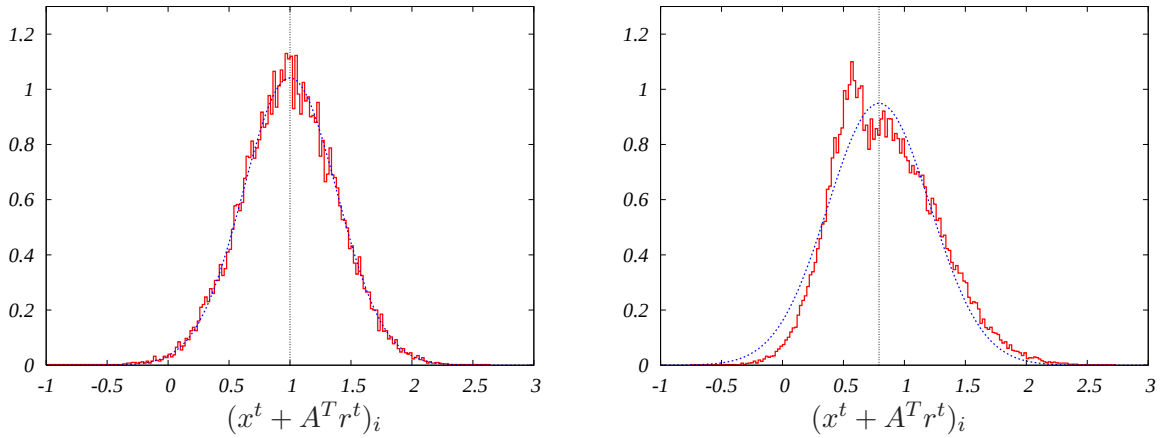

Figure 5: Distributions of un-thresholded estimates for AMP (left) and IST (right), after $t = 10$ iterations. These data were obtained using sensing matrices with $m = 2000$, $n = 4000$ and i.i.d. entries uniform in $\{+1/\sqrt{m}, -1/\sqrt{m}\}$. The signal x contained 500 non-zero entries uniform in $\{+1, -1\}$. A total of 40 instances was used to build the histograms. Blue lines are gaussian fits and vertical lines represent the fitted mean.

Definition 1. *The sequence of instances $\{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$ indexed by n is said to be a* converging sequence *if $x(n) \in \mathbb{R}^n$, $w(n) \in \mathbb{R}^m$, $A(n) \in \mathbb{R}^{m\times n}$ with $m = m(n)$ is such that $m/n \to \delta \in (0, \infty)$, and in addition the following conditions hold:*

(a) *The empirical distribution of the entries of $x(n)$ converges weakly to a probability measure p_0 on $\mathbb{R}$ with bounded second moment. Further $n^{-1}\sum_{i=1}^n x_i(n)^2 \to \mathbb{E}_{p_0}\{X_0^2\}$.*

(b) *The empirical distribution of the entries of $w(n)$ converges weakly to a probability measure p_W on $\mathbb{R}$ with bounded second moment. Further $m^{-1}\sum_{i=1}^m w_i(n)^2 \to \mathbb{E}_{p_W}\{W^2\}$.*

(c) *If $\{e_i\}_{1\le i\le n}$, $e_i \in \mathbb{R}^n$ denotes the standard basis, then $\max_{i\in[n]} \|A(n)e_i\|_2$, $\min_{i\in[n]} \|A(n)e_i\|_2 \to 1$, as $n \to \infty$ where $[n] \equiv \{1, 2, \dots, n\}$.*

As mentioned above, rigorous results have been proved only for a subclass of converging sequences, namely under the assumption that the matrices $A(n)$ have i.i.d. gaussian entries. However, numerical simulations show that the same limit behavior should apply within a much broader domain, including for instance random matrices with i.i.d. entries under an appropriate moment condition. This *universality* phenomenon is well-known in random matrix theory whereby asymptotic results initially estabilished for gaussian matrices where subsequently proved for a broad universality class. Rigorous evidence in this direction is presented in [KM10] where the normalized cost $\mathcal{C}(\widehat{x})/N$ is shown to have a limit as $N \to \infty$ which is universal with respect to random matrices A with iid entries. (More precisely, it is universal provided $\mathbb{E}\{A_{ij}\} = 0$, $\mathbb{E}\{A_{ij}^2\} = 1/n$ and $\mathbb{E}\{A_{ij}^6\} \le C/n^3$ for some uniform constant C.)

For a converging sequence of instances, and an arbitrary sequence of thresholds $\{\theta_t\}_{t\ge 0}$ (independent of n), the AMP iteration (5.1), (5.2) admits an high-dimensional limit which can be characterized exactly, provided Eq. (5.5) is used for fixing the Onsager term. This limit is given in

terms of the trajectory of a simple one-dimensional iteration termed *state evolution* which we will describe next.

Define the sequence $\{\tau_t^2\}_{t\geq 0}$ by setting $\tau_0^2 = \sigma^2 + \mathbb{E}\{X_0^2\}/\delta$ (for $X_0 \sim p_0$ and $\sigma^2 \equiv \mathbb{E}\{W^2\}$, $W \sim p_W$) and letting, for all $t \geq 0$:

$$\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \theta_t)\,, \tag{6.1}$$

$$\mathsf{F}(\tau^2, \theta) \equiv \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\{\,[\eta(X_0 + \tau Z; \theta) - X_0]^2\}\,, \tag{6.2}$$

where $Z \sim \mathsf{N}(0,1)$ is independent of X_0. Notice that the function F depends implicitly on the law p_0.

We say a function $\psi : \mathbb{R}^k \to \mathbb{R}$ is *pseudo-Lipschitz* if there exist a constant $L > 0$ such that for all $x, y \in \mathbb{R}^k$: $|\psi(x) - \psi(y)| \leq L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2$. (This is a special case of the definition used in [BM10b] where such a function is called pseudo-Lipschitz *of order 2*.)

The following theorem that was conjectured in [DMM09] and proved in [BM10b]. It shows that the behavior of AMP can be tracked by the above state evolution recursion.

Theorem 6.1 ([BM10b])**.** *Let $\{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$ be a converging sequence of instances with the entries of $A(n)$ iid normal with mean 0 and variance $1/m$. Let $\psi_1 : \mathbb{R} \to \mathbb{R}$, $\psi_2 : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz functions. Then, almost surely*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi_1\big(z_i^t\big) = \mathbb{E}\Big\{\psi_1\big(\tau_t Z\big)\Big\}\,, \tag{6.3}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi_2\big(x_i^{t+1}, x_i\big) = \mathbb{E}\Big\{\psi_2\big(\eta(X_0 + \tau_t Z; \theta_t), X_0\big)\Big\}\,, \tag{6.4}$$

where $Z \sim \mathsf{N}(0,1)$ is independent of $X_0 \sim p_0$.

Notice that this theorem holds for any choice of the sequence of thresholds $\{\theta_t\}$ and does not require –for instance– that the latter converge. Indeed [BM10b] proves a more general result that holds for a any choice of nonlinearities $\eta(\,\cdot\,;\,\cdot,)$ (not just soft-thresholding), under mild regularity assumptions, provided the AMP iteration is suitably modified.

Also, this theorem motivate both the use of soft thresholding, and the choice of the threshold level in Eq. (5.6) or (5.7). Indeed Eq. (6.3) states that the components of r^t are approximately i.i.d. $\mathsf{N}(0, \tau_t^2)$, and hence both definitions of $\widehat{\tau}_t$ in Eq. (5.6) or (5.7) provide consistent estimators of τ_t. Further, Eq. (6.3) implies that the components of the deviation $(x^t + A^T r^t - x)$ are also approximately i.i.d. $\mathsf{N}(0, \tau_t^2)$. In other words, the estimate $(x^t + A^T r^t)$ is equal to the actual signal plus noise of variance τ_t^2, as illustrated in Fig. 5. According to our discussion of scalar estimation in Section 3, the correct way of reducing the noise is to apply soft thresholding with threshold level $\alpha\tau_t$.

The choice $\theta_t = \alpha\tau_t$ with α fixed has another important advantage. In this case, the sequence $\{\tau_t\}_{t\geq 0}$ is determined by the one-dimensional homogeneous recursion

$$\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \alpha\tau_t)\,. \tag{6.5}$$

The function $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$ depends on the distribution of X_0 as well as on the other parameters of the problem. An example is plotted in Fig. (6). It turns out that the behavior shown here is generic: the function is always non-decreasing and concave. This remark allows to easily prove the following.

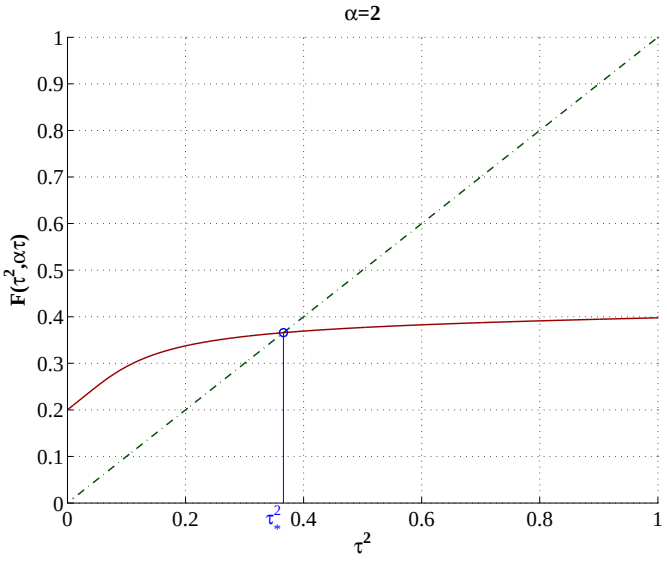

Figure 6: Mapping $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$ for $\alpha = 2$, $\delta = 0.64$, $\sigma^2 = 0.2$, $p_0(\{+1\}) = p_0(\{-1\}) = 0.064$ and $p_0(\{0\}) = 0.872$.

Proposition 6.2 ([DMM10b]). *Let $\alpha_{\min} = \alpha_{\min}(\delta)$ be the unique non-negative solution of the equation*

$$(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \frac{\delta}{2}\,, \tag{6.6}$$

with $\phi(z) \equiv e^{-z^2/2}/\sqrt{2\pi}$ the standard gaussian density and $\Phi(z) \equiv \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x$.

For any $\sigma^2 > 0$, $\alpha > \alpha_{\min}(\delta)$, the fixed point equation $\tau^2 = \mathsf{F}(\tau^2, \alpha\tau)$ admits a unique solution. Denoting by $\tau_* = \tau_*(\alpha)$ this solution, we have $\lim_{t\to\infty} \tau_t = \tau_*(\alpha)$.

It can also be shown that, under the choice $\theta_t = \alpha\tau_t$, convergence is exponentially fast unless the problem parameters take some 'exceptional' values (namely on the phase transition boundary discussed below).

6.2 The risk of the LASSO

State evolution provides a scaling limit of the AMP dynamics in the high-dimensional setting. By showing that AMP converges to the LASSO estimator, one can transfer this information to a scaling limit result of the LASSO estimator itself.

Before stating the limit, we have to describe a *calibration* mapping between the AMP parameter α (that defines the sequence of thresholds $\{\theta_t\}_{t\geq 0}$) and the LASSO regularization parameter λ. The connection was first introduced in [DMM10b].

We define the function $\alpha \mapsto \lambda(\alpha)$ on $(\alpha_{\min}(\delta), \infty)$, by

$$\lambda(\alpha) \equiv \alpha\tau_* \left[1 - \frac{1}{\delta}\mathbb{P}\big\{|X_0 + \tau_* Z| \geq \alpha\tau_*\big\}\right]\,, \tag{6.7}$$

where $\tau_* = \tau_*(\alpha)$ is the state evolution fixed point defined as per Proposition 6.2. Notice that this relation corresponds to the scaling limit of the general relation (5.3), provided we assume that the solution of the LASSO optimization problem (2.8) is indeed described by the fixed point of state evolution (equivalently, by its $t \to \infty$ limit). This follows by noting that $\theta_t \to \alpha\tau_*$ and that $\|x\|_0/n \to \mathbb{E}\{\eta'(X_0 + \tau_* Z; \alpha\tau_*)\}$. While this is just an interpretation of the definition (6.7), the result presented next implies that the interpretation is indeed correct.

In the following we will need to invert the function $\alpha \mapsto \lambda(\alpha)$. We thus define $\alpha : (0, \infty) \to (\alpha_{\min}, \infty)$ in such a way that

$$\alpha(\lambda) \in \big\{ a \in (\alpha_{\min}, \infty) \, : \, \lambda(a) = \lambda \big\} .$$

The fact that the right-hand side is non-empty, and therefore the function $\lambda \mapsto \alpha(\lambda)$ is well defined, is part of the main result of this section.

Theorem 6.3. *Let $\{x(n), w(n), A(n)\}_{n \in \mathbb{N}}$ be a converging sequence of instances with the entries of $A(n)$ iid normal with mean 0 and variance $1/m$. Denote by $\widehat{x}(\lambda)$ the* LASSO *estimator for instance $(x(n), w(n), A(n))$, with $\sigma^2, \lambda > 0$, and let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, almost surely*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi\big(\widehat{x}_i, x_i\big) = \mathbb{E}\Big\{ \psi\big(\eta(X_0 + \tau_* Z; \theta_*), X_0\big) \Big\} , \tag{6.8}$$

where $Z \sim \mathsf{N}(0,1)$ is independent of $X_0 \sim p_0$, $\tau_* = \tau_*(\alpha(\lambda))$ and $\theta_* = \alpha(\lambda)\tau_*(\alpha(\lambda))$.

Further, the function $\lambda \mapsto \alpha(\lambda)$ is well defined and unique on $(0, \infty)$.

The assumption of a converging problem-sequence is important for the result to hold, while the hypothesis of gaussian measurement matrices $A(n)$ is necessary for the proof technique to be applicable. On the other hand, the restrictions $\lambda, \sigma^2 > 0$, and $\mathbb{P}\{X_0 \neq 0\} > 0$ (whence $\tau_* \neq 0$ using Eq. (6.7)) are made in order to avoid technical complications due to degenerate cases. Such cases can be resolved by continuity arguments.

Let us now discuss some limitations of this result. Theorem 6.3 assumes that the entries of matrix A are iid gaussians. Further, our result is asymptotic, while and one might wonder how accurate it is for instances of moderate dimensions.

Numerical simulations were carried out in [DMM10b, BBM10] and suggest that the result is universal over a broader class of matrices and that is relevant already for n of the order of a few hundreds. As an illustration, we present in Figs. 7 and 8 the outcome of such simulations for two types of random matrices. Simulations with real data can be found in [BBM10]. We generated the signal vector randomly with entries in $\{+1, 0, -1\}$ and $\mathbb{P}(x_{0,i} = +1) = \mathbb{P}(x_{0,i} = -1) = 0.064$. The noise vector w was generated by using i.i.d. $\mathsf{N}(0, 0.2)$ entries.

We solved the LASSO problem (2.8) and computed estimator $\widehat{x}$ using `CVX`, a package for specifying and solving convex programs [GB10] and `OWLQN`, a package for solving large-scale versions of LASSO [AJ07]. We used several values of λ between 0 and 2 and N equal to 200, 500, 1000, and 2000. The aspect ratio of matrices was fixed in all cases to $\delta = 0.64$. For each case, the point (λ, MSE) was plotted and the results are shown in the figures. Continuous lines corresponds to the asymptotic prediction by Theorem 6.3 for $\psi(a, b) = (a - b)^2$, namely

$$\lim_{n \to \infty} \frac{1}{n} \|\widehat{x} - x\|^2 = \mathbb{E}\big\{ \big[\eta(X_0 + \tau_* Z; \theta_*) - X_0\big]^2 \big\} = \delta(\tau_*^2 - \sigma^2) .$$

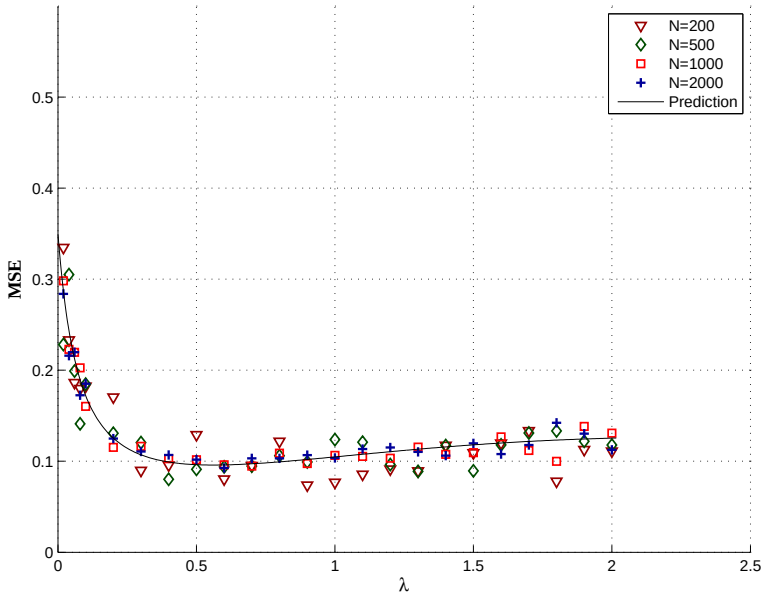

Figure 7: Mean square error (MSE) as a function of the regularization parameter λ compared to the asymptotic prediction for $\delta = 0.64$ and $\sigma^2 = 0.2$. Here the measurement matrix A has iid $\mathsf{N}(0, 1/m)$ entries. Each point in this plot is generated by finding the LASSO predictor $\widehat{x}$ using a measurement vector $y = Ax + w$ for an independent signal vector x_0, an independent noise vector w, and an independent matrix A.

The agreement is remarkably good already for n, m of the order of a few hundreds, and deviations are consistent with statistical fluctuations.

The two figures correspond to different entries distributions: (i) Random gaussian matrices with aspect ratio δ and iid $\mathsf{N}(0, 1/m)$ entries (as in Theorem 6.3); (ii) Random ± 1 matrices with aspect ratio δ. Each entry is independently equal to $+1/\sqrt{m}$ or $-1/\sqrt{m}$ with equal probability.

Notice that the asymptotic prediction has a minimum as a function of λ. The location of this minimum can be used to select the regularization parameter.

6.3 A decoupling principle

There exists a suggestive interpretation of the state evolution result in Theorem 6.1, as well as of the scaling limit of the LASSO estabilished in Theorem 6.3. *The estimation problem in the vector model $y = Ax + w$ reduces –asymptotically– to n uncoupled scalar estimation problems $\widetilde{y}_i = x_i + \widetilde{w}_i$.* However the noise variance is increased from σ^2 to τ_t^2 (or ${\tau_*}^2$ in the case of the LASSO), due to 'interference' between the original coordinates:

$$y = Ax + w \qquad \Leftrightarrow \qquad \begin{cases} \widetilde{y}_1 = x_1 + \widetilde{w}_1 \\ \widetilde{y}_2 = x_2 + \widetilde{w}_2 \\ \vdots \\ \widetilde{y}_n = x_n + \widetilde{w}_n \end{cases} . \tag{6.9}$$

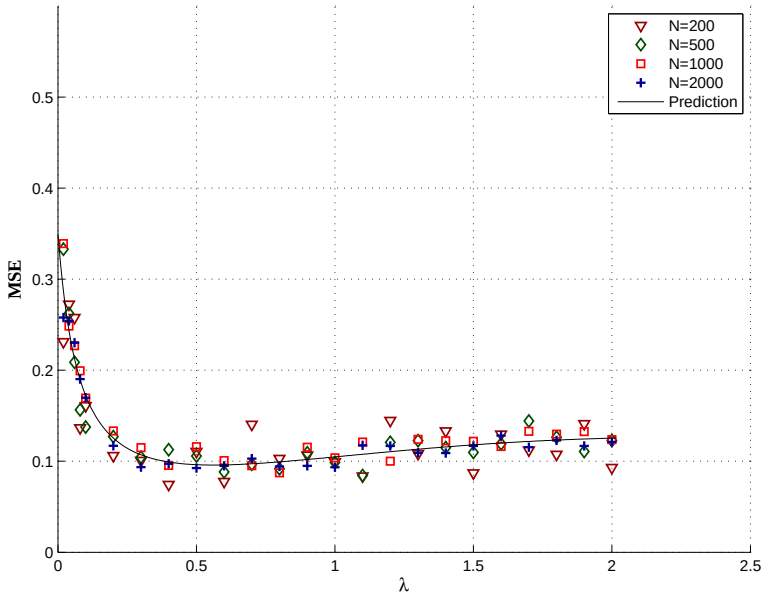

Figure 8: As in Fig. 7, but the measurement matrix A has iid entries that are equal to $\pm 1/\sqrt{m}$ with equal probabilities.

An analogous phenomenon is well known in statistical physics and probability theory and takes sometimes the name of 'correlation decay' [Wei05, GK06, MM09]. In the context of CDMA system analysis via replica method, the same phenomenon was also called 'decoupling principle' [Tan02, GV05].

Notice that the AMP algorithm gives a precise realization of this decoupling principle, since for each $i \in [n]$, and for each number of iterations t, it produces an estimate, namely $(x^t + A^T r^t)_i$ that can be considered a realization of the observation $\widetilde{y}_i$ above. Indeed Theorem 6.1 (see also discussion below the theorem) states that $(x^t + A^T r^t)_i = x_i + \widetilde{w}_i$ with $\widetilde{w}_i$ asymptotically gaussian with mean 0 and variance τ_t^2.

The fact that observations of distinct coordinates are asymptotically decoupled is stated precisely below.

Corollary 6.4 (Decoupling principle, [BM10b])**.** *Under the assumption of Theorem 6.1, fix $\ell \geq 2$, let $\psi : \mathbb{R}^{2\ell} \to \mathbb{R}$ be any Lipschitz function, and denote by E expectation with respect to a uniformly random subset of distinct indices $J(1), \dots, J(\ell) \in [n]$.*

Further, for some fixed $t > 0$, let $\widetilde{y}^t = x^t + A^T r^t \in \mathbb{R}^n$. Then, almost surely

$$\lim_{n\to\infty} \mathsf{E}\psi(\widetilde{y}^t_{J(1)}, \dots, \widetilde{y}^t_{J(\ell)}, x_{J(1)}, \dots, x_{J(\ell)}) = \mathbb{E}\big\{\psi\big(X_{0,1} + \tau_t Z_1, \dots, X_{0,\ell} + \tau_t Z_\ell, X_{0,1}, \dots, X_{0,\ell}\big)\big\},$$

for $X_{0,i} \sim p_0$ and $Z_i \sim \mathsf{N}(0,1)$, $i = 1, \dots, \ell$ mutually independent.

6.4 An heuristic derivation of state evolution

The state evolution recursion has a simple heuristic description, that is useful to present here since it clarifies the difficulties involved in the proof. In particular, this description brings up the key role played by the 'Onsager term' appearing in Eq. (5.2) [DMM09].

Consider again the recursion (5.1), (5.2) but introduce the following three modifications: (*i*) Replace the random matrix A with a new independent copy $A(t)$ at each iteration t; (*ii*) Correspondingly replace the observation vector y with $y^t = A(t)x_0 + w$; (*iii*) Eliminate the last term in the update equation for r^t. We thus get the following dynamics:

$$x^{t+1} = \eta(A(t)^T r^t + x^t; \theta_t)\,, \tag{6.10}$$

$$r^t = y^t - A(t)x^t\,, \tag{6.11}$$

where $A(0), A(1), A(2), \dots$ are iid matrices of dimensions $m \times n$ with i.i.d. entries $A_{ij}(t) \sim \mathsf{N}(0, 1/m)$. (Notice that, unlike in the rest of the article, we use here the argument of A to denote the iteration number, and not the matrix dimensions.)

This recursion is most conveniently written by eliminating r^t:

$$\begin{aligned} x^{t+1} &= \eta\big(A(t)^T y^t + (\mathbf{I} - A(t)^T A(t))x^t; \theta_t\big)\,, \\ &= \eta\big(x + A(t)^T w + B(t)(x^t - x); \theta_t\big)\,, \end{aligned} \tag{6.12}$$

where we defined $B(t) = \mathbf{I} - A(t)^* A(t) \in \mathbb{R}^{n\times n}$. Notice that this recursion does not correspond to any concrete algorithm, since the matrix A changes from iteration to iteration. It is nevertheless useful for developing intuition.

Using the central limit theorem, it is easy to show that each entry of $B(t)$ is approximately normal, with zero mean and variance $1/m$. Further, distinct entries are approximately pairwise independent. Therefore, if we let $\widetilde{\tau}_t^2 = \lim_{N\to\infty} \|x^t - x\|^2/n$, we obtain that $B(t)(x^t - x)$ converges to a vector with iid normal entries with 0 mean and variance $n\widetilde{\tau}_t^2/m = \widetilde{\tau}_t^2/\delta$. Notice that this is true because $A(t)$ is independent of $\{A(s)\}_{1\le s\le t-1}$ and, in particular, of $(x^t - x)$.

Conditional on w, $A(t)^T w$ is a vector of iid normal entries with mean 0 and variance $(1/m)\|w\|^2$ which converges by the law of large numbers to σ^2. A slightly longer exercise shows that these entries are approximately independent from the ones of $B(t)(x^t - x_0)$. Summarizing, each entry of the vector in the argument of η in Eq. (6.12) converges to $X_0 + \tau_t Z$ with $Z \sim \mathsf{N}(0,1)$ independent of X_0, and

$$\begin{aligned} \tau_t^2 &= \sigma^2 + \frac{1}{\delta}\widetilde{\tau}_t^2\,, \\ \widetilde{\tau}_t^2 &= \lim_{n\to\infty} \frac{1}{n}\|x^t - x\|^2\,. \end{aligned} \tag{6.13}$$

On the other hand, by Eq. (6.12), each entry of $x^{t+1} - x$ converges to $\eta(X_0 + \tau_t\, Z; \theta_t) - X_0$, and therefore

$$\widetilde{\tau}_{t+1}^2 = \lim_{n\to\infty} \frac{1}{n}\|x^{t+1} - x\|^2 = \mathbb{E}\big\{[\eta(X_0 + \tau_t\, Z; \theta_t) - X_0]^2\big\}\,. \tag{6.14}$$

Using together Eq. (6.13) and (6.14) we finally obtain the state evolution recursion, Eq. (6.1).

We conclude that state evolution would hold if the matrix A was drawn independently from the same gaussian distribution at each iteration. In the case of interest, A does not change across

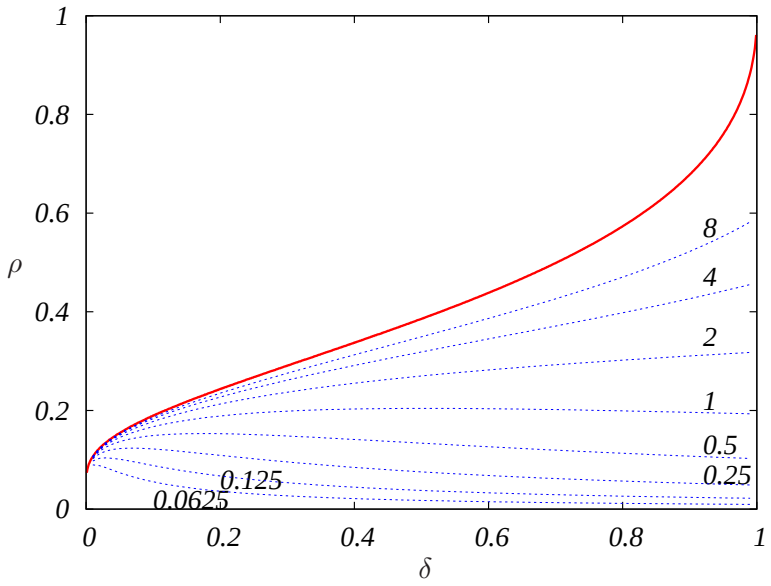

Figure 9: Noise sensitivity phase transition in the plane (δ, ρ) (here $\delta = m/n$ is the undersampling ratio and $\rho = \|x\|_0/m$ is the number of non-zero coefficients per measurement). Red line: The phase transition boundary $\rho = \rho_c(\delta)$. Blue lines: Level curves for the LASSO minimax $M^*(\delta, \rho)$. Notice that $M^*(\delta, \rho) \uparrow \infty$ as $\rho \uparrow \rho_c(\delta)$.

iterations, and the above argument falls apart because x^t and A are dependent. This dependency is non-negligible even in the large system limit $n \to \infty$. This point can be clarified by considering the IST alforithm fiven by Eqs. (5.9), (5.10). Numerical studies of iterative soft thresholding [MD10, DMM09] show that its behavior is dramatically different from the one of AMP and in particular *state evolution does not hold for IST*, even in the large system limit.

This is not a surprise: the correlations between A and x^t simply cannot be neglected. On the other hand, adding the Onsager term leads to an asymptotic cancelation of these correlations. As a consequence, state evolution holds for the AMP iteration.

6.5 The noise sensitivity phase transition

The formalism developed so far allows to extend the minimax analysis carried out in the scalar case in Section 3 to the vector estimation problem [DMM10b]. We define the LASSO mean square error per coordinate when the empirical distribution of the signal converges to p_0, as

$$\mathsf{MSE}(\sigma^2; p_0, \lambda) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\{\|\widehat{x}(\lambda) - x\|^2\}\,, \tag{6.15}$$

where the limit is taken along a converging sequence. This quantity can be computed using Theorem 6.3 for any specific distribution p_0.

We consider again the sparsity class $\mathcal{F}_\varepsilon$ with $\varepsilon = \rho\delta$. Hence $\rho = \|x\|_0/m$ measures the number of non-zero coordinates per measurement. Taking the worst case MSE over this class, and then the

minimum over the regularization parameter λ, we get a result that depends on ρ, δ, as well as on the noise level σ^2. The dependence on σ^2 must be linear because the class $\mathcal{F}_{\rho\delta}$ is scale invariant, and we obtain therefore

$$\inf_{\lambda} \sup_{p_0 \in \mathcal{F}_{\rho\delta}} \mathsf{MSE}(\sigma^2; p_0, \lambda) = M^*(\delta, \rho)\, \sigma^2\,, \tag{6.16}$$

for some function $(\delta, \rho) \mapsto M^*(\delta, \rho)$. We call this the LASSO minimax risk. It can be interpreted as the sensitivity (in terms of mean square error) of the LASSO estimator to noise in the measurements.

It ids clear that the prediction for $\mathsf{MSE}(\sigma^2; p_0, \lambda)$ provided by Theorem 6.3 can me used to characterize the LASSO minimax risk. What is remarkable is that the resulting formula is so simple.

Theorem 6.5 ([DMM10b])**.** *Assume the hypotheses of Theorem 6.3, and recall that $M^{\#}(\varepsilon)$ denotes the soft thresholding minimax risk over the class $\mathcal{F}_\varepsilon$. Further let $\rho_{\mathrm{c}}(\delta)$ be the unique solution of $\rho = M^{\#}(\rho\delta)$.*

Then for any $\rho < \rho_{\mathrm{c}}(\delta)$ the LASSO minimax risk is bounded and given by

$$M^*(\delta, \rho) = \frac{M^{\#}(\rho\delta)}{1 - M^{\#}(\rho\delta)/\delta}\,. \tag{6.17}$$

Viceversa, for any $\rho \geq \rho_{\mathrm{c}}(\delta)$, $M^*(\delta, \rho) = \infty$.

Figure 9 shows the location of the noise sensitivity boundary $\rho_{\mathrm{c}}(\delta)$ as well as the level lines of $M^*(\delta, \rho)$ for $\rho < \rho_{\mathrm{c}}(\delta)$. Above $\rho_{\mathrm{c}}(\delta)$ the LASSO MSE is not uniormly bounded in terms of the measurement noise σ^2. Other estimators (for instance one step of soft thresholding) can offer better stability guarantees in this region.

One remarkable fact is that the phase boundary $\rho = \rho_{\mathrm{c}}(\delta)$ coincides with the phase transition for $\ell_0 - \ell_1$ equivalence derived earlier by Donoho on the basis of random polytope geometry results by Affentranger-Schneider. The same phase transition was further studied in a series of papers by Donoho, Tanner and coworkers, in connection with the noiseless estimation problem. For $\rho < \rho_{\mathrm{c}}$ estimating x by ℓ_1-norm minimization returns the correct signal with high probability (over the choice of the random matrix A). For $\rho > \rho_{\mathrm{c}}(\delta)$, ℓ_1-minimization fails.

Here this phase transition is derived from a completely different perspective, and using a new method –the state evolution analysis of the AMP algorithm– which offers quantitative information about the noisy case as well i.e. to compute the value of $M^*(\delta, \rho)$ for $\rho < \rho_{\mathrm{c}}(\delta)$. Within the present approach, the line $\rho_{\mathrm{c}}(\delta)$ admits a very simple expresson. In parametric form, it is given by

$$\delta = \frac{2\phi(\alpha)}{\alpha + 2(\phi(\alpha) - \alpha\Phi(-\alpha))}\,, \tag{6.18}$$

$$\rho = 1 - \frac{\alpha\Phi(-\alpha)}{\phi(\alpha)}\,, \tag{6.19}$$

where ϕ and Φ are the gaussian density and gaussian distribution function, and $\alpha \in [0, \infty)$ is the parameter. Indeed α has a simple and practically important interpretation as well. Recall that the AMP algorithm uses a sequence of thresholds $\theta_t = \alpha\widehat{\tau}_t$, cf. Eqs. (5.6) and (5.7). How should the parameter α be fixed? A very simple prescription is obtained in the noiseless case. In order to achieve exact reconstruction for all $\rho < \rho_{\mathrm{c}}(\delta)$ for a given an undersampling ratio δ, α should be such that $(\delta, \rho_{\mathrm{c}}(\delta)) = (\delta(\alpha), \rho(\alpha))$ with functions $\alpha \mapsto \delta(\alpha)$, $\alpha \mapsto \rho(\alpha)$ defined as in Eq. (6.18), (6.19). In other words, this parametric expression yields each point of the phase boundary as a function of the threshold parameter used to achieve it via AMP.

6.6 Comparison with other analysis approaches

The analysis presented here is significantly different from for more standard approaches. We derived an *exact* characterization for the high-dimensional limit of the LASSO estimation problem under the assumption of converging sequences of random sensing matrices.

Alternative approaches assume an appropriate 'isometry', or 'incoherence' condition to hold for A. Under this condition upper bounds are proved for the mean square error. For instance Candes, Romberg and Tao [CRT06] prove that the mean square error is bounded by $C\sigma^2$ for some constant C. Work by Candes and Tao [CT07] on the analogous *Dantzig selector*, upper bounds the mean square error by $C\sigma^2(k/n)\log n$, with k the number of non-zero entries of the signal x.

These type of results are very robust but present two limitations: (*i*) They do not allow to distinguish reconstruction methods that differ by a constant factor (e.g. two different values of λ); (*ii*) The restricted isometry condition (or analogous ones) is quite restrictive. For instance, it holds for random matrices only under very strong sparsity assumptions. These restrictions are intrinsic to the worst-case point of view developed in [CRT06, CT07].

Guarantees have been proved for correct support recovery in [ZY06], under an incoherence assumption on A. While support recovery is an interesting conceptualization for some applications (e.g. model selection), the metric considered in the present paper (mean square error) provides complementary information and is quite standard in many different fields.

Close to the spirit of the treatment presented here, [RFG09] derived expressions for the mean square error under the same model considered here. Similar results were presented recently in [KWT09, GBS09]. These papers argue that a sharp asymptotic characterization of the LASSO risk can provide valuable guidance in practical applications. Unfortunately, these results were non-rigorous and were obtained through the famously powerful 'replica method' from statistical physics [MM09]. The approach discussed here offers two advantages over these recent developments: (*i*) It is completely *rigorous*, thus putting on a firmer basis this line of research; (*ii*) It is *algorithmic* in that the LASSO mean square error is shown to be equivalent to the one achieved by a low-complexity message passing algorithm.

7 Structured priors and more general regressions

The single most important advantage of the point of view based on graphical models is that it offers a unified disciplined approach to exploit structural information on the signal x. Such structural information can be of combinatorial type –as in 'model-based' compressed sensing– but can as well include probabilistic priors. Exploring such potential generalizations is –to a large extent– a future research program which is still in its infancy. Here we will briefly mention a few examples.

The case of block-sparse signals was already mentioned in Section 2. We write $x = (x_{B(1)}, x_{B(2)}, \dots, x_{B(\ell)})$ where $x_{B(i)} \in \mathbb{R}^{n/\ell}$ is a block for $\ell \in \{1, \dots, \ell\}$. Only a fraction $\varepsilon \in (0,1)$ of the blocks is non-vanishing. It is customary in this setting to replace the LASSO cost function with the following

$$C_{A,y}^{\text{Block}}(z) \equiv \frac{1}{2}\|y - Az\|^2 + \lambda \sum_{i=1}^{\ell} \|z_{B(i)}\|_2 \,. \tag{7.1}$$

The block-ℓ_2 regularization promotes block sparsity. Of course, the new regularization can be interpreted in terms of a new assumed prior that factorizes over blocks. An approximate message passing

algorithm suitable for this case is developed in [DM10]. Its analysis allows to generalize $\ell_0 - \ell_1$ phase transition curves to the block sparse case. This quantifies precisely the benefit of minimizing (7.1) over simple ℓ_1 penalization.

Tanaka and Raymond [TR10], and Som, Potter and Schniter and [SSS10] studied the case of signals with multiple level of sparsity. The simplest example consists of a signal $x = (x_{B(1)}, x_{B(2)})$, where $x_{B(1)} \in \mathbb{R}^{n_1}$, $x_{B(2)} \in \mathbb{R}^{n_2}$, $n_1 + n_2 = n$. Block $i \in \{1, 2\}$ has a fraction ε_i of non-zero entries, with $\varepsilon_1 \neq \varepsilon_2$. In the most complex case, one can consider a general factorized prior

$$p(\mathrm{d}x) = \prod_{i=1}^{n} p_i(\mathrm{d}x_i)\,,$$

where each $i \in [n]$ has a different sparsity parameter $\varepsilon_i \in (0, 1)$, and $p_i \in \mathcal{F}_{\varepsilon_i}$. In this case it is natural to use a weighted–ℓ_1 regularization, i.e. to minimize

$$\mathcal{C}^{\text{weight}}_{A,y}(z) \equiv \frac{1}{2}\|y - Az\|^2 + \lambda \sum_{i=1}^{n} w_i\,|z_i|\,, \tag{7.2}$$

for a suitable choice of the weights $w_1, \dots, w_n \geq 0$. The paper [TR10] studies the case $\lambda \to 0$ (equivalent to minimizing $\sum_i w_i|z_i|$ subject to $y = Az$), using non-rigorous statistical mechanics techniques that are equivalent to the state evolution approach presented here. Within a high-dimensional limit, it determines optimal tuning of the parameters w_i, for given sparsities ε_i. The paper [SSS10] follows instead the state approach explained in the present chapter, The authors develep a suitable AMP iteration and compute the optimal thresholds to be used by the algorithm. These are in correspondence with the optimal weights w_i mentioned above, albeit can be easily interpreted within the minimax framework develope in the previous pages.

The graphical model framework is particularly convenient for exploiting prior information that is probabilistin in nature. For instance Schniter [Sch10] study the case in which the signal x is generated by an Hidden Markov Model (HMM). In the simple case, the underlying Markov chain has two states indexed by $s_i \in \{0, 1\}$, and

$$p(\mathrm{d}x) = \sum_{s_1,\dots,s_n} \Big\{ \prod_{i=1}^{n} p(\mathrm{d}x_i|s_i) \cdot \prod_{i=1}^{n-1} p(s_{i+1}|s_i)\ \cdot p_1(s_1) \Big\}\,, \tag{7.3}$$

where $p(\,\cdot\,|0)$ and $p(\,\cdot\,|1)$ belong to two different sparsity classes $\mathcal{F}_{\varepsilon_0}$, $\mathcal{F}_{\varepsilon_1}$. For instance one can consider the case in which $\varepsilon_0 = 0$ and $\varepsilon_1 = 1$, i.e. the support of x coincides with the subset of coordinates such that $s_i = 1$. Message passing is used to perform reconstruction and AMP as a block in the algorithm to treat the constraint $y = Ax$.

Finally, many of the ideas developed here are not necessarily restricted to regularized linear regression. An important example is logistic regression, which is particularly suited for the case in which the measurements $y_1, \dots\ y_m$ are 0–1 valued. In the context of linear regression these are modeled as independent Bernoulli random variables with

$$p(y_a = 1|x) = \frac{e^{A_a^T x}}{1 + e^{A_a^T x}}\,, \tag{7.4}$$

with A_a a vector of 'features' that characterizes the a-th experiment. The objective is to learn the vector x of coefficients that encodes the relevance of each feature. A possible approach consists in

minimizing the regularized (negative) log-likelihood, that is

$$\mathcal{C}_{A,y}^{\text{LogReg}}(z) \equiv -\sum_{a=1}^{m} y_a (A_a^T z) + \sum_{a=1}^{m} \log\left(1 + e^{A_a^T z}\right) + \lambda \|z\|_1 \,, \tag{7.5}$$

The paper [BM10a] develops an approximate message passing algorithm for solving this minimization problem.

Acknowledgements

It is a pleasure to thank Mohsen Bayati, Jose Bento, David Donoho and Arian Maleki, with whom this research has been developed. This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211.

References

[AJ07] G. Andrew and G. Jianfeng, *Scalable training of l^1-regularized log-linear models*, Proceedings of the 24th international conference on Machine learning, 2007, pp. 33–40.

[BBM10] M. Bayati, J. Bento, and A. Montanari, *The LASSO risk: A comparison between theories and empirical results*, in preparation, 2010.

[BM10a] M. Bayati and A. Montanari, *Approximate message passing algorithms for generalized linear models*, in preparation, 2010.

[BM10b] ———, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory (2010), accepted, http://arxiv.org/pdf/1001.3448.

[CD95] S.S. Chen and D.L. Donoho, *Examples of basis pursuit*, Proceedings of Wavelet Applications in Signal and Image Processing III (San Diego, CA), 1995.

[CRT06] E. Candes, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics **59** (2006), 1207–1223.

[CT07] E. Candes and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Annals of Statistics **35** (2007), 2313–2351.

[DJ94a] D. L. Donoho and I. M. Johnstone, *Ideal spatial adaptation via wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[DJ94b] ———, *Minimax risk over l_p balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.

[DJHS92] D.L. Donoho, I.M. Johnstone, J.C. Hoch, and A.S. Stern, *Maximum entropy and the nearly black object*, Journal of the Royal Statistical Society, Series B (Methodological) **54** (1992), no. 1, 41–81.

[DM10] D. Donoho and A. Montanari, *Approximate message passing for reconstruction of block-sparse signals*, in preparation, 2010.

[DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.

[DMM10a] ———, *Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction*, Proceedings of IEEE Inform. Theory Workshop (Cairo), 2010.

[DMM10b] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, http://arxiv.org/abs/1004.1218, 2010.

[GB10] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 1.21*, `http://cvxr.com/cvx`, May 2010.

[GBS09] D. Guo, D. Baron, and S. Shamai, *A single-letter characterization of optimal noisy compressed sensing*, 47th Annual Allerton Conference (Monticello, IL), September 2009.

[GK06] D. Gamarnik and D. Katz, *Correlation decay and deterministic FPTAS for counting list-colorings of a graph*, 18th annual ACM-SIAM Symposium On Discrete Algorithm (New Orleans), 2006, pp. 1245–1254.

[GV05] D. Guo and S. Verdu, *Randomly Spread CDMA: Asymptotics via Statistical Physics*, IEEE Trans. on Inform. Theory **51** (2005), 1982–2010.

[Joh02] I. Johnstone, *Function Estimation and Gaussian Sequence Models*, Draft of a book, available at http://www-stat.stanford.edu/∼imj/based.pdf, 2002.

[KM10] S. Korada and A. Montanari, *Applications of Lindeberg Principle in Communications and Statistical Learning*, http://arxiv.org/abs/1004.0557, 2010.

[KWT09] Y. Kabashima, T. Wadayama, and T. Tanaka, *A typical reconstruction limit for compressed sensing based on lp-norm minimization*, J.Stat. Mech. (2009), L09003.

[MD10] A. Maleki and D. L. Donoho, *Optimally tuned iterative thresholding algorithm for compressed sensing*, IEEE Journal of Selected Topics in Signal Processing **4** (2010), 330–341.

[MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, Oxford, 2009.

[MR07] C. Moallemi and B. Van Roy, *Convergence of the min-sum algorithm for convex optimization*, 45th Annual Allerton Conference (Monticello, IL), September 2007.

[Pea88] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, San Francisco, 1988.

[RFG09] S. Rangan, A. K. Fletcher, and V. K. Goyal, *Asymptotic analysis of map estimation via the replica method and applications to compressed sensing*, PUT NIPS REF, 2009.

[SBB10] S. Sarvotham, D. Baron, and R. Baraniuk, *Bayesian Compressive Sensing via Belief Propagation*, IEEE Trans. on Signal Processing **58** (2010), 269–280.

[Sch10] P. Schniter, *Turbo Reconstruction of Structured Sparse Signals*, Proceedings of the Conference on Information Sciences and Systems (Princeton), 2010.

[SSS10] L.C. Potter S. Som and P. Schniter, *On Approximate Message Passing for Reconstruction of Non-Uniformly Sparse Signals*, Proceedings of the National Aereospace and Electronics Conference (Dayton, OH), 2010.

[Tan02] T. Tanaka, *A Statistical-Mechanics Approach to Large-System Analysis of CDMA Multiuser Detectors*, IEEE Trans. on Inform. Theory **48** (2002), 2888–2910.

[Tib96] R. Tibshirani, *Regression shrinkage and selection with the lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288.

[TR10] T. Tanaka and J. Raymond, *Optimal incorporation of sparsity information bt weighted L_1 optimization*, Proceedings of IEEE International Symposium on Inform. Theory (ISIT) (Austin), 2010.

[Wei05] D. Weitz, *Combinatorial criteria for uniqueness of Gibbs measures*, Rand. Struct. Alg. **27** (2005), 445475.

[ZY06] P. Zhao and B. Yu, *On model selection consistency of Lasso*, The Journal of Machine Learning Research **7** (2006), 2541–2563.