Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 3 (2957–2978)

MEAN FIELD ASYMPTOTICS IN HIGH-DIMENSIONAL STATISTICS: FROM EXACT RESULTS TO EFFICIENT ALGORITHMS

Andrea Montanari

Abstract

Modern data analysis challenges require building complex statistical models with massive numbers of parameters. It is nowadays commonplace to learn models with millions of parameters by using iterative optimization algorithms. What are typical properties of the estimated models? In some cases, the high-dimensional limit of a statistical estimator is analogous to the thermodynamic limit of a certain (disordered) statistical mechanics system. Building on mathematical ideas from the mean-field theory of disordered systems, exact asymptotics can be computed for high-dimensional statistical learning problems.

This theory suggests new practical algorithms and new procedures for statistical inference. Also, it leads to intriguing conjectures about the fundamental computational limits for statistical estimation.

1 Introduction

Natural and social sciences as well as engineering disciplines are nowadays blessed with abundant data which are used to construct ever more complex statistical models. This scenario requires new methodologies and new mathematical techniques to analyze these methods. In this article I will briefly overview some recent progress on two prototypical problems in this research area: high-dimensional regression and principal component analysis. This overview will be far from exhaustive, and will follow a viewpoint that builds on connections with mean field theory in mathematical physics and probability theory (see Section 5 for further context).

High-dimensional regression. We are given data points $(y_1, \boldsymbol{x}_1)$, $\dots (y_n, \boldsymbol{x}_n)$ that are independent draws from a common (unknown) distribution. Here $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature

The author was partially supported by grants NSF CCF-1714305 and NSF IIS-1741162.
MSC2010: primary 68Q87; secondary 62F10, 60K35, 82B44.

vector (or vector of covariates), and $y_i \in \mathbb{R}$ is a label or response variable. We would like to model the dependency of the response variable upon the feature vector as

$$y_i = \langle \boldsymbol{\theta}_0, \boldsymbol{x}_i \rangle + w_i \,, \tag{1-1}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is a vector of parameters (coefficients), and w_i captures non-linear dependence as well as random effects. This simple linear model (and its variants) has an impressive number of applications ranging from genomics Shevade and Keerthi [2003], to online commerce McMahan et al. [2013], to signal processing D. L. Donoho [2006] and Candès, Romberg, and Tao [2006].

Principal component analysis. We are given unlabeled data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, that are i.i.d. with zero mean and common covariance $\boldsymbol{\Sigma} \equiv \mathbb{E}\{\boldsymbol{x}_1 \boldsymbol{x}_1^{\mathsf{T}}\}$. We would like to estimate the directions of maximal variability of these data. Namely, denoting by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the ordered eigenvalues of $\boldsymbol{\Sigma}$ and by $\boldsymbol{v}_1(\boldsymbol{\Sigma}), \dots, \boldsymbol{v}_n(\boldsymbol{\Sigma})$ the correspondent eigenvectors, we would like to estimate $\boldsymbol{v}_1(\boldsymbol{\Sigma}), \dots, \boldsymbol{v}_k(\boldsymbol{\Sigma})$ for $k \ll d$ a fixed number. This task is a fundamental component of dimensionality reduction and clustering Kannan, S. Vempala, and Vetta [2004], and is often used in neuroscience Rossant et al. [2016] and genomics Abraham and Inouye [2014].

2 High-dimensional regression

Since Gauss [2011], least squares has been the method of choice for estimating the parameter vector $\boldsymbol{\theta}_0$ in the linear model (1-1). Least squares does not make assumptions on the coefficients $\boldsymbol{\theta}_0$, but implicitly assumes the errors w_i to be unbiased and all of roughly the same magnitude. In this is the case (for 'non-degenerate' features $\boldsymbol{x}_i$), consistent estimation is possible if and only if $n/p \gg 1$.

In contrast, many modern applications are characterized by a large amount of data, together with extremely complex models. In other words, both n and p are large and often comparable. The prototypical approach to this regime is provided by the following ℓ_1 regularized least squares problem, known as the Lasso Tibshirani [1996] or basis pursuit denoising Chen and D. L. Donoho [1995]:

$$\hat{\boldsymbol{\theta}}(\lambda; \boldsymbol{y}, \boldsymbol{X}) \equiv \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\} \,. \tag{2-1}$$

Here $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ is the vector of response variables, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the matrix whose i-th row is the i-th feature vector $\boldsymbol{x}_i$. Since the problem (2-1) is convex (and of a particularly simple form) it can be solved efficiently. In the following, we will drop the dependence of $\hat{\boldsymbol{\theta}}$ upon $\boldsymbol{y}, \boldsymbol{X}$ unless needed for clarity.

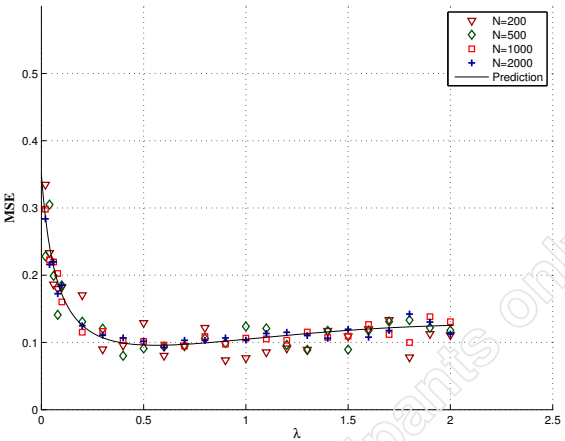

Figure 1: Mean square estimation error of the Lasso per dimension $\|\hat{\boldsymbol{\theta}}(\lambda)-\boldsymbol{\theta}_0\|_2^2/d$, as a function of the regularization parameter λ. Each point corresponds to a different instance of the problem (2-1) with symbols representing the dimension $d = N \in \{100, 500, 1000, 2000\}$. The number of samples is $n = d\delta$, with $\delta = 0.64$, and the noise level $\sigma^2 = 0.2 \cdot n$. The 'true' coefficients were generated with i.i.d. coordinates $\theta_{0,i} \in \{0, +1, -1\}$ and $\mathbb{P}(\theta_{0,i} = +1) = \mathbb{P}(\theta_{0,i} = +1) = 0.064$.

Over the last ten years, a sequence of beautiful works Candes and Tao [2007] and Bickel, Ritov, and Tsybakov [2009] has developed order-optimal bounds on the performance of the Lasso estimator. Analysis typically assumes that the data are generated according to model (1-1), with some vector $\boldsymbol{\theta}_0$, and i.i.d. noise $(w_i)_{i\le n}$: to be concrete we will assume here $w_i \sim \mathsf{N}(0, \sigma^2)$. For instance, if $\boldsymbol{\theta}_0$ has at most s_0 non-zero elements, and under suitable conditions on the matrix $\boldsymbol{X}$, it is known that (with high probability)

$$\|\hat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}_0\|_2^2 \le \frac{C\, s_0 \sigma^2}{n} \log d\,, \tag{2-2}$$

where C is a numerical constant.

This type of results give confidence in the use of the Lasso, and explain the origins of its effectiveness. However, they are not precise enough to compare different estimators with the same error rate or –say– different ways of selecting the regularization parameter λ. Also, they provide limited insight on the distribution of $\hat{\boldsymbol{\theta}}(\lambda)$, an issue that is crucial for statistical inference.

2.1 Exact asymptotics for the Lasso. In order to address these questions, a different type of analysis makes probabilistic assumptions about the feature vectors $\boldsymbol{x}_i$, and derives an asymptotically exact characterization of the high-dimensional estimator. In order to state a result of this type for the case of the Lasso, it is useful to introduce the proximal operator of the ℓ_1 norm (in one dimension):

$$\eta(y;\alpha) \equiv \arg\min_{x\in\mathbb{R}} \left\{\frac{1}{2}(y-x)^2 + \alpha|x|\right\}. \tag{2-3}$$

Explicitly, we have $\eta(y;\alpha) = (|y|-\alpha)\mathrm{sign}(y)$. We also note that the following simple consequence of the first-order stationarity conditions for problem (2-1) holds for any $\alpha > 0$:

$$\hat{\boldsymbol{\theta}}(\lambda) = \eta\big(\hat{\boldsymbol{\theta}}^{\mathrm{d}};\alpha\big), \quad \hat{\boldsymbol{\theta}}^{\mathrm{d}}(\alpha,\lambda) \equiv \hat{\boldsymbol{\theta}}(\lambda) + \frac{\alpha}{n\lambda}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}(\lambda)). \tag{2-4}$$

We say that a function $\psi : \mathbb{R}^d \to \mathbb{R}$ is pseudo-Lipschitz function of order k (and write $\psi \in \mathrm{PL}(k)$) if $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \le L(1 + (\|\boldsymbol{x}\|_2/\sqrt{d})^{k-1} + (\|\boldsymbol{y}\|_2/\sqrt{d})^{k-1})\|\boldsymbol{x}-\boldsymbol{y}\|_2$ for any $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d$. Also recall that a sequence of probability distributions ν_n on $\mathbb{R}^d$ converges in Wasserstein-k distance to ν if and only if $\int \psi(\boldsymbol{x})\nu_n(\mathrm{d}\boldsymbol{x}) \to \int \psi(\boldsymbol{x})\nu(\mathrm{d}\boldsymbol{x})$ for each $\psi \in \mathrm{PL}(k)$.

Theorem 1. *Consider a sequence of linear models (1-1) indexed by n, with $d = d(n)$ such that $\lim_{n\to\infty} n/d(n) = \delta \in (0,\infty)$, and let $\sigma = \sigma(n)$ be such that $\lim_{n\to\infty}\sigma(n)/\sqrt{n} = \sigma_0$. Assume $\boldsymbol{x}_i \sim \mathsf{N}(0,\boldsymbol{I}_d)$ and $w_i \sim \mathsf{N}(0,\sigma^2)$ independent and that the empirical distribution $d^{-1}\sum_{i=1}^d \delta_{\theta_{0,i}}$ converges in W_k to the law p_Θ of a random variable Θ.*

Let α_*, $\tau_*^2 \in \mathbb{R}_{>0}$ be the unique solution of the pair of equations

$$\lambda = \alpha\Big\{1 - \frac{1}{\delta}\mathbb{P}\big(|\Theta_* + \tau_* Z| \ge \alpha_*\big)\Big\}, \tag{2-5}$$

$$\tau_*^2 = \sigma_0^2 + \frac{1}{\delta}\mathbb{E}\big\{\big[\eta(\Theta + \tau_* Z;\alpha_*) - \Theta\big]^2\big\}, \tag{2-6}$$

where expectation is with respect to Θ and $Z \sim \mathsf{N}(0,1)$ independent. Then, taking $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{\mathrm{d}}(\alpha,\lambda)$, for any $\psi : \mathbb{R}^2 \to \mathbb{R}$, $\psi \in \mathrm{PL}(k)$, we have, almost surely,

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^d \psi(\theta_{0,i}, \hat{\theta}_i^{\mathrm{d}}) = \mathbb{E}\big\{\psi(\Theta, \Theta + \tau_* Z)\big\}. \tag{2-7}$$

The proof of this result Bayati and Montanari [2012] consists in introducing an iterative algorithm that converges rapidly to $\hat{\boldsymbol{\theta}}(\lambda)$ and can be analyzed exactly. Of course, the

existence of such an algorithm is of independent interest, cf. Section 4. Alternative proof techniques have been developed as well, and are briefly mentioned in the next section. All of these proofs take advantage in a crucial way of the fact that the optimization problem (2-1) is convex, which in turn is a choice dictated by computational tractability. However, for $\delta < 1$, the cost function is not strongly convex (since the kernel of $\boldsymbol{X}$ has dimension $n(1-\delta)$, with high probability), which poses interesting challenges.

Remark 2.1. A first obvious use of Theorem 1 is to derive asymptotic expressions for the risk of the Lasso. Using the stationarity condition (2-4) and choosing $\psi(x,y) = [x-\eta(y,\alpha_*)]^2$, we obtain

$$\lim_{n,p\to\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}(\lambda)-\boldsymbol{\theta}_0\|_2^2 = \mathbb{E}\{\big[\eta(\Theta+\tau_* Z;\alpha_*)-\Theta\big]^2\}\,. \tag{2-8}$$

For applications, this prediction has the disadvantage of depending on the asymptotic empirical distribution of the entries of $\boldsymbol{\theta}_0$, which is not known. One possible way to overcome this problem is to consider the worst case distribution D. L. Donoho, Maleki, and Montanari [2011]. Assuming $\boldsymbol{\theta}_0$ has at most $s_0 = p\varepsilon$ non-zero entries (and under the same assumptions of the last theorem), this results in the bound

$$\lim_{n,p\to\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}(\lambda)-\boldsymbol{\theta}_0\|_2^2 \le \frac{M(\varepsilon)}{1-M(\varepsilon)/\delta}\,\sigma_0^2\,. \tag{2-9}$$

Where $M(\varepsilon)$ is explicitly given in D. L. Donoho, Maleki, and Montanari [2011] and Montanari [2012], and behaves as $M(\varepsilon) = 2\varepsilon\log(1/\varepsilon)+O(\varepsilon)$ for small ε. This bound is tight in the sense that there exists sequences of vectors $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(n)$ for which the bound holds with equality.

Remark 2.2. Interestingly, Theorem 1 also characterizes the joint distribution of $\hat{\boldsymbol{\theta}}^{\mathrm{d}}$ and the true parameter vector. Namely $\hat{\theta}_i^{\mathrm{d}}$ is asymptotically Gaussian, with mean equal to the true parameter $\theta_{0,i}$ and variance τ_*^2. This is somewhat surprising, given that the Lasso estimator $\hat{\boldsymbol{\theta}} = \eta(\hat{\boldsymbol{\theta}}^{\mathrm{d}};\alpha_*)$ is highly non-Gaussian (in particular is $\hat{\theta}_i = 0$ for a positive fraction of the entries).

This Gaussian limit suggests a possible approach to statistical inference. In particular, a confidence interval for $\theta_{0,i}$ can be constructed by letting $J_i(c) = [\hat{\theta}_i^{\mathrm{d}} - c\tau_*, \hat{\theta}_i^{\mathrm{d}} + c\tau_*]$. The above theorem implies the following coverage guarantee

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{n} \mathbb{P}\big(\theta_{0,i}\in J_i(c)\big) = 1-2\Phi(-c)\,, \tag{2-10}$$

where $\Phi(x) \equiv \int e^{-t^2/2} \mathrm{d}t/\sqrt{2\pi}$ is the Gaussian distribution function. In other words, the confidence interval is valid on average Javanmard and Montanari [2014b].

Ideally, one would like stronger guarantees than in (2-10), for instance ensuring coverage for each coordinate, rather than on average over coordinates. Results of this type were proven in C.-H. Zhang and S. S. Zhang [2014], van de Geer, Bühlmann, Ritov, and Dezeure [2014], and Javanmard and Montanari [2014a, 2015] (these papers however do not address the regime $n/d \to \delta \in (0, \infty)$).

Remark 2.3. Theorem 1 assumes the entries of the design matrix $\boldsymbol{X}$ to be i.i.d. standard Gaussian. It is expected this result to enjoy some degree of universality, for instance with respect to matrices with i.i.d. entries with the same first two moments and sufficiently light tails. Universality results were proven in Korada and Montanari [2011], Bayati, Lelarge, and Montanari [2015], and Oymak and Tropp [2015], mainly focusing on the noiseless case $\sigma = 0$ which is addressed by solving the problem (2-1) in the limit $\lambda \to 0$ (equivalently, finding the solution of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}$ that minimizes $\|\boldsymbol{\theta}\|_1$). Classical tools of probability theory, in particular the moment method and Lindeberg swapping trick are successfully applied in this case.

Beyond matrices with i.i.d. entries, there is empirical evidence D. Donoho and Tanner [2009] and heuristic results Tulino, Caire, Verdu, and Shamai [2013] and Javanmard and Montanari [2014b] suggesting universality or (in some cases) generalizations of the prediction of Theorem 1.

2.2 Generalizations and comparisons. When the data $(y_i, \boldsymbol{x}_i)$, $1 \le i \le n$ contain outliers, the sum of square residuals $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$ in Equation (2-1) is overly influenced by such outliers resulting in poor estimates. Robust regression Huber and Ronchetti [2009] suggests to use the following estimator instead (focusing for simplicity on the un-regularized case):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho\big(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\big)\,, \tag{2-11}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is often chosen to be convex in order to ensure computational tractability. For instance, Huber [1964, 1973] advocated the use of $\rho(x) = \rho_{\text{Huber}}(x;)$ defined by $\rho_{\text{Huber}}(x; c) = x^2/2$ for $|x| \le c$ and $\rho_{\text{Huber}}(x; c) = c|x| - c^2/2$ otherwise. Results analogous to Theorem 1 were proven for robust estimators of the form (2-11) in Karoui [2013] and D. Donoho and Montanari [2016], following earlier conjectures in El Karoui, Bean, Bickel, Lim, and Yu [2013].

A second possibility for generalizing Theorem 1 is to modify the penalty function $\lambda\|\boldsymbol{\theta}\|_1$, and replacing it by $f(\boldsymbol{\theta})$ for $f : \mathbb{R}^d \to \mathbb{R}$ a convex function. General results

in this setting were proven in Chandrasekaran, Recht, Parrilo, and Willsky [2012] and Thrampoulidis, Oymak, and Hassibi [2015] via a different approach that builds on Gordon's minimax theorem Gordon [1988].

Finally, let us emphasize that sparsity of $\boldsymbol{\theta}_0$ –while motivating the Lasso estimator (2-1)– does not play any role in Theorem 1, which in fact holds for non-sparse $\boldsymbol{\theta}_0$ as well. Given this, it is natural to ask what is the best estimate for any given $\boldsymbol{\theta}_0$. Under the assumption of Theorem 1, it is natural to treat the $(\theta_i)_{i\le d}$ as i.i.d. draws with common distribution p_Θ. If this is the case, we can consider the posterior expectation estimator

$$\hat{\boldsymbol{\theta}}^{\text{Bayes}}(\boldsymbol{y},\boldsymbol{X}) \equiv \mathbb{E}_{p_\Theta}\big(\boldsymbol{\theta}\,\big|\,\boldsymbol{y},\boldsymbol{X}\big)\,. \tag{2-12}$$

The analysis of this estimator requires introducing two functions associated with the scalar problem of estimating $\Theta \sim p_\Theta$ from observations $Y = \sqrt{s}\,\Theta + Z$, $Z \sim \mathsf{N}(0,1)$:

$$\mathrm{I}(s) \equiv I(\Theta;\sqrt{s}\Theta+Z)\,,\quad \mathsf{mmse}(s) = \mathbb{E}\big\{\big[\Theta - \mathbb{E}(\Theta|\sqrt{s}\Theta+Z)\big]^2\big\}\,, \tag{2-13}$$

These two quantities are intimately related since $\frac{\mathrm{dI}}{\mathrm{d}s}(s) = \frac{1}{2}\,\mathsf{mmse}(s)$ Stam [1959] and Guo, Shamai, and Verdú [2005]. The following is a restatement of a theorem proved in Reeves and Pfister [2016],

Theorem 2. *Under the assumptions of Theorem 1, define the function* $\tau^2 \mapsto \Psi(\tau^2)$ *by*

$$\Psi(\tau^2) = \mathrm{I}(\tau^{-2}) + \frac{\delta}{2}\left(\log(\delta\tau^2) - \frac{\delta}{2} + \frac{\delta\sigma_0^2}{2\tau^2}\right)\,. \tag{2-14}$$

If, for $\sigma_0^2 > 0$, $\tau^2 \mapsto \Psi(\tau^2)$ *has at most three critical point and* $\tau^2_{\text{Bayes}} \equiv \arg\min_{\tau^2>0}\Psi(\tau^2)$ *is unique, then*

$$\lim_{n,d\to\infty}\frac{1}{d}\mathbb{E}\big\{\big\|\hat{\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{X}) - \boldsymbol{\theta}\big\|_2\big\} = \mathsf{mmse}(\tau^{-2}_{\text{Bayes}})\,. \tag{2-15}$$

A substantial generalization of this theorem was proved recently in Barbier, Macris, Dia, and Krzakala [2017], encompassing in particular a class of generalized linear models.

Notice that τ_{Bayes} must satisfy the following first-order stationarity condition (which is obtained by differentiating $\Psi(\cdot)$):

$$\tau^2_{\text{Bayes}} = \sigma^2 + \frac{1}{\delta}\,\mathsf{mmse}(\tau^{-2}_{\text{Bayes}})\,. \tag{2-16}$$

The form of this equation is tantalizingly similar to the one for the Lasso mean square error, cf. Equation (2-6). In both case the right-hand side is given in terms of the error in

estimating the scalar $\Theta \sim p_\Theta$ from noisy observations $Y = \Theta + \tau Z$. While Equation (2-6) corresponds to the error of proximal denoising using ℓ_1 norm, the Bayes estimation error appears in Equation (2-16).

2.3 Decoupling. A key property is shared by the Lasso and other convex estimators, as well as the Bayes-optimal estimators of Section 2.2. It will also hold for the message passing algorithms of Section 4 and it is sometimes referred to as 'decoupling'. Notice that Equation (2-7) of Theorem 1 can be interpreted as follows. By Equation (2-4), we can use the estimate $\hat{\boldsymbol{\theta}}$ to construct new 'pseudo-data' $\hat{\boldsymbol{\theta}}^{\mathrm{d}}$ with the following remarkable property. Each coordinate of the pseudo-data $\hat{\theta}^{\mathrm{d}}_i$ is approximately distributed as a Gaussian noisy observation of the true parameter $\theta_{0,i}$.

This naturally raises the question of the joint distribution of k coordinates $\hat{\theta}^{\mathrm{d}}_{i(1)}, \dots \hat{\theta}^{\mathrm{d}}_{i(k)}$. Decoupling occurs when these are approximately distributed as observations of $\theta_{i(1)}, \dots, \theta_{i(k)}$ with independent noise. For instance, in the case of Theorem 1, this can be formalized as

$$\lim_{n,d\to\infty} \frac{1}{d^k} \sum_{i(1),\dots,i(k)=1}^{d} \psi(\theta_{0,i(1)}, \dots, \theta_{0,i(k)}; \hat{\theta}^{\mathrm{d}}_{i(1)}, \dots, \hat{\theta}^{\mathrm{d}}_{i(1)}) = \\ = \mathbb{E}\{\psi(\Theta_1, \dots, \Theta_k; \Theta_1 + \tau_* Z_1, \dots, \Theta_k + \tau_* Z_k)\}, \tag{2-17}$$

where ψ is a bounded continuous function and $(\Theta_\ell)_{\ell\le k} \sim_{iid} p_\Theta$ independent of $(Z_\ell)_{\ell\le k} \sim \mathsf{N}(0,1)$. In this form, decoupling is in fact an immediate consequence of Equation (2-7), but other forms of decoupling are proved in the literature. (And sometimes the model of interest has to be perturbed in order to obtain decoupling.)

3 Principal component analysis

A standard model for principal component analysis assumes that the vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ are centered Gaussian, with covariance $\boldsymbol{\Sigma} = \boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{I}_n$, for $\boldsymbol{\theta}_0$ a fixed unknown vector. Equivalently, if we let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be the matrix whose i-th row is the vector $\boldsymbol{x}_i$, we have $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$, where $\boldsymbol{u} = (u_i)_{i\le n}$ is a vector with i.i.d. entries $u_i \sim \mathsf{N}(0,1)$, and $(W_{ij})_{i\le n, j\le d} \sim \mathsf{N}(0,1)$.

For the sake of simplicity, we shall consider here the symmetric version of this model. The data consists in a symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, where

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}, \tag{3-1}$$

where $\boldsymbol{W}$ is a noise matrix from the $\mathsf{GOE}(n)$ ensemble, namely $(W_{ij})_{i<j\le n} \sim \mathsf{N}(0,1/n)$ are independent of $(W_{ii})_{i\le n} \sim \mathsf{N}(0,2/n)$, and $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$. We further assume $\lambda \ge 0$ and $\|\boldsymbol{\theta}_0\|_2^n/n \to 1$ as $n \to \infty$. This normalization is chosen to make the problem nontrivial when $\lambda = \Theta(1)$.

We are asked to estimate $\boldsymbol{\theta}_0 \in \mathbb{R}^n$ from a single observation of the matrix $\boldsymbol{X}$. Spectral methods are –by far– the best studied approach to this problem, and the asymptotic spectral properties of $\boldsymbol{X}$ have been studied in exquisite detail across probability theory and statistics Baik, Ben Arous, and Péché [2005], Baik and Silverstein [2006], Féral and Péché [2007], Johnstone [2001], Paul [2007], Capitaine, Donati-Martin, and Féral [2009], Benaych-Georges and Nadakuditi [2011, 2012], and Knowles and Yin [2013]. In particular, letting $\hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X})$ denote the principal eigenvector of $\boldsymbol{X}$, we have

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|} = \begin{cases} 0 & \text{if } \lambda \le 1, \\ \sqrt{1-\lambda^{-2}} & \text{if } \lambda > 1. \end{cases} \tag{3-2}$$

In other words, the spectral estimator achieves a positive correlation with the unknown vector $\boldsymbol{\theta}_0$ provided $\lambda > 1$: this phenomenon is known as the BBAP phase transition Baik, Ben Arous, and Péché [2005].

From a statistical perspective, the principal eigenvector is known to be an asymptotically optimal estimator if no additional information is available about $\boldsymbol{\theta}_0$. In particular, it is asymptotically equivalent to the Bayes-optimal estimator when the prior of $\boldsymbol{\theta}_0$ is uniformly distributed on a sphere of radius $\sqrt{n}$. However, in many problems of interest, additional information is available on $\boldsymbol{\theta}_0$: exploiting this information optimally requires to move away from spectral methods and from the familiar grounds of random matrix theory.

3.1 $\mathbb{Z}_2$**-synchronization.** In some cases, all the entries of $\boldsymbol{\theta}_0$ are known to have equal magnitude. For instance, in the community detection problem we might be required to partition the vertices of a graph in two communities such that vertices are better connected within each part than across the partition. Under the so-called stochastic block model Decelle, Krzakala, Moore, and Zdeborová [2011] and Abbe [2017], the adjacency matrix of the graph is of the form (3-1) (albet with Bernoulli rather than Gaussian noise) whereby $\theta_{0,i} \in \{+1,-1\}$ is the label of vertex $i \in [n]$. Another motivation comes from group synchronization Wang and Singer [2013], which is a relative of model (3-1) whereby the unknowns $\theta_{0,i}$ are elements of a compact matrix group $\mathfrak{G}$. In the special case $\mathfrak{G} = \mathbb{Z}_2 = (\{+1,-1\}, \cdot)$, the resulting model is a special case of (3-1).

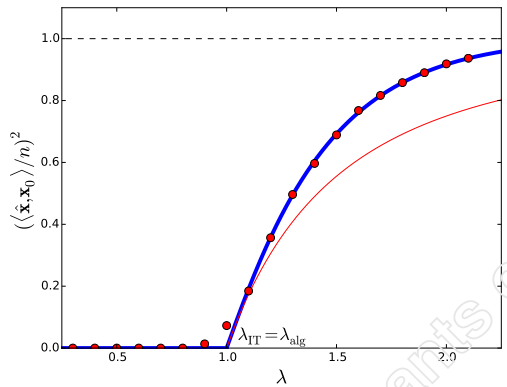

Figure 2: Estimation accuracy $(\langle \hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}, \boldsymbol{\theta}_0 \rangle / n)^2$ within the $\mathbb{Z}_2$-synchronization problem. Red circles: numerical simulations with the AMP algorithm (form matrices of dimension $n = 2000$ and $t = 200$ iterations). Continuous thick blue line: Bayes optimal estimation accuracy, cf. Theorem 3. Dashed blue line: other fixed points of state evolution. Red line: Accuracy achieved by principal component analysis.

The following theorem follows from Deshpande, Abbe, and Montanari [2017] and Montanari and Venkataramanan [2017] and provides an asymptotically exact characterization of optimal estimation in the $\mathbb{Z}_2$-synchronization problem, with respect to the metric in Equation (3-2).

Theorem 3. *Consider the model (3-1) and let $\gamma_* \in [0, \infty)$ denote the largest solution of*

$$\gamma = \lambda\big(1 - \mathsf{mmse}(\gamma)\big)\,, \tag{3-3}$$

where $\mathsf{mmse}(\,\cdot\,)$ is defined as in Equation (2-13), with $p_\Theta = (1/2)\delta_{+1} + (1/2)\delta_{-1}$.

Then, there exists an estimator $\hat{\boldsymbol{\theta}}^{\mathrm{Bayes}} : \boldsymbol{X} \mapsto \hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X})$ such that, almost surely,

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \sqrt{1 - \mathsf{mmse}(\gamma_*)}\,. \tag{3-4}$$

Further, this accuracy can be approximated within arbitrarily small (constant) additive error ε by a polynomial-time message passing algorithm, cf. Section 4. Finally, no estimator can achieve a better correlation than in Equation (3-4).

This prediction is illustrated in Figure 2. Notice that it undergoes a phase transition at the spectral threshold $\lambda = 1$. For $\lambda < 1$ no estimator can achieve a correlation that is bounded away from zero.

Remark 3.1. Substantial generalizations of the last theorem were proved in several papers Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová [2016], Lelarge and Miolane [2016], and Miolane [2017]. These generalization use new proof techniques inspired by mathematical spin glass theory and cover the case of vectors $\boldsymbol{\theta}$ whose entries have general distributions p_Θ, as well as the rectangular and higher rank cases.

In particularly, Theorem 3 holds almost verbatimly if $\boldsymbol{\theta}_0$ has i.i.d. entries with known distribution p_Θ such that $\int \theta^2\, p_\Theta(\mathrm{d}\theta) = 1$ and $\int \theta^4\, p_\Theta(\mathrm{d}\theta) < \infty$. One important difference is that in this more general setting, Equation (3-3) can have multiple solutions, and Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová [2016], Lelarge and Miolane [2016], and Miolane [2017] provide a way to select the 'correct' solution that is analogous to the one in Theorem 2.

Remark 3.2. As in the linear regression problem, the fixed point Equation (3-3) points at a connection between the high-dimensional estimation problem of Equation (3-1), where we are required to estimate n bits of information $\theta_{0,i} \in \{+1, -1\}$, to a much simpler scalar problem. The underlying mechanism is again the decoupling phenomenon of Section 2.3. An alternative viewpoint on the same phenomenon is provided by the analysis of message passing algorithms outlined in Section 4.

Remark 3.3. Replacing the minimum mean-square estimator $\mathbb{E}\{\Theta|Y\}$ with the optimal linear estimator $\widehat{\Theta}(Y) = aY$ in the definition of Equation (2-13) yields the general upper bound (for $\mathbb{E}\{\Theta^2\} = 1$) $\mathsf{mmse}(s) \le 1/(1+s)$. Substituting in Equations (3-3) and (3-4) this yields in turn

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\text{Bayes}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\text{Bayes}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \ge \sqrt{\left(1 - \frac{1}{\lambda^2}\right)_+}\,. \tag{3-5}$$

We thus recover the predicted accuracy of spectral methods, cf. 3-2. It is not hard to show that this inequality is strict unless the coordinates of $\boldsymbol{\theta}_0$ are asymptotically Gaussian. Figure 2 compares the Bayes optimal accuracy of Theorem 3 with this spectral lower bound.

While Theorem 3 states that there exists a message passing algorithm that essentially achieves Bayes-optimal performances, this type of algorithms can be sensitive to model misspecification. It is therefore interesting to consider other algorithmic approaches. One standard starting point is to consider the maximum likelihood estimator that is obtained

by solving the following optimization problem:

$$\begin{aligned} &\text{maximize} && \langle \boldsymbol{X}, \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} \rangle , \\ &\text{subject to} && \boldsymbol{\theta} \in \{+1, -1\}^n . \end{aligned} \tag{3-6}$$

Semidefinite programing (SDP) relaxations provide a canonical path to obtain a tractable algorithm for such combinatorial problems. A very popular relaxation for the present case Goemans and Williamson [1995] and Nesterov [1998] is the following program in the decision variable $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} &\text{maximize} && \langle \boldsymbol{X}, \boldsymbol{Q} \rangle , \\ &\text{subject to} && \boldsymbol{Q} \succeq \boldsymbol{0} , \\ & && Q_{ii} = 1 \ \text{ for all } i \in \{1, \dots, n\} . \end{aligned} \tag{3-7}$$

The matrix $\boldsymbol{Q}$ can be interpreted as a covariance matrix for a certain distribution on the vector $\boldsymbol{\theta}$. Once a solution $\boldsymbol{Q}_*$ of this SDP is computed, we can use it to produce an estimate $\hat{\boldsymbol{\theta}}^{\text{SDP}} \in \{+1, -1\}^n$ in many ways (this step is called 'rounding' in theoretical computer science). For instance, we can take the sign of its principal eigenvector: $\hat{\boldsymbol{\theta}} = \text{sign}(\boldsymbol{v}_1(\boldsymbol{Q}_*))$. There are many open questions concerning the SDP (Equation (3-7)). In particular Javanmard, Montanari, and Ricci-Tersenghi [2016] uses statistical physics methods to obtain close form expression for its asymptotic accuracy, that are still unproven. On the positive side, Montanari and Sen [2016] establishes the following positive result.

Theorem 4. *Let $\boldsymbol{X}$ be generated according to the model (3-1) with $\boldsymbol{\theta}_0 \in \{+1, -1\}^n$, and denote by $\boldsymbol{Q}_*$ the solution of the SDP (Equation (3-7)). Then there exists a rounding procedure that produces $\hat{\boldsymbol{\theta}}^{\text{SDP}} = \hat{\boldsymbol{\theta}}^{\text{SDP}}(\boldsymbol{Q}_*) \in \{+1, -1\}^n$ such that for any $\lambda > 1$ there exists $\varepsilon > 0$ such that, with high probability*

$$\frac{|\langle \hat{\boldsymbol{\theta}}^{\text{SDP}}, \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\text{SDP}}\|_2 \|\boldsymbol{\theta}_0\|_2} \ge \varepsilon . \tag{3-8}$$

In other words, semidefinite programming matches the optimal threshold.

3.2 The computation/information gap and the hidden clique problem. It is worth emphasizing one specific aspect of Theorem 3. Within the spiked matrix model (3-1), there exists a polynomia- time computable estimator that nearly achieves Bayes-optimal performances, despite the underlying estimation problem is combinatorial in nature: $\boldsymbol{\theta}_0 \in \{+1, -1\}^n$.

It is important to stress that the existence of a polynomial-time estimator for the problem (3-1) is far from being the norm, when changing the distribution p_Θ, and the signal-to-noise ratio λ. In certain cases, simple algorithms achieve nearly optimal performances. In others, even highly sophisticated approaches (for instance SDP relaxations from the sum-of-squares hierarchy Barak and Steurer [2014]) fail.

Developing a theory of which statistical estimation problems are solvable by polynomial-time algorithms is a central open problem in this area, and a very difficult one. For certain classes of problems, a bold conjecture was put forward on the basis of statistical physics insights.

In order to formulate this conjecture in the context of model (3-1), it is useful to state the following theorem from Montanari and Venkataramanan [2017] that concerns the case of a general distribution p_Θ of the entries of $\boldsymbol{\theta}_0$.

Theorem 5. *Consider –to be specific– model (3-1), with $\boldsymbol{\theta}_0$ having i.i.d. entries with known distribution p_Θ. Assume p_Θ and λ to be independent of n and known, with $\int \theta^2 p_\Theta(\mathrm{d}\theta) = 1$. If $\int \theta\, p_\Theta(\mathrm{d}\theta) = 0$, further assume $\lambda > 1$. Then there exists a polynomial time (message passing) algorithm that outputs an estimator $\hat{\boldsymbol{\theta}}^{\mathrm{AMP}} = \hat{\boldsymbol{\theta}}^{\mathrm{AMP}}(\boldsymbol{X})$ such that*

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{AMP}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{AMP}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \sqrt{1 - \mathsf{mmse}(\gamma_{\mathrm{AMP}})}\,. \tag{3-9}$$

where $\mathsf{mmse}(\cdot)$ is defined as in Equation (2-13) and γ_{AMP} is the smallest non-zero fixed point of Equation (3-3).

Within the setting of this theorem, *it is conjectured that Equation (3-9) is the optimal accuracy achieved by polynomial time estimators* Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová [2016], Lelarge and Miolane [2016], Lesieur, Krzakala, and Zdeborová [2017], and Montanari and Venkataramanan [2017]. Together with Remark 3.1, this provides a precise –albeit conjectural– picture of the gap between fundamental statistical limits (the Bayes optimal accuracy) and computationally efficient methods. This is sometimes referred to as the information-computation gap. The same phenomenon was pointed out earlier in other statistical estimation problems, e.g. in the context of error correcting codes Mézard and Montanari [2009].

The hidden clique problem is the prototypical example of a statistical estimation problem in which a large information-computation gap is present, and it is the problem for which this phenomenon is best studied. Nature generates a graph over n vertices as follows: a subset $S \subseteq [n]$ of size $|S| = k$ is chosen uniformly at random. Conditional on S,

for any pair of vertices $\{i, j\}$, an edge is added independently with probability

$$\mathbb{P}\big(\{i,j\} \in E \big| S\big) = \begin{cases} 1 & \text{if } \{i,j\} \subseteq S, \\ 1/2 & \text{otherwise.} \end{cases} \tag{3-10}$$

We are given one realization G such a graph, and are requested to identify the set S. In order to clarify the connection with the rank-one plus noise model (3-1), denote by $\boldsymbol{A}$ the $+/-$ adjacency matrix of G. This is the $n \times n$ matrix whose entry i, j is $A_{ij} = +1$ if $(i, j) \in E$ and -1 otherwise (in what follows, all matrices have diagonal entries equal to $+1$). Then it is easy to see that

$$\frac{1}{\sqrt{n}}\boldsymbol{A} = \lambda \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W} - \boldsymbol{W}_{S,S}\,, \tag{3-11}$$

$$\boldsymbol{\theta}_0 = \frac{1}{\sqrt{k}}\mathbf{1}_S\,, \quad \lambda = \frac{k}{\sqrt{n}}\,, \tag{3-12}$$

where $\boldsymbol{W}_{SS}$ is the restriction of matrix $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$ to rows/columns with index in S and $(W_{ij})_{i<j} \sim_{iid} \mathsf{Unif}(+1/\sqrt{n}, -1/\sqrt{n})$. This model has a few differences with respect to the one in Equation (3-1): (i) The noise is Radamacher instead of Gaussian; (ii) The term $\boldsymbol{W}_{SS}$ of noise is subtracted; (iii) The distribution of the entries of $\boldsymbol{\theta}_0$ is $p_\Theta = (k/n)\delta_{1/\sqrt{k}} + (1 - (k/n))\delta_0$; hence, for $\lambda = k/\sqrt{n}$ fixed, p_Θ depends on n. Of these differences, only the last one is really important for our purposes, and changes some qualitative features of the problem.

From a purely statistical point of view, the set S can be reconstructed with high probability provided that $k \geq 2(1+\varepsilon)\log_2(n)$, by searching over all subsets of k vertices. On the other hand, a variety of polynomial-time algorithms have been analyzed, including Monte Carlo Markov Chain Jerrum [1992], spectral algorithms Alon, Krivelevich, and Sudakov [1998], message passing algorithms Deshpande and Montanari [2015], semidefinite programming relaxations in the Feige and Krauthgamer [2003] and sum-of-squares Barak, Hopkins, Kelner, Kothari, Moitra, and Potechin [2016] hierarchies, statistical query models Feldman, Grigorescu, Reyzin, S. S. Vempala, and Xiao [2013]. Despite all of these efforts, no polynomial-time algorithms is known to be effective with high probability for $k \leq n^{1/2-\varepsilon}$, suggesting the possibility of a large information/computation gap for the hidden clique problem. As shown in Deshpande and Montanari [2015], this is consistent with the general picture emerging from statistical physics (although the hidden clique problem does not fit in the setting of the conjecture mentioned above).

4 Message passing algorithms

Message passing algorithms were already mentioned a few times in the previous pages and provide one natural class of algorithms to deal with random structures. Also, they are intimately connected to mean field approximations in statistical physics. Given an undirected graph $G = (V, E)$, we introduce the set of directed edges $\vec{E} = \{(i \to j) : (i, j) \in E\}$ (namely, for each edge $(i, j) \in E$, we introduce the two directed edges $(i \to j)$ and $(j \to i)$). A message passing algorithm operates on messages $(\nu^t_{i\to j})_{(i\to j)\in\vec{E}} \in \mathsf{M}^{\vec{E}}$ taking values in a set M, with t a time index. Messages are updated according to local rules:

$$\nu^{t+1}_{i\to j} = \Psi^{(t)}_{i\to j}\big(\nu^t_{k\to i} : k \in \partial i \setminus j\big), \tag{4-1}$$

In other words, a message outgoing vertex i at time $t+1$ is a function of messages ingoing the same vertex at time t, with the exception of the message along the same edge. Here all edges are updated synchronously: asynchronous schemes are of interest as well.

Notice that rather than an algorithm, (4-1) describes a general class of dynamical systems: we did not specify what the updating function $\Psi^{(t)}_{i\to j}$ are, what is the space M in which messages live, and not even what is the problem that we are trying to solve. We only insisted on locality and the 'non-backtracking information' condition: these turn out to be sufficient to lead to some interesting properties of the dynamical system (4-1) when the underlying graph is a tree or locally tree-like Richardson and Urbanke [2008].

Special forms of the dynamics (4-1) are used for Bayesian inference Koller and Friedman [2009], decoding in digital communications Richardson and Urbanke [2008], and combinatorial optimization Mézard and Montanari [2009]. To the best of my knowledge, the first appearance an algorithm of the form (4-1) (and its analysis) dates back to Gallager Ph.D. thesis on low-density parity check codes in the early sixties Gallager [1962]. As an analytical tool, recursions of this type have been in use in physics at least since Bethe's work in the thirties Bethe [1935].

At first sight, message passing algorithms might seem immaterial to the problems discussed in the rest of this paper: typically these are not associated to a locally tree-like graph (possibly with te exception of some sparse-graph versions of the hidden-clique problem Deshpande and Montanari [2015]). Somewhat surprisingly, there exists a natural class of algorithms whose datum is not a locally tree-like graph but a (dense) random matrix, and which can be considered a close relative of message passing algorithms. In fact, they can be thought as the limit of message passing algorithm when the average degree of the underlying graph diverges (see, for instance, Bayati and Montanari [2011]). These algorithms

are known as *approximate message passing*: for the sake of simplicity we will briefly discuss them in the case in which the data consists of a matrix $\boldsymbol{A} \sim \mathsf{GOE}(n)$. The algorithm operates on variables $\hat{\boldsymbol{\theta}}^t \in \mathbb{R}^{n\times k}$ where k is considered as fixad as $n \to \infty$. This state is updated according to

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{t+1} &= \boldsymbol{A}\, f_t(\hat{\boldsymbol{\theta}}^t) - f_{t-1}(\hat{\boldsymbol{\theta}}^{t-1})\mathsf{B}_t^{\mathsf{T}}, \\ \mathsf{B}_t &= \frac{1}{n}\sum_{i=1}^{m} \frac{\partial f_t}{\partial \hat{\boldsymbol{\theta}}_i^t}(\hat{\boldsymbol{\theta}}_i^t). \end{aligned} \tag{4-2}$$

Here $f_t : \mathbb{R}^k \to \mathbb{R}^k$ is a Lipschitz continuous function and we denote by $f_t(\hat{\boldsymbol{\theta}}^t) \in \mathbb{R}^{n\times k}$ the matrix that is obtained by applying f_t row-by-row to $\hat{\boldsymbol{\theta}}^t$. The i-th row of $\hat{\boldsymbol{\theta}}^t$ is denoted by $\hat{\boldsymbol{\theta}}_i^t$ and, by convention, $\mathsf{B}_0 = 0$. Once again, Equation (4-2) does not specify the update functions f_t, nor the problem we are trying to solve: rather it define a class of dynamical systems. However, special cases can be developed for Bayesian inference, statistical estimation, optimization, and so on.

In the Bayesian case, the functions $f_t(\cdot)$ take the form of conditional expectations with respect to certain distributions, and the fixed point version of the iteration (4-2) dates back to the work of Thouless, Anderson, Palmer (TAP) on mean field spin glasses Thouless, Anderson, and Palmer [1977]. Iterative solutions of the TAP equations were studied among others in Bolthausen [2014]. The general (non-Bayesian) formulation was developed and analyzed in D. L. Donoho, Maleki, and Montanari [2009] and Bayati and Montanari [2011], with the original motivation being its application to compressed sensing.

Crucially, the recursion (4-2) admits an asymptotically exact characterization in the limit $n \to \infty$ with t fixed. This type of analysis is known as *state evolution*.

Theorem 6. *Consider the AMP iteration (4-2) with f_t Lipschitz continuous, $\boldsymbol{A} \sim \mathsf{GOE}(n)$, and deterministic initialization $\hat{\boldsymbol{\theta}}^0$ such that $\lim_{n\to\infty} f_0(\hat{\boldsymbol{\theta}}^0)^{\mathsf{T}} f_0(\hat{\boldsymbol{\theta}}_0)/n = \boldsymbol{\Sigma}_0 \in \mathbb{R}^{k\times k}$. Define the sequence $\boldsymbol{\Sigma}_t \in \mathbb{R}^{k\times k}$ via the recursion:*

$$\boldsymbol{\Sigma}_{t+1} = \mathbb{E}\{f_t(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{g})\, f_t(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{g})\}, \tag{4-3}$$

where expectation is with respect to $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_k)$. Then, for any t and any test function $\psi : \mathbb{R}^k \to \mathbb{R}$ that is continuous and with at most quadratic growth at infinity, the following holds almost surely

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\boldsymbol{\theta}}_i^t) = \mathbb{E}\{\psi(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{g})\}. \tag{4-4}$$

A theorem of this type was first proved in the case of the TAP equations for the Sherrington-Kirkpatrick model in Bolthausen [2014] and then in general in Bayati and Montanari [2011]. Generalizations have also been proved for matrices with non-i.i.d. entries Javanmard and Montanari [2013], non-Gaussian random matrices Bayati, Lelarge, and Montanari [2015], non-separable functions f_t Berthier, Montanari, and Nguyen [2017], invariant matrix ensembles Schniter, Rangan, and Fletcher [2016], non-asymptotic settings Rush and Venkataramanan [2016], non-deterministic initializations Montanari and Venkataramanan [2017].

This type of analysis is used to prove the algorithmic part of Theorem 3, as well as algorithmic versions of the other theorems in this paper.

5 Context and conclusion

For the greatest part of the last century, mean field theory has been an important tool used by physicists to understand the behavior of systems with a large number of degrees of freedom Landau [1937]. Classical mean field theory describes homogeneous states, e.g. the state of a fluid in which each molecule interacts with the average environment created by all the other molecules. Starting in the late seventies, a new class mean-field ideas was developed to deal with heterogeneous states, where all particles look statistically the same, but typical configurations are highly heterogeneous, as is the case with disordered solids and spin glasses Kirkpatrick and Sherrington [1978] and Parisi [1979]. This opened the way to applying the same tools to a variety of probabilistic models without apparent connection to physics, including combinatorial optimization and neural networks (see Mézard, Parisi, and Virasoro [1987] for seminal papers in this direction).

Over the last few years, this circle of ideas has gone through a spectacular renaissance for at least three reasons: (*i*) Mathematical methods have been developed to prove (part of) physicists' predictions Talagrand [2007], Panchenko [2013], and Ding, Sly, and Sun [2015]; (*ii*) Structural insights from physics have unveiled new computational phenomena; (*iii*) New applications of these techniques have emerged within high-dimensional statistics and machine learning, generating interest across several communities.

This brief overview focused on the last two points, and hopefully will provide the reader with an entrypoint in this rapidly evolving literature.

References

Emmanuel Abbe (2017). "Community detection and stochastic block models: recent developments". arXiv: 1703.10146 (cit. on p. 2965).

Gad Abraham and Michael Inouye (2014). "Fast principal component analysis of large-scale genome-wide data". *PloS one* 9.4, e93766 (cit. on p. 2958).

Noga Alon, Michael Krivelevich, and Benny Sudakov (1998). "Finding a large hidden clique in a random graph". In: *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*. ACM, New York, pp. 594–598. MR: 1642973 (cit. on p. 2970).

Jinho Baik, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". *Ann. Probab.* 33.5, pp. 1643–1697. MR: 2165575 (cit. on p. 2965).

Jinho Baik and Jack W. Silverstein (2006). "Eigenvalues of large sample covariance matrices of spiked population models". *J. Multivariate Anal.* 97.6, pp. 1382–1408. MR: 2279680 (cit. on p. 2965).

Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin (2016). "A nearly tight sum-of-squares lower bound for the planted clique problem". In: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, pp. 428–437. MR: 3631005 (cit. on p. 2970).

Boaz Barak and David Steurer (2014). "Sum-of-squares proofs and the quest toward optimal algorithms". arXiv: 1404.5236 (cit. on p. 2969).

Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová (2016). "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Advances in Neural Information Processing Systems*, pp. 424–432 (cit. on pp. 2967, 2969).

Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala (2017). "Mutual Information and Optimality of Approximate Message-Passing in Random Linear Estimation". arXiv: 1701.05823 (cit. on p. 2963).

Mohsen Bayati, Marc Lelarge, and Andrea Montanari (2015). "Universality in polytope phase transitions and message passing algorithms". *Ann. Appl. Probab.* 25.2, pp. 753–822. MR: 3313755 (cit. on pp. 2962, 2973).

Mohsen Bayati and Andrea Montanari (2011). "The dynamics of message passing on dense graphs, with applications to compressed sensing". *IEEE Trans. Inform. Theory* 57.2, pp. 764–785. MR: 2810285 (cit. on pp. 2971–2973).

– (2012). "The LASSO risk for Gaussian matrices". *IEEE Trans. Inform. Theory* 58.4, pp. 1997–2017. MR: 2951312 (cit. on p. 2960).

Florent Benaych-Georges and Raj Rao Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". *Adv. Math.* 227.1, pp. 494–521. MR: 2782201 (cit. on p. 2965).

– (2012). "The singular values and vectors of low rank perturbations of large rectangular random matrices". *J. Multivariate Anal.* 111, pp. 120–135. MR: 2944410 (cit. on p. 2965).

Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen (2017). "State Evolution for Approximate Message Passing with Non-Separable Functions". arXiv: 1708.03950 (cit. on p. 2973).

Hans A Bethe (1935). "Statistical theory of superlattices". *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 150.871, pp. 552–575 (cit. on p. 2971).

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov (2009). "Simultaneous analysis of lasso and Dantzig selector". *Ann. Statist.* 37.4, pp. 1705–1732. MR: 2533469 (cit. on p. 2959).

Erwin Bolthausen (2014). "An iterative construction of solutions of the TAP equations for the Sherrington-Kirkpatrick model". *Comm. Math. Phys.* 325.1, pp. 333–366. MR: 3147441 (cit. on pp. 2972, 2973).

Emmanuel J. Candès, Justin Romberg, and Terence Tao (2006). "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". *IEEE Trans. Inform. Theory* 52.2, pp. 489–509. MR: 2236170 (cit. on p. 2958).

Emmanuel Candes and Terence Tao (2007). "The Dantzig selector: statistical estimation when p is much larger than n". *Ann. Statist.* 35.6, pp. 2313–2351. MR: 2382644 (cit. on p. 2959).

Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral (2009). "The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations". *Ann. Probab.* 37.1, pp. 1–47. MR: 2489158 (cit. on p. 2965).

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky (2012). "The convex geometry of linear inverse problems". *Found. Comput. Math.* 12.6, pp. 805–849. MR: 2989474 (cit. on p. 2963).

Scott Chen and David L Donoho (1995). "Examples of basis pursuit". In: *Wavelet Applications in Signal and Image Processing III*. Vol. 2569. International Society for Optics and Photonics, pp. 564–575 (cit. on p. 2958).

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová (2011). "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". *Physical Review E* 84.6, p. 066106 (cit. on p. 2965).

Yash Deshpande, Emmanuel Abbe, and Andrea Montanari (2017). "Asymptotic mutual information for the balanced binary stochastic block model". *Inf. Inference* 6.2, pp. 125–170. MR: 3671474 (cit. on p. 2966).

Yash Deshpande and Andrea Montanari (2015). "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time". *Found. Comput. Math.* 15.4, pp. 1069–1128. MR: 3371378 (cit. on pp. 2970, 2971).

Jian Ding, Allan Sly, and Nike Sun (2015). "Proof of the satisfiability conjecture for large k [extended abstract]". In: *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 59–68. MR: 3388183 (cit. on p. 2973).

David L. Donoho (2006). "Compressed sensing". *IEEE Trans. Inform. Theory* 52.4, pp. 1289–1306. MR: 2241189 (cit. on p. 2958).

David L. Donoho, Arian Maleki, and Andrea Montanari (2009). "Message Passing Algorithms for Compressed Sensing". *Proceedings of the National Academy of Sciences* 106, pp. 18914–18919 (cit. on p. 2972).

– (2011). "The noise-sensitivity phase transition in compressed sensing". *IEEE Trans. Inform. Theory* 57.10, pp. 6920–6941. MR: 2882271 (cit. on p. 2961).

David Donoho and Andrea Montanari (2016). "High dimensional robust M-estimation: asymptotic variance via approximate message passing". *Probab. Theory Related Fields* 166.3-4, pp. 935–969. MR: 3568043 (cit. on p. 2962).

David Donoho and Jared Tanner (2009). "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing". *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 367.1906. With electronic supplementary materials available online, pp. 4273–4293. MR: 2546388 (cit. on p. 2962).

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu (2013). "On robust regression with high-dimensional predictors". *Proceedings of the National Academy of Sciences* 110.36, pp. 14557–14562 (cit. on p. 2962).

Uriel Feige and Robert Krauthgamer (2003). "The probable value of the Lovász-Schrijver relaxations for maximum independent set". *SIAM J. Comput.* 32.2, pp. 345–370. MR: 1969394 (cit. on p. 2970).

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao (2013). "Statistical algorithms and a lower bound for detecting planted cliques". In: *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*. ACM, New York, pp. 655–664. MR: 3210827 (cit. on p. 2970).

Delphine Féral and Sandrine Péché (2007). "The largest eigenvalue of rank one deformation of large Wigner matrices". *Comm. Math. Phys.* 272.1, pp. 185–228. MR: 2291807 (cit. on p. 2965).

R. G. Gallager (1962). "Low-density parity-check codes". *IRE Trans.* IT-8, pp. 21–28. MR: 0136009 (cit. on p. 2971).

Carl Friedrich Gauss (2011). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Cambridge Library Collection. Reprint of the 1809 original. Cambridge University Press, Cambridge, pp. xii+228+21. MR: 2858122 (cit. on p. 2958).

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". *Ann. Statist.* 42.3, pp. 1166–1202. MR: 3224285 (cit. on p. 2962).

Michel X. Goemans and David P. Williamson (1995). "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". *J. Assoc. Comput. Mach.* 42.6, pp. 1115–1145. MR: 1412228 (cit. on p. 2968).

Y. Gordon (1988). "On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$". In: *Geometric aspects of functional analysis (1986/87)*. Vol. 1317. Lecture Notes in Math. Springer, Berlin, pp. 84–106. MR: 950977 (cit. on p. 2963).

Dongning Guo, Shlomo Shamai, and Sergio Verdú (2005). "Mutual information and minimum mean-square error in Gaussian channels". *IEEE Trans. Inform. Theory* 51.4, pp. 1261–1282. MR: 2241490 (cit. on p. 2963).

Peter J. Huber (1964). "Robust estimation of a location parameter". *Ann. Math. Statist.* 35, pp. 73–101. MR: 0161415 (cit. on p. 2962).

– (1973). "Robust regression: asymptotics, conjectures and Monte Carlo". *Ann. Statist.* 1, pp. 799–821. MR: 0356373 (cit. on p. 2962).

Peter J. Huber and Elvezio M. Ronchetti (2009). *Robust statistics*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, xvi+354 pp. + loose erratum. MR: 2488795 (cit. on p. 2962).

Adel Javanmard and Andrea Montanari (2013). "State evolution for general approximate message passing algorithms, with applications to spatial coupling". *Inf. Inference* 2.2, pp. 115–144. MR: 3311445 (cit. on p. 2973).

– (2014a). "Confidence intervals and hypothesis testing for high-dimensional regression". *J. Mach. Learn. Res.* 15, pp. 2869–2909. MR: 3277152 (cit. on p. 2962).

– (2014b). "Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory". *IEEE Trans. Inform. Theory* 60.10, pp. 6522–6554. MR: 3265038 (cit. on p. 2962).

– (2015). "De-biasing the Lasso: Optimal Sample Size for Gaussian Designs". arXiv: 1508.02757 (cit. on p. 2962).

Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi (2016). "Phase transitions in semidefinite relaxations". *Proc. Natl. Acad. Sci. USA* 113.16, E2218–E2223. MR: 3494080 (cit. on p. 2968).

Mark Jerrum (1992). "Large cliques elude the Metropolis process". *Random Structures Algorithms* 3.4, pp. 347–359. MR: 1179827 (cit. on p. 2970).

Iain M. Johnstone (2001). "On the distribution of the largest eigenvalue in principal components analysis". *Ann. Statist.* 29.2, pp. 295–327. MR: 1863961 (cit. on p. 2965).

Ravi Kannan, Santosh Vempala, and Adrian Vetta (2004). "On clusterings: good, bad and spectral". *J. ACM* 51.3, pp. 497–515. MR: 2145863 (cit. on p. 2958).

Noureddine El Karoui (2013). "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results". arXiv: 1311.2445 (cit. on p. 2962).

Scott Kirkpatrick and David Sherrington (1978). "Infinite-ranged models of spin-glasses". *Physical Review B* 17.11, pp. 4384–4403 (cit. on p. 2973).

Antti Knowles and Jun Yin (2013). "The isotropic semicircle law and deformation of Wigner matrices". *Comm. Pure Appl. Math.* 66.11, pp. 1663–1750. MR: 3103909 (cit. on p. 2965).

Daphne Koller and Nir Friedman (2009). *Probabilistic graphical models*. Adaptive Computation and Machine Learning. Principles and techniques. MIT Press, Cambridge, MA, pp. xxxvi+1231. MR: 2778120 (cit. on p. 2971).

Satish Babu Korada and Andrea Montanari (2011). "Applications of the Lindeberg principle in communications and statistical learning". *IEEE Trans. Inform. Theory* 57.4, pp. 2440–2450. MR: 2809100 (cit. on p. 2962).

Lev Davidovich Landau (1937). "On the theory of phase transitions". *Ukr. J. Phys.* 7, pp. 19–32 (cit. on p. 2973).

Marc Lelarge and Léo Miolane (2016). "Fundamental limits of symmetric low-rank matrix estimation". arXiv: 1611.03888 (cit. on pp. 2967, 2969).

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová (2017). "Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications". *J. Stat. Mech. Theory Exp.* 7, pp. 073403, 86. arXiv: 1701.00858. MR: 3683819 (cit. on p. 2969).

H Brendan McMahan et al. (2013). "Ad click prediction: a view from the trenches". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1222–1230 (cit. on p. 2958).

Marc Mézard and Andrea Montanari (2009). *Information, physics, and computation*. Oxford Graduate Texts. Oxford University Press, Oxford, pp. xiv+569. MR: 2518205 (cit. on pp. 2969, 2971).

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro (1987). *Spin glass theory and beyond*. Vol. 9. World Scientific Lecture Notes in Physics. World Scientific Publishing Co., Inc., Teaneck, NJ, pp. xiv+461. MR: 1026102 (cit. on p. 2973).

Léo Miolane (2017). "Fundamental limits of low-rank matrix estimation: the non-symmetric case". arXiv: 1702.00473 (cit. on p. 2967).

Andrea Montanari (2012). "Graphical Models Concepts in Compressed Sensing". In: *Compressed Sensing: Theory and Applications*. Ed. by Y.C. Eldar and G. Kutyniok. Cambridge University Press (cit. on p. 2961).

Andrea Montanari and Subhabrata Sen (2016). "Semidefinite programs on sparse random graphs and their application to community detection". In: *STOC'16—Proceedings of*

the 48th Annual ACM SIGACT Symposium on Theory of Computing. ACM, New York, pp. 814–827. MR: 3536616 (cit. on p. 2968).

Andrea Montanari and Ramji Venkataramanan (2017). "Estimation of Low-Rank Matrices via Approximate Message Passing". arXiv: 1711.01682 (cit. on pp. 2966, 2969, 2973).

Yu. Nesterov (1998). "Semidefinite relaxation and nonconvex quadratic optimization". *Optim. Methods Softw.* 9.1-3, pp. 141–160. MR: 1618100 (cit. on p. 2968).

Samet Oymak and Joel A. Tropp (2015). "Universality laws for randomized dimension reduction, with applications". arXiv: 1511.09433 (cit. on p. 2962).

Dmitry Panchenko (2013). *The Sherrington-Kirkpatrick model*. Springer Monographs in Mathematics. Springer, New York, pp. xii+156. MR: 3052333 (cit. on p. 2973).

Giorgio Parisi (1979). "Infinite number of order parameters for spin-glasses". *Phys. Rev. Lett.* 43.23, p. 1754 (cit. on p. 2973).

Debashis Paul (2007). "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". *Statist. Sinica* 17.4, pp. 1617–1642. MR: 2399865 (cit. on p. 2965).

Galen Reeves and Henry D Pfister (2016). "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact". In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, pp. 665–669 (cit. on p. 2963).

Tom Richardson and Rüdiger Urbanke (2008). *Modern coding theory*. Cambridge University Press, Cambridge, pp. xvi+572. MR: 2494807 (cit. on p. 2971).

Cyrille Rossant et al. (2016). "Spike sorting for large, dense electrode arrays". *Nature neuroscience* 19.4, pp. 634–641 (cit. on p. 2958).

Cynthia Rush and Ramji Venkataramanan (2016). "Finite-sample analysis of approximate message passing". In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, pp. 755–759 (cit. on p. 2973).

Philip Schniter, Sundeep Rangan, and Alyson K Fletcher (2016). "Vector approximate message passing for the generalized linear model". In: *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, pp. 1525–1529 (cit. on p. 2973).

Shirish Krishnaj Shevade and S Sathiya Keerthi (2003). "A simple and efficient algorithm for gene selection using sparse logistic regression". *Bioinformatics* 19.17, pp. 2246–2253 (cit. on p. 2958).

A. J. Stam (1959). "Some inequalities satisfied by the quantities of information of Fisher and Shannon". *Information and Control* 2, pp. 101–112. MR: 0109101 (cit. on p. 2963).

Michel Talagrand (2007). "Mean field models for spin glasses: some obnoxious problems". In: *Spin glasses*. Vol. 1900. Lecture Notes in Math. Springer, Berlin, pp. 63–80. MR: 2309598 (cit. on p. 2973).

David J Thouless, Philip W Anderson, and Robert G Palmer (1977). "Solution of'solvable model of a spin glass'". *Philosophical Magazine* 35.3, pp. 593–601 (cit. on p. 2972).

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi (2015). "Regularized linear regression: A precise analysis of the estimation error". In: *Conference on Learning Theory*, pp. 1683–1709 (cit. on p. 2963).

Robert Tibshirani (1996). "Regression shrinkage and selection via the lasso". *J. Roy. Statist. Soc. Ser. B* 58.1, pp. 267–288. MR: 1379242 (cit. on p. 2958).

Antonia M Tulino, Giuseppe Caire, Sergio Verdu, and Shlomo Shamai (2013). "Support recovery with sparsely sampled free random matrices". *IEEE Transactions on Information Theory* 59.7, pp. 4243–4271 (cit. on p. 2962).

Lanhui Wang and Amit Singer (2013). "Exact and stable recovery of rotations for robust synchronization". *Inf. Inference* 2.2, pp. 145–193. MR: 3311446 (cit. on p. 2965).

Cun-Hui Zhang and Stephanie S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76.1, pp. 217–242. MR: 3153940 (cit. on p. 2962).

Received 2017-12-01.

Andrea Montanari
Department of Electrical Engineering and Department of Statistics
Stanford University
montanari@stanford.edu