The LASSO risk for gaussian matrices

Mohsen Bayati* and Andrea Montanari*,†

Abstract

We consider the problem of learning a coefficient vector $x_0 \in \mathbb{R}^N$ from noisy linear observation $y = Ax_0 + w \in \mathbb{R}^n$. In many contexts (ranging from model selection to image processing) it is desirable to construct a sparse estimator $\widehat{x}$. In this case, a popular approach consists in solving an ℓ_1-penalized least squares problem known as the LASSO or Basis Pursuit DeNoising (BPDN).

For sequences of matrices A of increasing dimensions, with independent gaussian entries, we prove that the normalized risk of the LASSO converges to a limit, and we obtain an explicit expression for this limit. Our result is the first rigorous derivation of an explicit formula for the asymptotic mean squared error of the LASSO for random instances. The proof technique is based on the analysis of AMP, a recently developed efficient algorithm, that is inspired from graphical models ideas.

Simulations on real data matrices suggest that our results can be relevant in a broad array of practical applications.

1 Introduction

Let $x_0 \in \mathbb{R}^N$ be an unknown vector, and assume that a vector $y \in \mathbb{R}^n$ of noisy linear measurements of x_0 is available. The problem of reconstructing x_0 from such measurements arises in a number of disciplines, ranging from statistical learning to signal processing. In many contexts the measurements are modeled by

$$y = Ax_0 + w\,, \tag{1.1}$$

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

The LASSO or Basis Pursuit Denoising (BPDN) is a method for reconstructing the unknown vector x_0 given y, A, and is particularly useful when one seeks sparse solutions. For given A, y, one considers the cost functions $\mathcal{C}_{A,y} : \mathbb{R}^N \to \mathbb{R}$ defined by

$$\mathcal{C}_{A,y}(x) = \frac{1}{2}\,\|y - Ax\|^2 + \lambda\|x\|_1\,, \tag{1.2}$$

with $\lambda > 0$. The original signal is estimated by

$$\widehat{x}(\lambda; A, y) = \mathrm{argmin}_x\, \mathcal{C}_{A,y}(x)\,. \tag{1.3}$$

*Department of Electrical Engineering, Stanford University

†Department of Statistics, Stanford University

In what follows we shall often omit the arguments A, y (and occasionally λ) from the above notations. We will also use $\widehat{x}(\lambda; N)$ to emphasize the N-dependence. Further $\|v\|_p \equiv (\sum_{i=1}^m v_i^p)^{1/p}$ denotes the ℓ_p-norm of a vector $v \in \mathbb{R}^m$ (the subscript p will often be omitted if $p = 2$).

A large and rapidly growing literature is devoted to developing fast algorithms for solving the optimization problem (1.3) and characterizing the performances and optimality of the estimator $\widehat{x}$. We refer to Section 1.3 for an unavoidably incomplete overview.

Despite such substantial effort, and many remarkable achievements, our understanding of (1.3) is not even comparable to the one we have of more classical topics in statistics and estimation theory. For instance, the best bound on the mean squared error (MSE) of the estimator (1.3), i.e. on the quantity $N^{-1}\|\widehat{x} - x_0\|^2$, was proved by Candes, Romberg and Tao [CRT06] (who in fact did not consider the LASSO but a related optimization problem). Their result estimates the mean squared error only up to an unknown numerical multiplicative factor. Work by Candes and Tao [CT07] on the analogous *Dantzig selector*, upper bounds the mean squared error up to a factor $C \log N$, under somewhat different assumptions.

The objective of this paper is to complement this type of 'rough but robust' bounds by proving *asymptotically exact* expressions for the mean square error. Our asymptotic result holds almost surely for sequences of random matrices A with fixed aspect ratio and independent gaussian entries. While this setting is admittedly specific, the careful study of such matrix ensembles has a long tradition both in statistics and communications theory and has spurred many insights [Joh06, Tel99]. Further, we carried out simulations on real data matrices with continuous entries (gene expression data) and binary feature matrices (hospital medical records). The results appear to be quite encouraging.

Although our rigorous results are asymptotic in the problem dimensions, numerical simulations have shown that they are accurate already on problems with a few hundreds of variables. Further, they seem to enjoy a remarkable *universality* property and to hold for a fairly broad family of matrices [DMM10]. Both these phenomena are analogous to ones in random matrix theory, where delicate asymptotic properties of gaussian ensembles were subsequently proved to hold for much broader classes of random matrices. Also, asymptotic statements in random matrix theory have been replaced over time by concrete probability bounds in finite dimensions. Of course the optimization problem (1.2) is not immediately related to spectral properties of the random matrix A. As a consequence, universality and non-asymptotic results in random matrix theory cannot be directly exported to the present problem. Nevertheless, we expect such developments to be foreseeable.

Our proofs are based on the analysis of an efficient iterative algorithm first proposed by [DMM09], and called AMP, for approximate message passing. The algorithm is inspired by belief-propagation on graphical models; although the resulting iteration is significantly simpler (and scales linearly in the number of nodes). Extensive simulations [DMM10] showed that, in a number of settings, AMP performances are statistically indistinguishable to the ones of LASSO, while its complexity is essentially as low as the one of the simplest greedy algorithms.

The proof technique just described is new. Earlier literature analyzes the convex optimization problem (1.3) –or similar problems– by a clever construction of an approximate optimum, or of a dual witness. Such constructions are largely explicit. Here instead we prove an asymptotically exact characterization of a rather non-trivial iterative algorithm. The algorithm is then proved to converge to the exact optimum.

1.1 Definitions

In order to define the AMP algorithm, we denote by $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ the soft thresholding function

$$\eta(x;\theta) = \begin{cases} x-\theta & \text{if } x > \theta, \\ 0 & \text{if } -\theta \le x \le \theta, \\ x+\theta & \text{otherwise.} \end{cases} \tag{1.4}$$

The algorithm constructs a sequence of estimates $x^t \in \mathbb{R}^N$, and residuals $z^t \in \mathbb{R}^n$, according to the iteration

$$\begin{aligned} x^{t+1} &= \eta(A^* z^t + x^t; \theta_t), \\ z^t &= y - Ax^t + \frac{1}{\delta} z^{t-1} \left\langle \eta'(A^* z^{t-1} + x^{t-1}; \theta_{t-1}) \right\rangle , \end{aligned} \tag{1.5}$$

initialized with $x^0 = 0 \in \mathbb{R}^N$. Here A^* denotes the transpose of matrix A, $\delta \equiv n/N$, and $\eta'(\,\cdot\,;\,\cdot\,)$ is the derivative of the soft thresholding function with respect to its first argument. Given a scalar function f and a vector $u \in \mathbb{R}^m$, we let $f(u)$ denote the vector $(f(u_1), \dots, f(u_m)) \in \mathbb{R}^m$ obtained by applying f componentwise. Finally $\langle u \rangle \equiv m^{-1} \sum_{i=1}^m u_i$ is the average of the vector $u \in \mathbb{R}^m$.

As already mentioned, we will consider sequences of instances of increasing sizes, along which the LASSO behavior has a non-trivial limit.

Definition 1. *The sequence of instances $\{x_0(N), w(N), A(N)\}_{N \in \mathbb{N}}$ indexed by N is said to be a* converging sequence *if $x_0(N) \in \mathbb{R}^N$, $w(N) \in \mathbb{R}^n$, $A(N) \in \mathbb{R}^{n \times N}$ with $n = n(N)$ is such that $n/N \to \delta \in (0, \infty)$, and in addition the following conditions hold:*

(a) *The empirical distribution of the entries of $x_0(N)$ converges weakly to a probability measure p_{X_0} on $\mathbb{R}$ with bounded second moment. Further $N^{-1} \sum_{i=1}^N x_{0,i}(N)^2 \to \mathbb{E}_{p_{X_0}}\{X_0^2\}$.*

(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 $n^{-1} \sum_{i=1}^n w_i(N)^2 \to \mathbb{E}_{p_W}\{W^2\}$.*

(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 $[N] \equiv \{1, 2, \dots, N\}$.*

Let us stress that our proof only applies to a subclass of converging sequences, namely for gaussian measurement matrices $A(N)$. The notion of converging sequences is however important since it defines a class of problem instances to which the ideas developed below might be generalizable. Also, while the measurement matrices $A(N)$ will be random, the signal $x_0(N)$, and noise vectors $w(N)$ will be deterministic.

For a converging sequence of instances, and an arbitrary sequence of thresholds $\{\theta_t\}_{t \ge 0}$ (independent of N), the asymptotic behavior of the recursion (1.5) can be characterized as follows.

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

$$\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \theta_t)\,, \tag{1.6}$$

$$\mathsf{F}(\tau^2, \theta) \equiv \sigma^2 + \frac{1}{\delta} \mathbb{E}\{\, [\eta(X_0 + \tau Z; \theta) - X_0]^2 \}\,, \tag{1.7}$$

where $Z \sim \mathsf{N}(0,1)$ is independent of X_0. Notice that the function F depends implicitly on the law p_{X_0}. We will see later that the quantity $A^* z^t + x^t$ has the same distribution as $X_0 + \tau_t Z$. In other words, τ_t^2 is the MSE of the estimator $A^* z^t + x^t$ for x_0.

We say a function $\psi : \mathbb{R}^2 \to \mathbb{R}$ is *pseudo-Lipschitz* if there exist a constant $L > 0$ such that for all $x, y \in \mathbb{R}^2$: $|\psi(x) - \psi(y)| \leq L(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2$. (This is a special case of the definition used in [BM11] where such a function is called pseudo-Lipschitz *of order 2*.)

The next proposition that was conjectured in [DMM09] and proved in [BM11] shows that the behavior of AMP can be tracked by the above one dimensional recursion. We often refer to this prediction by *state evolution*.

Theorem 1.1 ([BM11])**.** *Let $\{x_0(N), w(N), A(N)\}_{N\in\mathbb{N}}$ be a converging sequence of instances with the entries of $A(N)$ iid normal with mean 0 and variance $1/n$ and let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\big(x_i^{t+1}, x_{0,i}\big) = \mathbb{E}\Big\{\psi\big(\eta(X_0 + \tau_t Z; \theta_t), X_0\big)\Big\}\,, \tag{1.8}$$

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

In order to establish the connection with the LASSO, a specific policy has to be chosen for the thresholds $\{\theta_t\}_{t\geq 0}$. Throughout this paper we will take $\theta_t = \alpha\tau_t$ with α is fixed. In other words, the sequence $\{\tau_t\}_{t\geq 0}$ is given by the recursion

$$\tau_{t+1}^2 = \mathsf{F}(\tau_t^2, \alpha\tau_t)\,. \tag{1.9}$$

This choice enjoys several convenient properties [DMM09]. In particular the sequence $\{\tau_t\}$ always converges to the largest solution of the fixed point equation $\tau^2 = \mathsf{F}(\tau^2, \alpha\tau)$. Further, it is a very natural choice from an intuitive point of view. Consider indeed the AMP recursion (1.5). At each step we construct a vector of 'effective observations' $y^t = x^t + A^* z^t \in \mathbb{R}^N$. This can be regarded as a noisy version of the signal x_0, whereby each entry of x_0 has been corrupted by Gaussian noise with mean 0 and variance τ_t^2. Indeed, as witnessed by Theorem 1.1, y^t is asymptotically distributed as $x_0 + w_t$ with $w^t \sim \mathsf{N}(0, \tau_t \mathbf{I}_{N\times N})$ (this statement holds in the sense of finite-dimensional marginals). Hence, it is very natural to obtain a refined estimate by applying the soft thresholding denoiser $\eta(\,\cdot\,;\theta_t)$ componentwise to y_t, which is exactly what happens in the first equation in (1.5). This denoiser shrinks component y_i^t to 0 if $|y_i^t| \leq \theta_t$. The interpretation is that any entry above θ_t is regarded as pure noise. Obviously this suggests to choose θ_t proportional to the standard deviation of the effective noise, τ_t. This is indeed confirmed by a careful mathematical analysis: choosing $\theta_t = \alpha\tau_t$ is minimax optimal, for a suitable choice of the proportionality constant α [DJ94, DJ98, DMM09].

Let us finally discuss why there should be any relation at all between the AMP algorithm (1.5) and the solution of the LASSO. Assume that $\theta_t \to \theta$, and that (x, z) is a fixed point of the corresponding AMP iteration. Let $\omega = \delta^{-1}\langle\eta'(x + A^* z; \theta)\rangle$. Then the fixed point condition reads

$$x = \eta(x + A^* z; \theta)\,, \tag{1.10}$$

$$z = y - Ax + \omega\, z\,. \tag{1.11}$$

Notice that $x = \eta(r; \theta)$ if and only if there exists $v(x) \in \partial\|x\|_1$ such that $x + \theta v(x) = r$ (here ∂f denotes the subgradient of the function f). It follows that the fixed point condition can be rewritten as

$$A^*(y - Ax) = \theta(1-\omega)\, v\,, \qquad v \in \partial \|x\|_1\,. \tag{1.12}$$

Comparing with the stationarity condition for the LASSO cost function (1.2) we obtain the following.

Lemma 1.2. *Any fixed point $x^t = x$ of the AMP iteration with $\theta_t = \theta$ is a minimizer of the LASSO cost function with*

$$\lambda = \theta\Big\{1 - \frac{1}{\delta}\langle \eta'(x + A^* z; \theta)\rangle\Big\}\,. \tag{1.13}$$

1.2 Main result

Before stating our results, we have to describe a *calibration* mapping between α and λ that was introduced in [DMM10]. This mapping is necessary since in the analysis of AMP α plays the role of λ. In other words, it can be viewed as regularization parameter and controls sparsity of AMP estimates. In particular, we will show that there exist a one-to-one (monotone) function between values of α and λ.

1.2.1 Calibration between α and λ

Let us start by stating some convenient properties of the state evolution recursion.

Proposition 1.3 ([DMM09])**.** *Let $\alpha_{\min} = \alpha_{\min}(\delta)$ be the unique non-negative solution of the equation*

$$(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \frac{\delta}{2}\,, \tag{1.14}$$

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 $\lim_{t\to\infty} \tau_t = \tau_*(\alpha)$. Further the convergence takes place for any initial condition and is monotone. Finally $\big|\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau^2, \alpha\tau)\big| < 1$ at $\tau = \tau_*$.

For greater convenience of the reader, a proof of this statement is provided in Appendix A.1.

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

$$\lambda(\alpha) \equiv \alpha\tau_*\left[1 - \frac{1}{\delta}\mathbb{E}\big\{\eta'(X_0 + \tau_* Z; \alpha\tau_*)\big\}\right]\,. \tag{1.15}$$

This function defines a correspondence (calibration) between the threshold $\alpha\tau_*$ and the regularization parameter λ. It should be intuitively clear that larger λ corresponds to larger thresholds and hence larger α since both cases yield smaller estimates of x_0. The specific choice in Eq. (1.15) is motivated by Lemma 1.2.

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

$$\alpha(\lambda) \in \big\{\, a \in (\alpha_{\min}, \infty)\,:\, \lambda(a) = \lambda \big\}\,. \tag{1.16}$$

The next result implies that the set on the right-hand side is non-empty and therefore the function $\lambda \mapsto \alpha(\lambda)$ is well defined.

Proposition 1.4 ([DMM10]). *The function $\alpha \mapsto \lambda(\alpha)$ is continuous on the interval $(\alpha_{\min}, \infty)$ with $\lambda(\alpha_{\min}+) = -\infty$ and $\lim_{\alpha\to\infty} \lambda(\alpha) = \infty$.*

Therefore the function $\lambda \mapsto \alpha(\lambda)$ satisfying Eq. (1.16) exists.

A proof of this statement is provided in Section A.2. We will denote by $\mathcal{A} = \alpha((0,\infty))$ the image of the function α. Notice that the definition of α is *a priori* not unique. We will see that uniqueness follows from our main theorem.

Examples of the mappings $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$, $\alpha \mapsto \tau_*(\alpha)$ and $\alpha \mapsto \lambda(\alpha)$ are presented in Figures 1, 2, and 3 respectively.

1.2.2 Main results

We can now state our main result.

Theorem 1.5. *Let $\{x_0(N), w(N), A(N)\}_{N\in\mathbb{N}}$ be a converging sequence of instances with the entries of $A(N)$ iid normal with mean 0 and variance $1/n$. Denote by $\widehat{x}(\lambda; N)$ the* LASSO *estimator for instance $(x_0(N), w(N), A(N))$, with $\sigma^2, \lambda > 0$, $\mathbb{P}\{X_0 \neq 0\} > 0$ and let $\psi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\big(\widehat{x}_i, x_{0,i}\big) = \mathbb{E}\Big\{\psi\big(\eta(X_0 + \tau_* Z; \theta_*), X_0\big)\Big\}\,, \tag{1.17}$$

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

Let us emphasize oonce more that the vectors $x_0(N)$, $w(N)$ are deterministic in this statement, and 'almost surely' is understood with respect to the choice of $A(N)$.

As a corollary, using function $\psi(a,b) \equiv (a-b)^2$ we obtain:

Corollary 1.6. *Assume the hypothesis of Theorem 1.5. Let $\widehat{x}(\lambda; N)$ be the* LASSO *estimator for instance $(x_0(N), w(N), A(N))$. Then, almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \|x_0 - \widehat{x}(\lambda; N)\|^2 = \mathbb{E}\Big\{\,[\eta(X_0 + \tau_* Z; \theta_*) - X_0]^2\,\Big\} = \delta(\tau_*^2 - \sigma^2)\,,$$

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

As a second corollary of Theorem 1.5, the function $\lambda \mapsto \alpha(\lambda)$ is indeed uniquely defined.

Corollary 1.7. *For any $\lambda, \sigma^2 > 0$ there exists a unique $\alpha > \alpha_{\min}$ such that $\lambda(\alpha) = \lambda$ (with the function $\alpha \to \lambda(\alpha)$ defined as in Eq. (1.15).*

Hence the function $\lambda \mapsto \alpha(\lambda)$ is continuous non-decreasing with $\alpha((0,\infty)) \equiv \mathcal{A} = (\alpha_0, \infty)$.

The proof of this corollary (which uses Theorem 1.5) is provided in Appendix A.3.

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 correct. On the other hand, the restrictions $\lambda, \sigma^2 > 0$, and $\mathbb{P}\{X_0 \neq 0\} > 0$ (whence $\tau_* \neq 0$ using Eq. (1.15)) are made in order to avoid technical complications due to degenerate cases. Such cases can be resolved by continuity arguments.

We prove Theorem 1.5 by proving the following result in Section 3.

Theorem 1.8. *Assume the hypotheses of Theorem 1.5. Let $\widehat{x}(\lambda; N)$ be the LASSO estimator for instance $(x_0(N), w(N), A(N))$, and denote by $\{x^t(N)\}_{t\geq 0}$ the sequence of estimates produced by* AMP. *Then*

$$\lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\|x^t(N)-\widehat{x}(\lambda;N)\|_2^2=0\,, \tag{1.18}$$

almost surely.

Let us emphasize that the statement of Theorem 1.8 requires taking the limit of infinite dimensions $N \to \infty$ *before* the limit of an infinite number of iterations $t \to \infty$. In this sense it is (informally speaking) a statement about the high-dimensional limit behavior, for a large-but-finite number of iterations. Although this is not a common setting within mathematical optimization, we think that it is particularly compelling from a compressed sensing point of view. It implies that, for any finite tolerance $\varepsilon > 0$, there exists a finite number of iterations $t_*(\varepsilon)$ such that for any fixed $t \geq t_*(\varepsilon)$, AMP has mean squared error at most ε larger than the LASSO, *with high probability as* $N \to \infty$. Further, closer analysis of the state evolution recursion [DMM09, DMM10] implies that $t_*(\varepsilon) \leq C \log(1/\varepsilon)$ for some constant C independent of the dimension, and the signal x_0, provided the under-sampling ratio δ is larger than a phase transition value δ_c. Notice that taking the high dimensional point of view yields us a considerably faster convergence than the optimum rate at fixed dimension, namely $t_*(\varepsilon) \leq C/\sqrt{\varepsilon}$ [BT09].

1.3 Related work

The LASSO was introduced in [Tib96, CD95]. Several papers provide performance guarantees for the LASSO or similar convex optimization methods [CRT06, CT07], by proving upper bounds on the resulting mean squared error. These works assume an appropriate 'isometry' condition to hold for A. While such condition hold with high probability for some random matrices, it is often difficult to verify them explicitly. Further, it is only applicable to very sparse vectors x_0. 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 appropriate '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 squared error) provides complementary information and is quite standard in many different fields.

Closer to the spirit of this paper [RFG09] derived expressions for the mean squared 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. For instance, it can be used to evaluate competing optimization methods on large scale applications, or to tune the regularization parameter λ.

Unfortunately, these results were non-rigorous and were obtained through the famously powerful 'replica method' from statistical physics [MM09].

Let us emphasize that the present paper 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 squared error is shown to be equivalent to the one achieved by a low-complexity message passing algorithm.

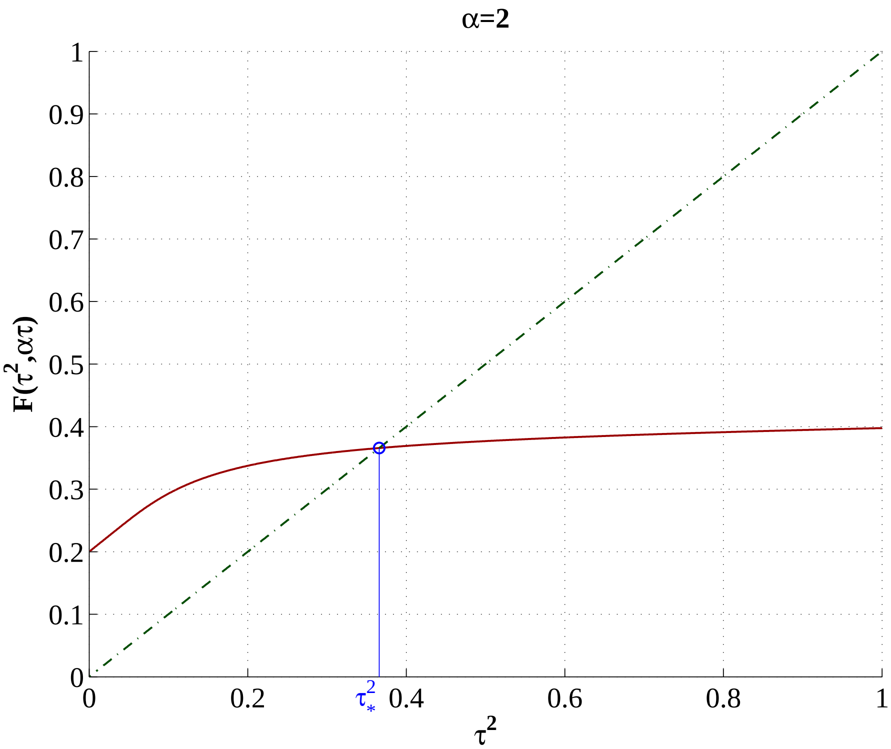

Figure 1: Mapping $\tau^2 \mapsto \mathsf{F}(\tau^2, \alpha\tau)$ for $\alpha = 2$, $\delta = 0.64$, $\sigma^2 = 0.2$, $p_{X_0}(\{+1\}) = p_{X_0}(\{-1\}) = 0.064$ and $p_{X_0}(\{0\}) = 0.872$.

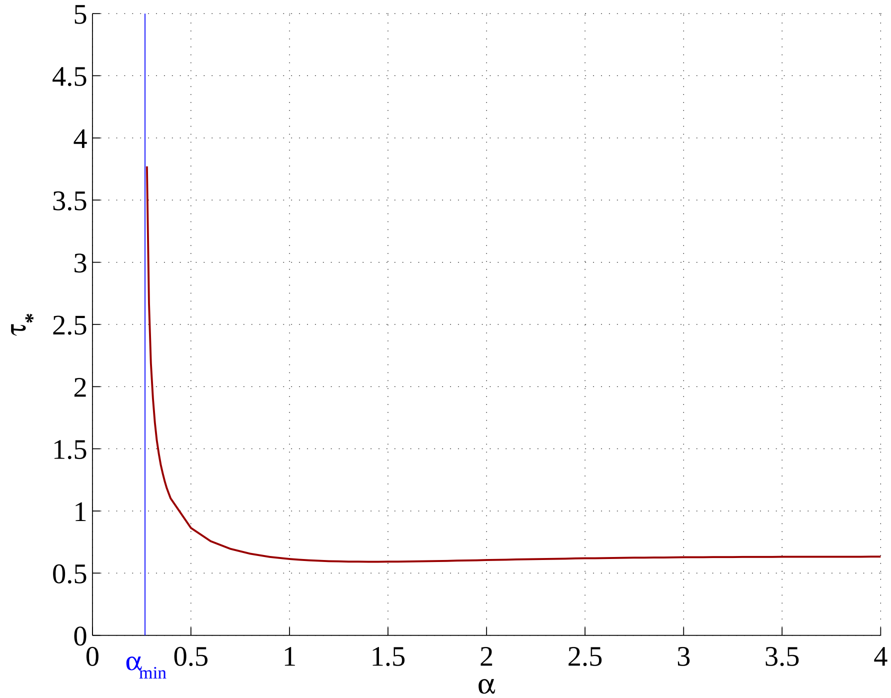

Figure 2: Mapping $\alpha \mapsto \tau_*(\alpha)$ for the same parameters δ, σ^2 and distribution p_{X_0} as in Figure 1.

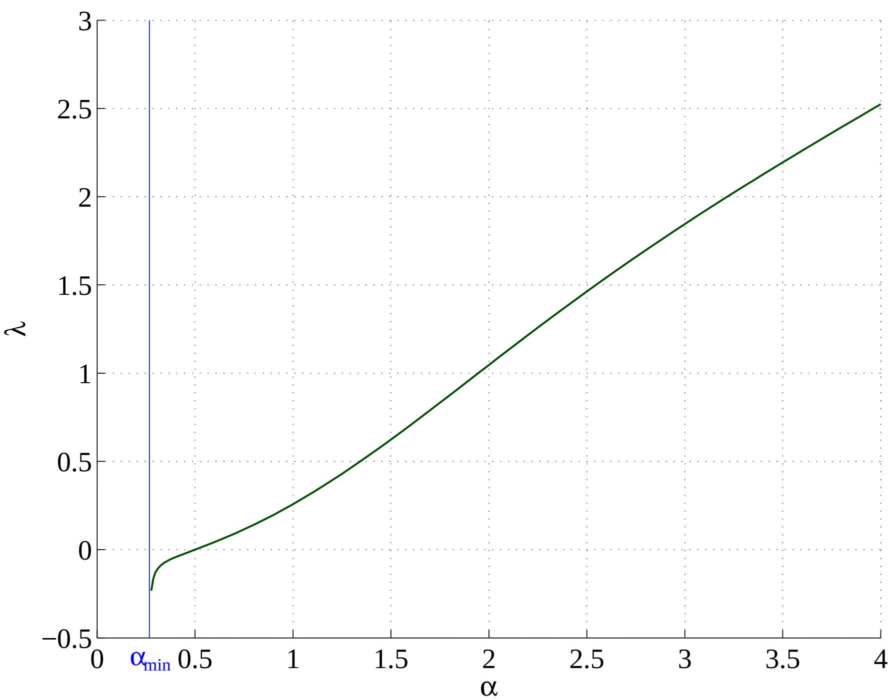

Figure 3: Mapping $\alpha \mapsto \lambda(\alpha)$ for the same parameters δ, σ^2 and distribution p_{X_0} as in Figure 1.

2 Numerical illustrations

Theorem 1.5 assumes that the entries of matrix A have iid gaussian distribution. We expect however the mean squared error prediction to be robust and hold for much larger family of matrices. 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.)

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 [DMM10, BBM10] and suggest that the result is robust and relevant already for N of the order of a few hundreds. As an illustration, we present in Figures 4-7 the outcome of such simulations for four types of real data and random matrices. 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 obtained the optimum 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 [AG07]. 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 Corollary 1.6, namely $\delta(\tau_*^2 - \sigma^2)$.

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

The four figures correspond to measurement matrices A:

- Figure 4: Data consist of 2253 measurements of expression level of 7077 genes.From this matrix we took sub-matrices A of aspect ratio δ for each N. The entries were continuous variables. We standardized all columns of A to have mean 0 and variance 1.

- Figure 5: From a data set of 1932 patient records we extracted 4833 binary features describing demographic information, medical history, lab results, medications etc. The 0-1 matrix was sparse (with only 3.1% non-zero entries). Similar to (i), for each N, the sub-matrices A with aspect ratio δ were selected and standardized.

- Figure 6: Random gaussian matrices with aspect ratio δ and iid $\mathsf{N}(0, 1/n)$ entries (as in Theorem 1.5);

- Figure 7: Random ± 1 matrices with aspect ratio δ. Each entry is independently equal to $+1/\sqrt{n}$ or $-1/\sqrt{n}$ with equal probability.

Notice the behavior appears to be essentially indistinguishable. Also the asymptotic prediction has a minimum as a function of λ. The location of this minimum can be used to select the regularization parameter. Further empirical analysis is presented in [BBM11].

3 A structural property and proof of the main results

The rest of the paper is devoted to the proof of Theorem 1.8. Section 3.2 proves a structural property that is the key tool in this proof. Section 3.3 uses this property together with a few lemmas to prove

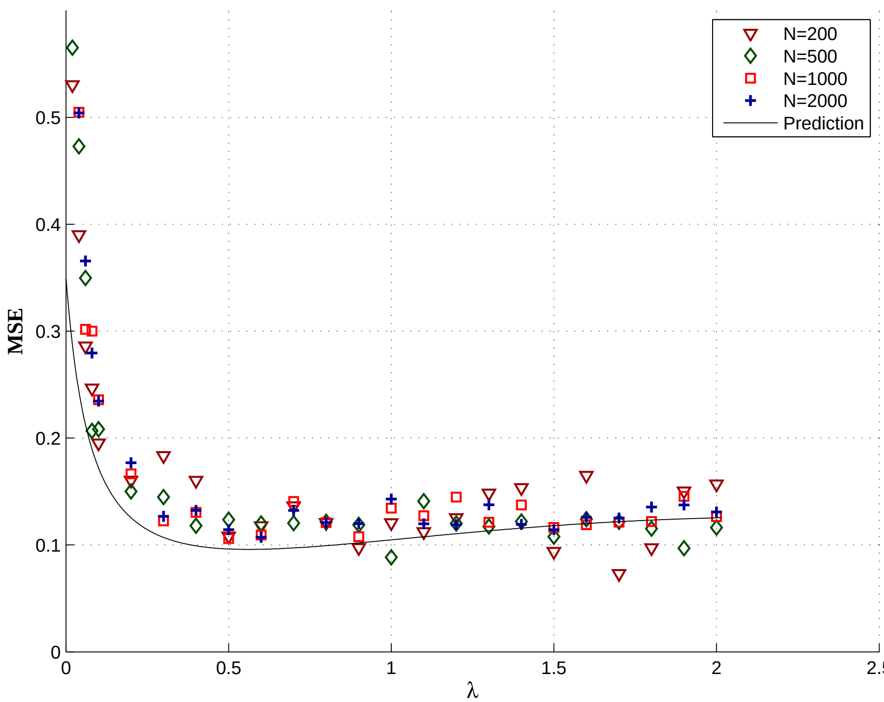

Figure 4: Mean squared error (MSE) as a function of the regularization parameter λ compared to the asymptotic prediction for $\delta = .64$ and $\sigma^2 = .2$. Here the measurement matrix A is a real valued (standardized) matrix of gene expression data. Each point in these plots is generated by finding the LASSO predictor $\widehat{x}$ using a measurement vector $y = Ax_0 + w$ for an independent signal vector x_0 and an independent noise vector w.

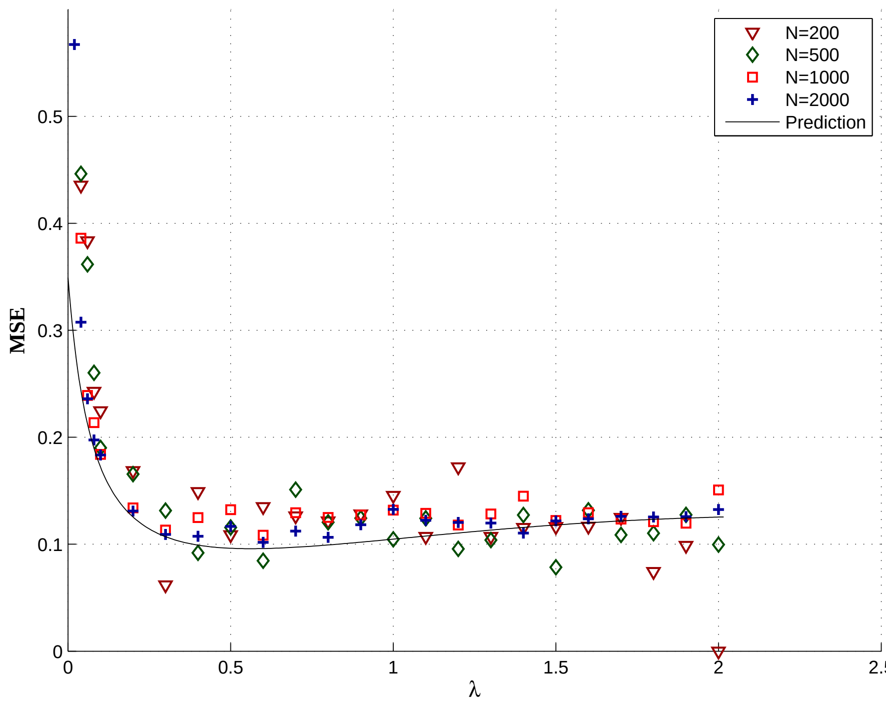

Figure 5: As in Figure 4, but the measurement matrix A is a (standardized) 0-1 feature matrix of hospital records.

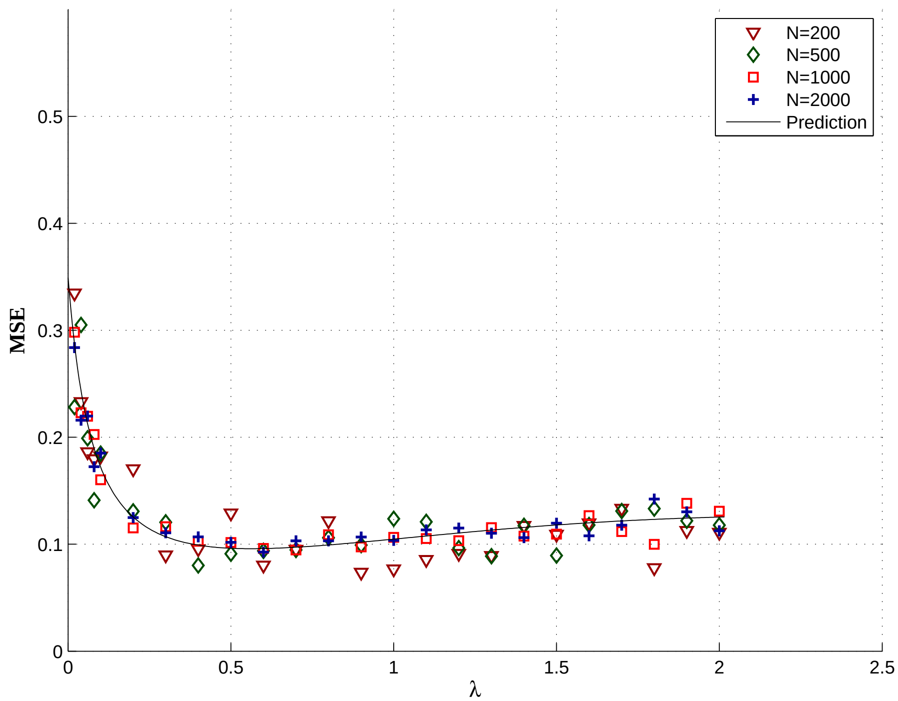

Figure 6: As in Figure 4, but the measurement matrix A has iid $\mathsf{N}(0, 1/n)$ entries. Additionally, each point in this plot uses an independent matrix A.

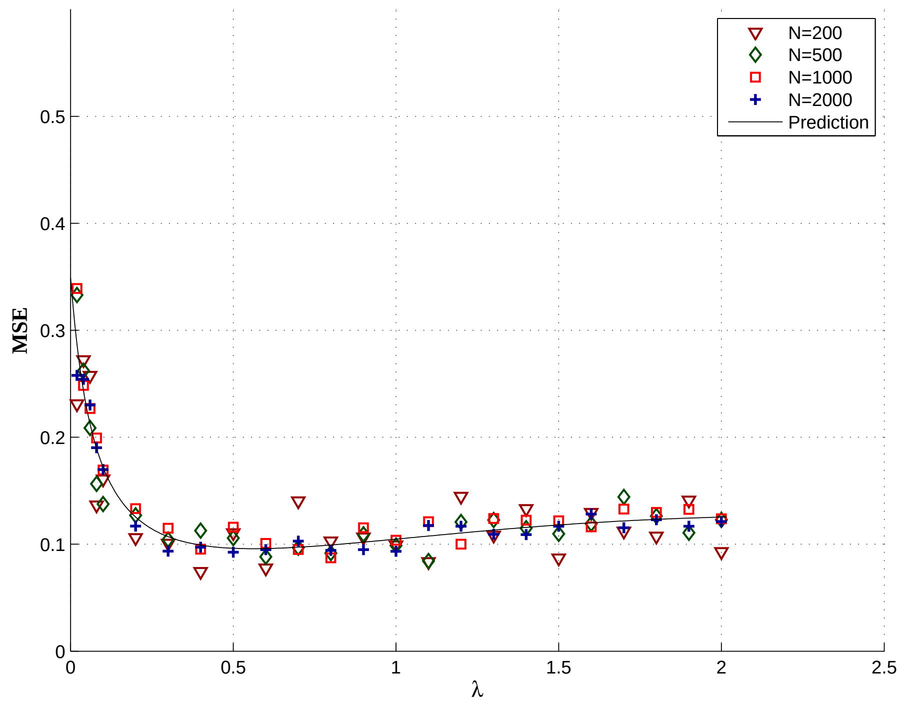

Figure 7: As in Figure 4, but the measurement matrix A has iid entries that are equal to $\pm 1/\sqrt{n}$ with equal probabilities. Similar to Figure 6, each point in this plot uses an independent matrix A.

Theorem 1.8

The proof of Theorem 1.5 follows immediately from Theorem 1.8.

Proof of Theorem 1.5. For any $t \geq 0$, we have, by the pseudo-Lipschitz property of ψ,

$$
\begin{aligned}
\left|\frac{1}{N}\sum_{i=1}^{N}\psi\big(x_i^{t+1},x_{0,i}\big)-\frac{1}{N}\sum_{i=1}^{N}\psi\big(\widehat{x}_i,x_{0,i}\big)\right| &\leq \frac{L}{N}\sum_{i=1}^{N}|x_i^{t+1}-\widehat{x}_i|\big(1+2|x_{0,i}|+|x_i^{t+1}|+|\widehat{x}_i|\big)\\
&\leq \frac{L}{N}\,\|x^{t+1}-\widehat{x}\|_2\sqrt{\sum_{i=1}^{N}\big(1+2|x_{0,i}|+|x_i^{t+1}|+|\widehat{x}_i|\big)^2}\\
&\leq L\frac{\|x^{t+1}-\widehat{x}\|_2}{\sqrt{N}}\sqrt{4+\frac{8\|x_0\|_2^2}{N}+\frac{4\|x^{t+1}\|_2^2}{N}+\frac{4\|\widehat{x}\|_2^2}{N}}\,,
\end{aligned}
$$

where the second inequality follows by Cauchy-Schwarz. Next we take the limit $N \to \infty$ followed by $t \to \infty$. The first term vanishes by Theorem 1.8. For the second term, note that $\|x_0\|_2^2/N$ remains bounded since (x_0, w, A) is a converging sequence. The two terms $\|x^{t+1}\|_2^2/N$ and $\|\widehat{x}\|_2^2/N$ also remain bounded in this limit because of state evolution (as proved in Lemma 3.2 below).

We then obtain

$$
\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi\big(\widehat{x}_i,x_{0,i}\big)=\lim_{t\to\infty}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi\big(x_i^{t+1},x_{0,i}\big)=\mathbb{E}\Big\{\psi\big(\eta(X_0+\tau_*Z;\theta_*),X_0\big)\Big\}\,,
$$

where we used Theorem 1.1 and Proposition 1.3. □

3.1 Some notations

Before continuing, we introduce some useful notations. For any non-empty subset S of $[m]$ and any $k \times m$ matrix M we refer by M_S to the k by $|S|$ sub-matrix of M that contains only the columns of M corresponding to S. The same notation is used for vectors $v \in \mathbb{R}^m$: v_S is the vector $(v_i : i \in S)$. For any vector $v \in \mathbb{R}^m$ we denote support of v by

$$
\mathrm{supp}(v) \equiv \{i \,|\, v_i \neq 0\}\,.
$$

We will also use the following scalar product for $u, v \in \mathbb{R}^m$:

$$
\langle u, v\rangle \equiv \frac{1}{m}\sum_{i=1}^{m} u_i\, v_i\,. \tag{3.1}
$$

For a matrix M we denote its minimum and maximum singular values by $\sigma_{\min}(M)$, $\sigma_{\max}(M)$ respectively. We also denote the minimum non-zero singular value of M by $\hat{\sigma}_{\min}(M)$.

The subgradient of a convex function $f : \mathbb{R}^m \to \mathbb{R}$ at point $x \in \mathbb{R}^m$ is denoted by $\partial f(x)$. In particular, remember that the subgradient of the ℓ_1 norm, $x \mapsto \|x\|_1$ is given by

$$
\partial\|x\|_1 = \big\{v \in \mathbb{R}^m \text{ such that } |v_i| \leq 1\,\forall i \text{ and } x_i \neq 0 \Rightarrow v_i = \mathrm{sign}(x_i)\big\}\,. \tag{3.2}
$$

We will generally be interested in sequences of events $\{\mathcal{E}_N\}$ indexed by the problem dimensions N. It is understood throughout that the underlying probability space is the one generated by the random matrices $A(N)$, which we take to be independent across different N. We say that such a sequence of events holds *eventually almost surely* (as $N \to \infty$) if[1] there exists a random variable N_0 such that: (i) N_0 is almost surely finite; (ii) The events $\mathcal{E}_N$ hold for all $N \geq N_0$.

3.2 A structural property of the LASSO cost function

One main challenge in the proof of Theorem 1.5 lies in the fact that the function $x \mapsto \mathcal{C}_{A,y}(x)$ is not –in general– strictly convex. Hence there can be, in principle, vectors x of cost very close to the optimum and nevertheless far from the optimum.

The following Lemma provides conditions under which this does not happen.

Lemma 3.1. *There exists a function $\xi(\varepsilon, c_1, \dots, c_5)$ such that the following happens.*

If x, $r \in \mathbb{R}^N$ satisfy the following conditions

1. $\|r\|_2 \leq c_1\sqrt{N}$;
2. $\mathcal{C}(x+r) \leq \mathcal{C}(x)$;
3. *There exists* $\mathrm{sg}(\mathcal{C}, x) \in \partial\mathcal{C}(x)$ *with* $\|\mathrm{sg}(\mathcal{C}, x)\|_2 \leq \sqrt{N}\,\varepsilon$;
4. *Let* $v \equiv (1/\lambda)[A^*(y - Ax) + \mathrm{sg}(\mathcal{C}, x)] \in \partial\|x\|_1$, *and* $S(c_2) \equiv \{i \in [N] : \; |v_i| \geq 1 - c_2\}$. *Then, for any* $S' \subseteq [N]$, $|S'| \leq c_3 N$, *we have* $\sigma_{\min}(A_{S(c_2)\cup S'}) \geq c_4$;
5. *The maximum singular value of A is bounded:* $\sigma_{\max}(A)^2 \leq c_5$.

Then $\|r\|_2 \leq \sqrt{N}\,\xi(\varepsilon, c_1, \dots, c_5)$. *Further for any* $c_1, \dots, c_5 > 0$, $\xi(\varepsilon, c_1, \dots, c_5) \to 0$ *as* $\varepsilon \to 0$.

Further, if $\ker(A) = \{0\}$, *the same conclusion holds under assumptions 1, 2, 3, 5.*

Proof. Throughout the proof we denote $\xi_1, \xi_2, \dots$ functions of the constants $c_1, \dots, c_5 > 0$ and of ε such that $\xi_i(\varepsilon) \to 0$ as $\varepsilon \to 0$ (we shall omit the dependence of ξ_i on ε).

Let $S = \mathrm{supp}(x) \subseteq [N]$. We have

$$
\begin{aligned}
0 &\overset{(a)}{\geq} \Big(\frac{\mathcal{C}(x+r) - \mathcal{C}(x)}{N}\Big) \\
&\overset{(b)}{=} \lambda\Big(\frac{\|x_S + r_S\|_1 - \|x_S\|_1}{N}\Big) + \frac{\lambda\|r_{\overline{S}}\|_1 + \frac{1}{2}\|y - Ax - Ar\|_2^2 - \frac{1}{2}\|y - Ax\|_2^2}{N} \\
&\overset{(c)}{=} \lambda\Big(\frac{\|x_S + r_S\|_1 - \|x_S\|_1}{N} - \langle \mathrm{sign}(x_S), r_S\rangle\Big) + \lambda\Big(\frac{\|r_{\overline{S}}\|_1}{N} - \langle v_{\overline{S}}, r_{\overline{S}}\rangle\Big) + \lambda\langle v, r\rangle - \langle y - Ax, Ar\rangle + \frac{\|Ar\|_2^2}{2N} \\
&\overset{(d)}{=} \lambda\Big(\frac{\|x_S + r_S\|_1 - \|x_S\|_1}{N} - \langle \mathrm{sign}(x_S), r_S\rangle\Big) + \lambda\Big(\frac{\|r_{\overline{S}}\|_1}{N} - \langle v_{\overline{S}}, r_{\overline{S}}\rangle\Big) + \langle \mathrm{sg}(\mathcal{C}, x), r\rangle + \frac{\|Ar\|_2^2}{2N}\,,
\end{aligned}
$$

where (a) follows from hypothesis (2), (c) from the fact that $v_S = \mathrm{sign}(x_S)$ since $v \in \partial\|x\|_1$ which gives

$$\langle \mathrm{sign}(x_S), r_S\rangle + \langle v_{\overline{S}}, r_{\overline{S}}\rangle = \langle v_S, r_S\rangle + \langle v_{\overline{S}}, r_{\overline{S}}\rangle = \langle v, r\rangle\,,$$

[1]Formally, if $\mathbb{P}(\cup_{\overline{N}\geq 1} \cap_{N \geq \overline{N}} \mathcal{E}_N) = 1$.

and (d) follows from the definition of (v).

Using hypothesis (1) and (3), we get by Cauchy-Schwarz

$$\lambda\Big(\frac{\|x_S + r_S\|_1 - \|x_S\|_1}{N} - \langle \mathrm{sign}(x_S), r_S\rangle\Big) + \lambda\Big(\frac{\|r_{\overline{S}}\|_1}{N} - \langle v_{\overline{S}}, r_{\overline{S}}\rangle\Big) + \frac{\|Ar\|_2^2}{2N} \le c_1\varepsilon\,. \tag{3.3}$$

Each of the three terms on the left-hand side is non-negative. The third one is trivial. The first one is non-negative since

$$\frac{\sum_{i\in S}\Big\{(x_i+r_i)\mathrm{sign}(x_i+r_i) - x_i\mathrm{sign}(x_i) - r_i\mathrm{sign}(x_i)\Big\}}{N} \quad = \quad \frac{\sum_{i\in S}(x_i+r_i)\big[\mathrm{sign}(x_i+r_i)-\mathrm{sign}(x_i)\big]}{N}\,,$$

and each $(x_i+r_i)\,[\mathrm{sign}(x_i+r_i)-\mathrm{sign}(x_i)]$ is either equal to 0 (when $\mathrm{sign}(x_i)=\mathrm{sign}(x_i+r_i)$) or equal to $2|x_i+r_i|$ otherwise. The second term in (3.3) is also non-negative since $|r_i|-v_ir_i = |r_i|[1-v_i\mathrm{sign}(r_i)]$ and $1\ge v_i\,\mathrm{sign}(r_i)$ since $|v_i|\le 1$ by definition of subgradient. Therefore,

$$\begin{aligned} \frac{\|r_{\overline{S}}\|_1}{N} - \langle v_{\overline{S}}, r_{\overline{S}}\rangle \quad &\le \quad \xi_1(\varepsilon)\,, && (3.4)\\ \|Ar\|_2^2 \quad &\le \quad N\xi_1(\varepsilon)\,. && (3.5) \end{aligned}$$

Let $V_{\|}\subseteq\mathbb{R}^N$ be the subspace of $\mathbb{R}^N$ spanned by the right singular vectors of A with singular values $\sigma_i\le c_4/2$ (including –eventually– the null space of A), and denote by $V_\perp$ the orthogonal complement of $V_{\|}$. Hence $V_\perp$ is spanned by right singular vectors of A with singular value $\sigma_i > c_4/2$. Let $\mathrm{P}_{\|}$ and $\mathrm{P}_\perp$ denote the orthogonal projectors on $V_{\|}$ and $V_\perp$. Write $r = r^\perp + r^{\|}$, with $r^{\|} = \mathrm{P}_{\|}r\in V_{\|}$ and $r^\perp = \mathrm{P}_\perp r \in V_\perp$. Also, write $A = A_{\|} + A_\perp \equiv A\mathrm{P}_{\|} + A\mathrm{P}_\perp$ (note that $A_{\|}$ and $A_\perp$ have orthogonal column spaces).

It follows from Eq. (3.5) that

$$\|A_{\|}r^{\|}\|_2^2 \le N\xi_1(\varepsilon)\,, \qquad \|A_\perp r^\perp\|_2^2 \le N\xi_1(\varepsilon)\,. \tag{3.6}$$

Since $\|A_\perp r^\perp\|_2^2 \ge (c_4^2/4)\|r^\perp\|_2^2$, we have

$$\|r^\perp\|_2^2 \le \frac{4N\xi_1(\varepsilon)}{c_4^2}\,. \tag{3.7}$$

In the case $V_{\|}=\{0\}$, the proof is concluded. In the case $V_{\|}\ne\{0\}$, we need to prove an analogous bound for $r^{\|}$. From Eq. (3.4) together with $\|r^\perp_{\overline{S}}\|_1 \le \sqrt{N}\|r^\perp_{\overline{S}}\|_2 \le \sqrt{N}\|r^\perp\|_2 \le (2N/c_4)\sqrt{\xi_1(\varepsilon)}$, we get

$$\begin{aligned} &\frac{\|r^{\|}_{\overline{S}}\|_1}{N} - \langle v_{\overline{S}}, r^{\|}_{\overline{S}}\rangle \le \xi_2(\varepsilon)\,. && (3.8)\\ &A_\perp r^{\|} = 0\,, && (3.9) \end{aligned}$$

Where (3.9) follows immediately from definition of $A_\perp$ and $r^{\|}$. Now, notice that $\overline{S}(c_2)\subseteq\overline{S}$. From Eq. (3.8) and definition of $S(c_2)$ it follows that

$$\begin{aligned} \|r^{\|}_{\overline{S}(c_2)}\|_1 &\le \frac{\|r^{\|}_{\overline{S}(c_2)}\|_1 - N\langle v_{\overline{S}(c_2)}, r^{\|}_{\overline{S}(c_2)}\rangle}{c_2} && (3.10)\\ &\le Nc_2^{-1}\xi_2(\varepsilon)\,. && (3.11) \end{aligned}$$

In particular, inequality (3.11) relies on the fact that the right hand side of (3.10) can be written as $\sum_{i\in\overline{S}(c_2)} |r_i^{\|}|(1-v_i\text{sign}(r_i^{\|})|$ where each summand is non-negative, therefore the summation increases by replacing $\sum_{i\in\overline{S}(c_2)}$ with $\sum_{i\in\overline{S}}$. Next, let us first consider the case $|\overline{S}(c_2)| \ge Nc_3/2$. Then partition $\overline{S}(c_2) = \cup_{\ell=1}^K S_\ell$, where $(Nc_3/2) \le |S_\ell| \le Nc_3$, and for each $i \in S_\ell$, $j\in S_{\ell+1}$, $|r_i^{\|}| \ge |r_j^{\|}|$. Also define $\overline{S}_+ \equiv \cup_{\ell=2}^K S_\ell \subseteq \overline{S}(c_2)$. Since, $|r_i^{\|}| \le \|r^{\|}_{S_{\ell-1}}\|_1/|S_{\ell-1}|$ holds for any $i\in S_\ell$, we have

$$\begin{aligned}
\|r^{\|}_{\overline{S}_+}\|_2^2 &= \sum_{\ell=2}^K \|r^{\|}_{S_\ell}\|_2^2 \le \sum_{\ell=2}^K |S_\ell|\Big(\frac{\|r^{\|}_{S_{\ell-1}}\|_1}{|S_{\ell-1}|}\Big)^2\\
&\le \frac{4}{Nc_3}\sum_{\ell=2}^K \|r^{\|}_{S_{\ell-1}}\|_1^2 \le \frac{4}{Nc_3}\Big(\sum_{\ell=2}^K \|r^{\|}_{S_{\ell-1}}\|_1\Big)^2\\
&\le \frac{4}{Nc_3}\,\|r^{\|}_{\overline{S}(c_2)}\|_1^2 \le \frac{4\xi_2(\varepsilon)^2}{c_2^2 c_3}\,N \equiv N\xi_3(\varepsilon)\,.
\end{aligned}$$

To conclude the proof, it is sufficient to prove an analogous bound for $\|r^{\|}_{S_+}\|_2^2$ with $S_+ = [N]\backslash\overline{S}_+ = S(c_2)\cup S_1$. Since $|S_1| \le Nc_3$, we have by hypothesis (4) that $\sigma_{\min}(A_{S_+}) \ge c_4$. By Eq. (3.9) we have $A_{\|}r^{\|} = Ar^{\|} = A_{S_+}r^{\|}_{S_+} + A_{\overline{S}_+}r^{\|}_{\overline{S}_+}$. Therefore

$$c_4^2\|r^{\|}_{S_+}\|_2^2 \le \|A_{S_+}r^{\|}_{S_+}\|_2^2 = \|A_{\overline{S}_+}r^{\|}_{\overline{S}_+} - A_{\|}r^{\|}\|_2^2 \le 2c_5\|r^{\|}_{\overline{S}_+}\|_2^2 + 2\frac{c_4^2}{4}\|r^{\|}\|_2^2\,.$$

In the last step we used triangular inequality together with the fact that $\sigma_{\max}(A_{\overline{S}_+})^2 \le c_5$ (by assumption (5)) and $\sigma_{\max}(A_{\|}) \le c_4/2$ (by construction). Using $\|r^{\|}\|_2^2 = \|r^{\|}_{S_+}\|_2^2 + \|r^{\|}_{\overline{S}_+}\|_2^2$, we get

$$\frac{c_4^2}{2}\|r^{\|}_{S_+}\|_2^2 \le \Big(2c_5 + \frac{c_4^2}{2}\Big)\|r^{\|}_{\overline{S}_+}\|_2^2 \le \Big(2c_5+\frac{c_4^2}{2}\Big)N\xi_3(\varepsilon)\,.$$

This finishes the proof when $|\overline{S}(c_2)| \ge Nc_3/2$. Note that if this assumption does not hold then we can take $\overline{S}_+ = \emptyset$ and $S_+ = [N]$. Hence, the result follows as a special case of above. □

3.3 Proof of Theorem 1.8

The proof is based on a series of Lemmas that are used to check the assumptions of Lemma 3.1

The first one is an upper bound on the ℓ_2–norm of AMP estimates, and of the LASSO estimate. Its proof is deferred to Section 5.1.

Lemma 3.2. *Under the conditions of Theorem 1.5, assume $\lambda > 0$ and $\alpha = \alpha(\lambda)$. Denote by $\widehat{x}(\lambda;N)$ the* LASSO *estimator and by $\{x^t(N)\}$ the sequence of* AMP *estimates. Then there is a constant* B *such that for all $t\ge 0$, almost surely*

$$\lim_{t\to\infty}\lim_{N\to\infty}\langle x^t(N), x^t(N)\rangle < \mathsf{B}, \tag{3.12}$$

$$\lim_{N\to\infty}\langle \widehat{x}(\lambda;N), \widehat{x}(\lambda;N)\rangle < \mathsf{B}. \tag{3.13}$$

The second Lemma implies that the estimates of AMP are approximate minima, in the sense that the cost function $\mathcal{C}$ admits a small subgradient at x^t, when t is large. The proof is deferred to Section 5.2.

Lemma 3.3. *Under the conditions of Theorem 1.5, for all t there exists a subgradient $\mathrm{sg}(\mathcal{C}, x^t)$ of $\mathcal{C}$ at point x^t such that almost surely,*

$$\lim_{t\to\infty} \lim_{N\to\infty} \frac{1}{N} \|\mathrm{sg}(\mathcal{C}, x^t)\|^2 = 0. \tag{3.14}$$

The next lemma implies that sub-matrices of A constructed using the first t iterations of the AMP algorithm are non-singular (more precisely, have singular values bounded away from 0). The proof can be found in Section 5.3.

Lemma 3.4. *Let $S \subseteq [N]$ be measurable on the σ-algebra $\mathfrak{S}_t$ generated by $\{z^0, \dots, z^{t-1}\}$ and $\{x^0 + A^* z^0, \dots, x^{t-1} + A^* z^{t-1}\}$ and assume $|S| \le N(\delta - c)$ for some $c > 0$. Then there exists $a_1 = a_1(c) > 0$ (independent of t) and $a_2 = a_2(c, t) > 0$ (depending on t and c) such that*

$$\min_{S'} \big\{ \sigma_{\min}(A_{S \cup S'}) \, : \, S' \subseteq [N] \, , \, |S'| \le a_1 N \big\} \ge a_2 \, , \tag{3.15}$$

eventually almost surely as $N \to \infty$.

We will apply this lemma to a specific choice of the set S. Namely, defining

$$v^t \equiv \frac{1}{\theta_{t-1}} (x^{t-1} + A^* z^{t-1} - x^t) \, , \tag{3.16}$$

we will then consider the set

$$S_t(\gamma) \equiv \big\{ \, i \in [N] \, : \, |v^t_i| \ge 1 - \gamma \, \big\} \, , \tag{3.17}$$

for $\gamma \in (0, 1)$. Our last lemma shows that this sequence of sets $S_t(\gamma)$ 'converges' in the following sense. The proof can be found in Section 5.4.

Lemma 3.5. *Fix $\gamma \in (0, 1)$ and let the sequence $\{S_t(\gamma)\}_{t \ge 0}$ be defined as in Eq. (3.17) above. For any $\xi > 0$ there exists $t_* = t_*(\xi, \gamma) < \infty$ such that, for all $t_2 \ge t_1 \ge t_*$ fixed, we have*

$$|S_{t_2}(\gamma) \setminus S_{t_1}(\gamma)| < N\xi \, , \tag{3.18}$$

eventually almost surely as $N \to \infty$.

The above two lemmas imply the following.

Proposition 3.6. *There exist constants $\gamma_1 \in (0, 1)$, γ_2, $\gamma_3 > 0$ and $t_{\min} < \infty$ such that, for any $t \ge t_{\min}$,*

$$\min_{S_1} \big\{ \sigma_{\min}(A_{S_t(\gamma_1) \cup S'}) \, : \, S' \subseteq [N] \, , \, |S'| \le \gamma_2 N \big\} \ge \gamma_3 \, , \tag{3.19}$$

eventually almost surely as $N \to \infty$.

Proof. First notice that, for any fixed γ, the set $S_t(\gamma)$ is measurable on $\mathfrak{S}_t$. Indeed by Eq. (1.5) $\mathfrak{S}_t$ contains $\{x^0, \dots, x^t\}$ as well, and hence it contains v^t which is a linear combination of $x^{t-1} + A^* z^{t-1}$, x^t. Finally $S_t(\gamma)$ is obviously a measurable function of v^t.

Using Lemma F.3(b) the empirical distribution of $(x_0 - A^* z^{t-1} - x^{t-1}, x_0)$ converges weakly to $(\tau_{t-1} Z, X_0)$ for $Z \sim \mathsf{N}(0,1)$ independent of $X_0 \sim p_{X_0}$. (Following the notation of [BM11], we let $h^t = x_0 - A^* z^{t-1} - x^{t-1}$.) Therefore, for any constant γ we have almost surely

$$\lim_{N\to\infty} \frac{|S_t(\gamma)|}{N} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\left\{\frac{1}{\theta_{t-1}}\left|x_i^{t-1} + [A^* z^{t-1}]_i - x_i^t\right| \ge 1-\gamma\right\}} \tag{3.20}$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\left\{\frac{1}{\theta_{t-1}}\left|x_{0,i} - h_i^t - \eta(x_{0,i} - h_i^t, \theta_{t-1})\right| \ge 1-\gamma\right\}} \tag{3.21}$$

$$= \mathbb{P}\left\{ \frac{1}{\theta_{t-1}} \big| X_0 + \tau_{t-1} Z - \eta(X_0 + \tau_{t-1} Z, \theta_{t-1}) \big| \ge 1 - \gamma \right\}. \tag{3.22}$$

The last equality follows from the weak convergence of the empirical distribution of $\{(h_i, x_{0,i})\}_{i\in[N]}$ (from Lemma F.3(b), which takes the same form as Theorem 1.8), together with the absolute continuity of the distribution of $|X_0 + \tau_{t-1} Z - \eta(X_0 + \tau_{t-1} Z, \theta_{t-1})|$.

Now, combining

$$\Big| X_0 + \tau_{t-1} Z - \eta(X_0 + \tau_{t-1} Z, \theta_{t-1}) \Big| = \begin{cases} \theta_{t-1} & \text{when} \quad |X_0 + \tau_{t-1} Z| \ge \theta_{t-1}\,, \\ |X_0 + \tau_{t-1} Z| & \text{otherwise}\,, \end{cases}$$

and Eq. (3.22) we obtain almost surely

$$\lim_{N\to\infty} \frac{|S_t(\gamma)|}{N} = \mathbb{E}\left\{\eta'(X_0 + \tau_{t-1} Z, \theta_{t-1})\right\} + \mathbb{P}\Big\{(1-\gamma) \le \frac{1}{\theta_{t-1}} |X_0 + \tau_{t-1} Z| \le 1\Big\}. \tag{3.23}$$

It is easy to see that the second term $\mathbb{P}\{1 - \gamma \le (1/\theta_{t-1})|X + \tau_{t-1} Z| \le 1\}$ converges to 0 as $\gamma \to 0$. On the other hand, using Eq. (1.15) and the fact that $\lambda(\alpha) > 0$ the first term will be strictly smaller than δ for large enough t. Hence, we can choose constants $\gamma_1 \in (0,1)$ and $c > 0$ such that

$$|S_t(\gamma_1)| < N(\delta - c)\,. \tag{3.24}$$

eventually almost surely as $N \to \infty$, for all fixed t larger than some $t_{\min,1}(c)$.

For any $t \ge t_{\min,1}(c)$ we can apply Lemma 3.4 for some $a_1(c)$, $a_2(c,t) > 0$. Fix $c > 0$ and let $a_1 = a_1(c)$ be fixed as well. Let $t_{\min} = \max(t_{\min,1}, t_*(a_1/2, \gamma_1))$ (with $t_*(\,\cdot\,)$ defined as per Lemma 3.5). Take $a_2 = a_2(c, t_{\min})$. Obviously $t \mapsto a_2(c,t)$ is non-increasing. Then we have, by Lemma 3.4

$$\min\left\{\sigma_{\min}(A_{S_{t_{\min}}(\gamma_1)\cup S'}) \;:\; S' \subseteq [N]\,,\; |S'| \le a_1 N\right\} \ge a_2\,, \tag{3.25}$$

and by Lemma 3.5

$$|S_t(\gamma_1) \setminus S_{t_{\min}}(\gamma_1)| \le N a_1/2, \tag{3.26}$$

where both events hold eventually almost surely as $N \to \infty$. The claim follows with $\gamma_2 = a_1(c)/2$ and $\gamma_3 = a_2(c, t_{\min})$. □

We are now in position to prove Theorem 1.8.

Proof of Theorem 1.8. We apply Lemma 3.1 to $x = x^t$, the AMP estimate and $r = \widehat{x} - x^t$ the distance from the LASSO optimum. The thesis follows by checking conditions 1–5. Namely we need to show that there exists constants $c_1, \dots, c_5 > 0$ and, for each $\varepsilon > 0$ some $t = t(\varepsilon)$ exists such that 1–5 hold eventually almost surely as $N \to \infty$.

Condition 1 holds by Lemma 3.2.

Condition 2 is immediate since $x + r = \widehat{x}$ minimizes $\mathcal{C}(\,\cdot\,)$.

Condition 3 follows from Lemma 3.3 with ε arbitrarily small for t large enough.

Condition 4. Notice that this condition only needs to be verified for $\delta < 1$.

Take $v = v^t$ as defined in Eq. (3.16). Using the definition (1.5), it is easy to check that $|v_i^t| \le 1$ if $x_i^t = 0$ and $v_i^t = \mathrm{sign}(x_i^t)$ otherwise. In other words $v^t \in \partial\|x\|_1$ as required. Further by inspection of the proof of Lemma 3.3, it follows that $v^t = (1/\lambda)[A^*(y - Ax^t) + \mathrm{sg}(\mathcal{C}, x^t)]$, with $\mathrm{sg}(\mathcal{C}, x^t)$ the subgradient bounded in that lemma (cf. Eq. (5.3)). The condition then holds by Proposition 3.6.

Condition 5 follows from standard limit theorems on the singular values of Wishart matrices (cf. Theorem F.2). □

4 State evolution estimates

This section contains a reminder of the state-evolution method developed in [BM11]. For greater convenience of the reader, we also restate two lemmas from [BM11] (namely, Lemmas F.3 and F.3) in appendix F.3. We will use these two Lemmas throughout our analysis.

We also state some extensions of those results that will be proved in the appendices.

4.1 State evolution

AMP, cf. Eq. (1.5) is a special case of the general iterative procedure given by Eq. (3.1) of [BM11]. This takes the general form

$$\begin{aligned} h^{t+1} &= A^* m^t - \xi_t\, q^t\,, \qquad & m^t = g_t(b^t, w)\,,\\ b^t &= A\, q^t - \lambda_t m^{t-1}\,, \qquad & q^t = f_t(h^t, x_0)\,, \end{aligned} \tag{4.1}$$

where $\xi_t = \langle g'(b^t, w)\rangle$, $\lambda_t = \frac{1}{\delta}\langle f_t'(h^t, x_0)\rangle$ (both derivatives are with respect to the first argument).

This reduction can be seen by defining

$$h^{t+1} = x_0 - (A^* z^t + x^t)\,, \tag{4.2}$$

$$q^t = x^t - x_0\,, \tag{4.3}$$

$$b^t = w - z^t\,, \tag{4.4}$$

$$m^t = -z^t\,, \tag{4.5}$$

where

$$f_t(s, x_0) = \eta_{t-1}(x_0 - s) - x_0\,, \qquad g_t(s, w) = s - w\,, \tag{4.6}$$

and the initial condition is $q^0 = -x_0$.

Regarding h^t, b^t as column vectors, the equations for $b^0, \dots, b^{t-1}$ and $h^1, \dots, h^t$ can be written in matrix form as:

$$\underbrace{\left[h^1 + \xi_0 q^0 | h^2 + \xi_1 q^1 | \cdots | h^t + \xi_{t-1} q^{t-1}\right]}_{X_t} = A^* \underbrace{\left[m^0 | \dots | m^{t-1}\right]}_{M_t}, \tag{4.7}$$

$$\underbrace{\left[b^0 | b^1 + \lambda_1 m^0 | \cdots | b^{t-1} + \lambda_{t-1} m^{t-2}\right]}_{Y_t} = A \underbrace{\left[q^0 | \dots | q^{t-1}\right]}_{Q_t}. \tag{4.8}$$

or in short $Y_t = AQ_t$ and $X_t = A^* M_t$.

Following [BM11], we define $\mathfrak{S}_t$ as the σ-algebra generated by $b^0, \dots, b^{t-1}$, $m^0, \dots, m^{t-1}$, $h^1, \dots, h^t$, and $q^0, \dots, q^t$. The conditional distribution of the random matrix A given the σ-algebra $\mathfrak{S}_t$, is given by

$$A|_{\mathfrak{S}_t} \stackrel{\text{d}}{=} E_t + \mathcal{P}_t(\tilde{A}). \tag{4.9}$$

Here $\tilde{A} \stackrel{\text{d}}{=} A$ is a random matrix independent of $\mathfrak{S}_t$, and $E_t = \mathbb{E}(A|\mathfrak{S}_t)$ is given by

$$E_t = Y_t(Q_t^* Q_t)^{-1} Q_t^* + M_t(M_t^* M_t)^{-1} X_t^* - M_t(M_t^* M_t)^{-1} M_t^* Y_t (Q_t^* Q_t)^{-1} Q_t^*. \tag{4.10}$$

Further, $\mathcal{P}_t$ is the orthogonal projector onto subspace $\mathrm{V}_t = \{A | AQ_t = 0, A^* M_t = 0\}$, defined by

$$\mathcal{P}_t(\tilde{A}) = P_{M_t}^{\perp} \tilde{A} P_{Q_t}^{\perp}.$$

Here $P_{M_t}^{\perp} = I - P_{M_t}$, $P_{Q_t}^{\perp} = I - P_{Q_t}$, and P_{Q_t}, P_{M_t} are orthogonal projector onto column spaces of Q_t and M_t respectively.

Before proceeding, it is convenient to introduce the notation

$$\omega_t \equiv \frac{1}{\delta} \langle \eta'(A^* z^{t-1} + x^{t-1}; \theta_{t-1}) \rangle$$

to denote the coefficient of z^{t-1} in Eq. (1.5). Using $h^t = x_0 - A^* z^{t-1} - x^{t-1}$ and Lemma F.3(b) (proved in [BM11]) we get, almost surely,

$$\lim_{N \to \infty} \omega_t = \omega_t^{\infty} \equiv \frac{1}{\delta} \mathbb{E}\big[\eta'(X_0 + \tau_{t-1} Z; \theta_{t-1})\big]. \tag{4.11}$$

Notice that the function $\eta'(\,\cdot\,; \theta_{t-1})$ is discontinuous and therefore Lemma F.3(b) does not apply immediately. On the other hand, this implies that the empirical distribution of $\{(A^* z_i^{t-1} + x_i^{t-1}, x_{0,i})\}_{1 \le i \le N}$ converges weakly to the distribution of $(X_0 + \tau_{t-1} Z, X_0)$. The claim follows from the fact that $X_0 + \tau_{t-1} Z$ has a density, together with the standard properties of weak convergence.

4.2 Some consequences and generalizations

We begin with a simple calculation, that will be useful.

Lemma 4.1. *If $\{z^t\}_{t \ge 0}$ are the* AMP *residuals, then, almost surely,*

$$\lim_{n \to \infty} \frac{1}{n} \|z^t\|^2 = \tau_t^2. \tag{4.12}$$

Proof. Using representation (4.5) and Lemma F.3(b)(c), we get

$$\lim_{n\to\infty} \frac{1}{n} \|z^t\|^2 \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} \|m^t\|^2 \overset{\text{a.s.}}{=} \lim_{N\to\infty} \frac{1}{N} \|h^{t+1}\|^2 = \tau_t^2 \,.$$

□

Next, we need to generalize state evolution to compute large system limits for functions of x^t, x^s, with $t \neq s$. To this purpose, we define the covariances $\{\mathsf{R}_{s,t}\}_{s,t\geq 0}$ recursively by

$$\mathsf{R}_{s+1,t+1} = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big\{[\eta(X_0+Z_s;\theta_s) - X_0]\,[\eta(X_0+Z_t;\theta_t) - X_0]\Big\}\,, \tag{4.13}$$

with (Z_s, Z_t) jointly gaussian, independent from $X_0 \sim p_{X_0}$ with zero mean and covariance given by $\mathbb{E}\{Z_s^2\} = \mathsf{R}_{s,s}$, $\mathbb{E}\{Z_t^2\} = \mathsf{R}_{t,t}$, $\mathbb{E}\{Z_s Z_t\} = \mathsf{R}_{s,t}$. The boundary condition is fixed by letting $\mathsf{R}_{0,0} = \sigma^2 + \mathbb{E}\{X_0^2\}/\delta$ and

$$\mathsf{R}_{0,t+1} = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big\{[\eta(X_0+Z_t;\theta_t) - X_0]\,(-X_0)\Big\}\,, \tag{4.14}$$

with $Z_t \sim \mathsf{N}(0, \mathsf{R}_{t,t})$ independent of X_0. This determines by the above recursion $\mathsf{R}_{t,s}$ for all $t \geq 0$ and for all $s \geq 0$.

With these definition, we have the following generalization of Theorem 1.1.

Theorem 4.2. *Let $\{x_0(N), w(N), A(N)\}_{N\in\mathbb{N}}$ be a converging sequence of instances with the entries of $A(N)$ iid normal with mean 0 and variance $1/n$ and let $\psi : \mathbb{R}^3 \to \mathbb{R}$ be a pseudo-Lipschitz function. Then, for all $s \geq 0$ and $t \geq 0$ almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\big(x_i^s + (A^* z^s)_i, x_i^t + (A^* z^t)_i, x_{0,i}\big) = \mathbb{E}\Big\{\psi\big(X_0+Z_s, X_0+Z_t, X_0\big)\Big\}\,, \tag{4.15}$$

where (Z_s, Z_t) jointly gaussian, independent from $X_0 \sim p_{X_0}$ with zero mean and covariance given by $\mathbb{E}\{Z_s^2\} = \mathsf{R}_{s,s}$, $\mathbb{E}\{Z_t^2\} = \mathsf{R}_{t,t}$, $\mathbb{E}\{Z_s Z_t\} = \mathsf{R}_{s,t}$.

Notice that the above implies in particular, for any pseudo-Lipschitz function $\psi : \mathbb{R}^3 \to \mathbb{R}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\big(x_i^{s+1}, x_i^{t+1}, x_{0,i}\big) = \mathbb{E}\Big\{\psi\big(\eta(X_0+Z_s;\theta_s), \eta(X_0+Z_t;\theta_t), X_0\big)\Big\}\,. \tag{4.16}$$

Clearly this result reduces to Theorem 1.1 in the case $s = t$ by noting that $\mathsf{R}_{t,t} = \tau_t^2$. The general proof can be found in Appendix B.

The following lemma implies that, asymptotically for large N, the AMP estimates converge.

Lemma 4.3. *Under the condition of Theorem 1.5, the estimates $\{x^t\}_{t\geq 0}$ and residuals $\{z^t\}_{t\geq 0}$ of* AMP *almost surely satisfy*

$$\lim_{t\to\infty} \lim_{N\to\infty} \frac{1}{N}\|x^t - x^{t-1}\|^2 = 0\,, \qquad \lim_{t\to\infty} \lim_{N\to\infty} \frac{1}{N}\|z^t - z^{t-1}\|^2 = 0\,. \tag{4.17}$$

The proof is deferred to Appendix C.

5 Proofs of auxiliary lemmas

5.1 Proof of Lemma 3.2

In order to bound the norm of x^t, we use state evolution, Theorem 1.1, for the function $\psi(a,b) = a^2$,

$$\lim_{t\to\infty}\lim_{N\to\infty}\langle x^t, x^t\rangle \overset{\text{a.s.}}{=} \mathbb{E}\left\{\eta(X_0 + \tau_* Z;\theta_*)^2\right\}$$

for $Z \sim \mathsf{N}(0,1)$ and independent of $X_0 \sim p_{X_0}$. The expectation on the right hand side is bounded and hence $\lim_{t\to\infty}\lim_{N\to\infty}\langle x^t, x^t\rangle$ is bounded.

For $\widehat{x}$, first note that

$$\begin{aligned}
\frac{1}{N}\mathcal{C}(\widehat{x}) \le \frac{1}{N}\mathcal{C}(0) &= \frac{1}{2N}\|y\|^2 \\
&= \frac{1}{2N}\|Ax_0 + w\|^2 \\
&\le \frac{\|w\|^2 + \sigma_{\max}(A)^2\|x_0\|^2}{N} \le \mathsf{B}_1. \qquad (5.1)
\end{aligned}$$

The last bound holds almost surely as $N \to \infty$, using standard asymptotic estimate on the singular values of random matrices (cf. Theorem F.2) implying that $\sigma_{\max}(A)$ has a bounded limit almost surely, together with the fact that (x_0, w, A) is a converging sequence.

Now, decompose $\widehat{x}$ as $\widehat{x} = \widehat{x}_{\|} + \widehat{x}_{\perp}$ where $\widehat{x}_{\|} \in \ker(A)$ and $\widehat{x}_{\perp} \in \ker(A)^{\perp}$ (the orthogonal complement of $\ker(A)$). Since, $\widehat{x}_{\|}$ belongs to the random subspace $\ker(A)$ with dimension $N - n = N(1-\delta)$, Kashin theorem (cf. Theorem F.1) implies that there exists a positive constant $c_1 = c_1(\delta)$ such that

$$\begin{aligned}
\frac{1}{N}\|\widehat{x}\|^2 &= \frac{1}{N}\|\widehat{x}_{\|}\|^2 + \frac{1}{N}\|\widehat{x}_{\perp}\|^2 \\
&\le c_1\left(\frac{\|\widehat{x}_{\|}\|_1}{N}\right)^2 + \frac{1}{N}\|\widehat{x}_{\perp}\|^2 .
\end{aligned}$$

Hence, by using triangle inequality and Cauchy-Schwarz, we get

$$\begin{aligned}
\frac{1}{N}\|\widehat{x}\|^2 &\le 2c_1\left(\frac{\|\widehat{x}\|_1}{N}\right)^2 + 2c_1\left(\frac{\|\widehat{x}_{\perp}\|_1}{N}\right)^2 + \frac{1}{N}\|\widehat{x}_{\perp}\|^2 \\
&\le 2c_1\left(\frac{\|\widehat{x}\|_1}{N}\right)^2 + \frac{2c_1+1}{N}\|\widehat{x}_{\perp}\|^2 .
\end{aligned}$$

By definition of cost function we have $\|\widehat{x}\|_1 \le \lambda^{-1}\mathcal{C}(\widehat{x})$. Further, limit theorems for the eigenvalues of Wishart matrices (cf. Theorem F.2) imply that there exists a constant $c = c(\delta)$ such that asymptotically almost surely $\|\widehat{x}_{\perp}\|^2 \le c\,\|A\widehat{x}_{\perp}\|^2$. Therefore (denoting by $c_i:\ i = 2,3,4$ bounded constants), we have

$$\begin{aligned}
\frac{1}{N}\|\widehat{x}\|^2 &\le 2c_1\left(\frac{\|\widehat{x}\|_1}{N}\right)^2 + \frac{c_2}{N}\|A\widehat{x}_{\perp}\|^2 \\
&\le 2c_1\left(\frac{\|\widehat{x}\|_1}{N}\right)^2 + \frac{2c_2}{N}\|y - A\widehat{x}_{\perp}\|^2 + \frac{2c_2}{N}\|y\|^2 \\
&\le c_3\left(\frac{\mathcal{C}(\widehat{x})}{N}\right)^2 + 2c_2\frac{\mathcal{C}(\widehat{x})}{N} + \frac{2c_2}{N}\|Ax_0 + w\|^2 .
\end{aligned}$$

The claim follows by using the Eq. (5.1) to bound $\mathcal{C}(\widehat{x})/N$ and using $\|Ax_0+w\|^2 \le \sigma_{\max}(A)^2\|x_0\|^2 + \|w\|^2 \le 2N\mathsf{B}_1$ to bound the last term. □

5.2 Proof of Lemma 3.3

First note that equation $x^t = \eta(A^*z^{t-1} + x^{t-1};\theta_{t-1})$ of AMP implies

$$
\begin{aligned}
x_i^t + \theta_{t-1}\,\mathrm{sign}(x_i^t) &= [A^*z^{t-1}]_i + x_i^{t-1}, \qquad &&\text{if } x_i^t \neq 0\,,\\
\Big|[A^*z^{t-1}]_i + x_i^{t-1}\Big| &\le \theta_{t-1}, \qquad &&\text{if } x_i^t = 0\,.
\end{aligned}
\tag{5.2}
$$

Therefore, the vector $\mathrm{sg}(\mathcal{C},x^t) \equiv \lambda\, s^t - A^*(y - Ax^t)$ where

$$
s_i^t = \begin{cases} \mathrm{sign}(x_i^t) & \text{if } x_i^t \neq 0\,,\\[2ex] \frac{1}{\theta_{t-1}}\Big\{[A^*z^{t-1}]_i + x_i^{t-1}\Big\} & \text{otherwise,}\end{cases}
\tag{5.3}
$$

is a valid subgradient of $\mathcal{C}$ at x^t. On the other hand, $y - Ax^t = z^t - \omega_t z^{t-1}$. We finally get

$$
\begin{aligned}
\mathrm{sg}(\mathcal{C},x^t) &= \frac{1}{\theta_{t-1}}\left[\lambda\theta_{t-1}s^t - \theta_{t-1}A^*(z^t - \omega_t z^{t-1})\right]\\
&= \frac{1}{\theta_{t-1}}\left[\lambda\theta_{t-1}s^t - \theta_{t-1}(1-\omega_t)A^*z^{t-1}\right] - A^*(z^t - z^{t-1})\\
&= \underbrace{\frac{1}{\theta_{t-1}}\left[\lambda\theta_{t-1}s^t - \lambda A^*z^{t-1}\right]}_{(I)} - A^*(z^t - z^{t-1}) + \frac{[\lambda - \theta_{t-1}(1-\omega_t)]}{\theta_{t-1}}\,A^*z^{t-1}\,.
\end{aligned}
$$

It is straightforward to see from Eqs. (5.2) and (5.3) that $(I) = \lambda(x^{t-1} - x^t)$. Hence,

$$
\frac{1}{\sqrt{N}}\|\mathrm{sg}(\mathcal{C},x^t)\| \le \frac{\lambda}{\theta_{t-1}\sqrt{N}}\|x^t - x^{t-1}\| + \frac{\sigma_{\max}(A)}{\sqrt{N}}\|z^t - z^{t-1}\| + \frac{|\lambda - \theta_{t-1}(1-\omega_t)|}{\theta_{t-1}}\,\frac{1}{\sqrt{N}}\|z^{t-1}\|\,.
$$

By Lemma 4.3, and the fact that $\sigma_{\max}(A)$ is almost surely bounded as $N\to\infty$ (cf. Theorem F.2), we deduce that the two terms $\lambda\|x^t - x^{t-1}\|/(\theta_{t-1}\sqrt{N})$ and $\sigma_{\max}(A)\|z^t - z^{t-1}\|^2/\sqrt{N}$ converge to 0 when $N\to\infty$ and then $t\to\infty$. For the third term, using state evolution (see Lemma 4.1), we obtain $\lim_{N\to\infty}\|z^{t-1}\|^2/N < \infty$. Finally, using the calibration relation Eq. (1.15), we get

$$
\lim_{t\to\infty}\lim_{N\to\infty}\left|\frac{\lambda - \theta_{t-1}(1-\omega_t)}{\theta_{t-1}}\right| \overset{\text{a.s.}}{=} \frac{1}{\theta_*}\left|\lambda - \theta_*(1 - \frac{1}{\delta}\mathbb{E}\left\{\eta'(X_0 + \tau_* Z;\theta_*)\right\})\right| = 0\,,
$$

which finishes the proof. □

5.3 Proof of Lemma 3.4

The proof uses the representation (4.9), together with the expression (4.10) for the conditional expectation. Apart from the matrices Y_t, Q_t, X_t, M_t introduced there, we will also use

$$
B_t \equiv \Big[b^0\Big|b^1\Big|\cdots\Big|b^{t-1}\Big]\,,\qquad H_t \equiv \Big[h^1\Big|h^2\Big|\cdots\Big|h^t\Big]\,.
$$

In this section, since t is fixed, we will drop everywhere the subscript t from such matrices.

We state below a somewhat more convenient description.

Lemma 5.1. *For any $v \in \mathbb{R}^N$, we have*

$$Av|_{\mathfrak{S}} \stackrel{d}{=} Y(Q^*Q)^{-1}Q^*P_Q v + M(M^*M)^{-1}X^*P_Q^{\perp} v + P_M^{\perp}\tilde{A}P_Q^{\perp} v\,. \tag{5.4}$$

Proof. It is clearly sufficient to prove that, for $v = v_{\parallel} + v_{\perp}$, $P_Q v_{\parallel} = v_{\parallel}$, $P_Q^{\perp} v_{\perp} = v_{\perp}$, we have

$$Ev_{\parallel} = Y(Q^*Q)^{-1}Q^*v_{\parallel}\,, \qquad Ev_{\perp} = M(M^*M)^{-1}X^*v_{\perp}\,. \tag{5.5}$$

The first identity is an easy consequence of the fact that $X^*Q = M^*AQ = M^*Y$, while the second one follows immediately from $Q^*v_{\perp} = 0$. □

The following fact (see Appendix D for a proof) will be used several times.

Lemma 5.2. *For any t there exists $c > 0$ such that, for $R \in \{Q^*Q;\, M^*M;\, X^*X;\, Y^*Y\}$, eventually almost surely as $N \to \infty$,*

$$c \le \lambda_{\min}(R/N) \le \lambda_{\max}(R/N) \le 1/c\,. \tag{5.6}$$

Given the above remarks, we will immediately see that Lemma 3.4 is implied by the following statement.

Lemma 5.3. *Let $S \subseteq [N]$ be given such that $|S| \le N(\delta - \gamma)$, for some $\gamma > 0$. Then there exists $\alpha_1 = \alpha_1(\gamma) > 0$ (independent of t) and $\alpha_2 = \alpha_2(\gamma, t) > 0$ (depending on t and γ) such that*

$$\mathbb{P}\Big\{ \min_{\|v\|=1,\, \mathrm{supp}(v)\subseteq S} \big\|Ev + P_M^{\perp}\tilde{A}P_Q^{\perp}v\big\| \le \alpha_2 \Big| \mathfrak{S}_t \Big\} \le e^{-N\alpha_1}\,,$$

eventually almost surely as $N \to \infty$. (With $Ev = Y(Q^*Q)^{-1}Q^*P_Q v + M(M^*M)^{-1}X^*P_Q^{\perp} v$.)

In the next section we will show that this lemma implies Lemma 3.4. We will then prove the lemma just stated.

5.3.1 Lemma 5.3 implies Lemma 3.4

By Borel-Cantelli, it is sufficient to show that, for S measurable on $\mathfrak{S}_t$ and $|S| \le N(\delta - c)$ there exist $a_1 = a_1(c) > 0$ and $a_2 = a_2(c, t) > 0$, such that

$$\mathbb{P}\Big\{ \min_{|S'| \le a_1 N}\ \min_{\|v\|=1, \mathrm{supp}(v)\subseteq S\cup S'} \|Av\| < a_2 \Big\} \le \frac{1}{N^2}\,,$$

for all N large enough. Conditioning on $\mathfrak{S}_t$ and using the union bound, this probability can be estimated as

$$\mathbb{E}\Big\{\mathbb{P}\Big\{ \min_{|S'| \le a_1 N}\ \min_{\|v\|=1, \mathrm{supp}(v)\subseteq S\cup S'} \|Av\| < a_2 \Big| \mathfrak{S}_t \Big\}\Big\} \le$$
$$\le e^{Nh(a_1)}\mathbb{E}\Big\{ \max_{|S'| \le a_1 N} \mathbb{P}\Big\{ \min_{\|v\|=1, \mathrm{supp}(v)\subseteq S\cup S'} \|Av\| < a_2 \Big| \mathfrak{S}_t \Big\}\Big\}\,,$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. The union bound calculation indeed proceeds as follows

$$\begin{aligned}\mathbb{P}\{\min_{|S'|\le Na_1} \mathsf{X}_{S'} < a_2 \big| \mathfrak{S}_t\} &\le \sum_{|S'|\le Na_1} \mathbb{P}\{\mathsf{X}_{S'} < a_2 \big| \mathfrak{S}_t\} \\ &\le \Big[\sum_{k=1}^{Na_1} \binom{N}{k}\Big] \max_{|S'|\le Na_1} \mathbb{P}\{\mathsf{X}_{S'} < a_2 \big| \mathfrak{S}_t\} \\ &\le e^{Nh(a_1)} \max_{|S'|\le Na_1} \mathbb{P}\{\mathsf{X}_{S'} < a_2 \big| \mathfrak{S}_t\}\,,\end{aligned}$$

where $\mathsf{X}_{S'} = \min_{\|v\|=1, \mathrm{supp}(v)\subseteq S\cup S'} \|Av\|$. Now, fix $a_1 < c/2$ in such a way that $h(a_1) \le \alpha_1(c/2)/2$ (with α_1 defined as per Lemma 5.3). Further choose $a_2 = \alpha_2(c/2,t)/2$. The above probability is then upper bounded by

$$e^{N\alpha_1(c/2)/2}\, \mathbb{E}\Big\{ \max_{|S''|\le N(\delta - c/2)} \mathbb{P}\Big\{ \min_{\|v\|=1, \mathrm{supp}(v)\subseteq S''} \|Av\| < \frac{1}{2}\alpha_2(c/2,t) \Big| \,\mathfrak{S}_t\Big\}\Big\}\,.$$

Finally, applying Lemma 5.3 and using Lemma 5.1 to estimate Av, we get, for all N large enough,

$$e^{N\alpha_1/2}\, \mathbb{E}\big\{ \max_{|S''|\le N(\delta - c/2)} e^{-N\alpha_1}\big\} \le \frac{1}{N^2}\,.$$

This finishes the proof. □

5.3.2 Proof of Lemma 5.3

We begin with the following Pythagorean inequality.

Lemma 5.4. *Let $S \subseteq [N]$ be given such that $|S| \le N(\delta - \gamma)$, for some $\gamma > 0$. Recall that $Ev = Y(Q^*Q)^{-1}Q^*P_Q v + M(M^*M)^{-1}X^*P_Q^{\perp} v$ and consider the event*

$$\mathcal{E}_1 \equiv \Big\{ \big\|Ev + P_M^{\perp}\tilde{A}P_Q^{\perp}v\big\|^2 \ge \frac{\gamma}{4\delta}\big\|Ev - P_M\tilde{A}P_Q^{\perp}v\big\|^2 + \frac{\gamma}{4\delta}\big\|\tilde{A}P_Q^{\perp}v\big\|^2 \;\; \forall v \;\; s.t. \;\; \|v\| = 1 \;\; and \;\; \mathrm{supp}(v) \subseteq S\Big\}.$$

Then there exists $a = a(\gamma) > 0$ such that $\mathbb{P}\{\mathcal{E}_1|\mathfrak{S}_t\} \ge 1 - e^{-Na}$.

Proof. We claim that the following inequality holds for all $v \in \mathbb{R}^N$, that satisfy $\|v\| = 1$ and $\mathrm{supp}(v) \subseteq S$, with the probability claimed in the statement

$$|(Ev - P_M\tilde{A}P_Q^{\perp}v\,,\, \tilde{A}P_Q^{\perp}v)| \;\; \le \;\; \sqrt{1 - \frac{\gamma}{2\delta}}\, \|Ev - P_M\tilde{A}P_Q^{\perp}v\|\, \|\tilde{A}P_Q^{\perp}v\|\,. \tag{5.7}$$

Here the notation (u,v) refers to the usual scalar product u^*v of vectors u and v of the same dimension. Assuming that the claim holds, we have indeed

$$\begin{aligned}\big\|Ev + P_M^{\perp}\tilde{A}P_Q^{\perp}v\big\|^2 &\ge \big\|Ev - P_M\tilde{A}P_Q^{\perp}v\big\|^2 + \big\|\tilde{A}P_Q^{\perp}v\big\|^2 - 2|(Ev - P_M\tilde{A}P_Q^{\perp}v\,,\, \tilde{A}P_Q^{\perp}v)| \\ &\ge \big\|Ev\big\|^2 + \big\|P_M^{\perp}\tilde{A}P_Q^{\perp}v\big\|^2 - 2\sqrt{1 - \frac{\gamma}{2\delta}}\, \|Ev - P_M\tilde{A}P_Q^{\perp}v\|\, \|\tilde{A}P_Q^{\perp}v\| \\ &\ge \Big(1 - \sqrt{1 - \frac{\gamma}{2\delta}}\Big)\Big\{\big\|Ev - P_M\tilde{A}P_Q^{\perp}v\big\|^2 + \big\|\tilde{A}P_Q^{\perp}v\big\|^2\Big\}\,,\end{aligned}$$

which implies the thesis.

In order to prove the claim (5.7), we notice that for any v, the unit vector $\tilde{A}P_Q^\perp v/\|\tilde{A}P_Q^\perp v\|$ belongs to the random linear space $\mathrm{im}(\tilde{A}P_Q^\perp P_S)$. Here P_S is the orthogonal projector onto the subspace of vectors supported on S. Further $\mathrm{im}(\tilde{A}P_Q^\perp P_S)$ is a uniformly random subspace of dimension at most $N(\delta-\gamma)$. Also, the normalized vector $(Ev - P_M\tilde{A}P_Q^\perp v)/\|Ev - P_M\tilde{A}P_Q^\perp v\|$ belongs to the linear space of dimension at most $2t$ spanned the columns of M and of B. The claim follows then from a standard concentration-of-measure argument. In particular applying Proposition E.1 for

$$m = n, \quad m\lambda = N(\delta-\gamma), \quad d = 2t \;\; \textit{and} \;\; \varepsilon = \sqrt{1-\frac{\gamma}{2\delta}} - \sqrt{1-\frac{\gamma}{\delta}}$$

yields

$$\left(\frac{Ev - P_M\tilde{A}P_Q^\perp v}{\|Ev - P_M\tilde{A}P_Q^\perp v\|}\,,\; \frac{\tilde{A}P_Q^\perp v}{\|\tilde{A}P_Q^\perp v\|}\right) \le \sqrt{\lambda} + \varepsilon = \sqrt{1-\frac{\gamma}{2\delta}}\,.$$

(Notice that in Proposition E.1 is stated for the equivalent case of a random sub-space of fixed dimension d, and a subspace of dimension scaling linearly with the ambient one.) □

Next we estimate the term $\|\tilde{A}P_Q^\perp v\|^2$ in the above lower bound.

Lemma 5.5. *Let $S \subseteq [N]$ be given such that $|S| \le N(\delta-\gamma)$, for some $\gamma > 0$. Then there exists constant $c_1 = c_1(\gamma)$, $c_2 = c_2(\gamma)$ such that the event*

$$\mathcal{E}_2 \equiv \Big\{ \big\|\tilde{A}P_Q^\perp v\big\| \ge c_1(\gamma)\|P_Q^\perp v\| \;\; \forall v \;\; \textit{such that} \;\; \mathrm{supp}(v) \subseteq S \Big\}\,,$$

holds with probability $\mathbb{P}\{\mathcal{E}_2|\mathfrak{S}_t\} \ge 1 - e^{-Nc_2}$.

Proof. Let V be the linear space $V = \mathrm{im}(P_Q^\perp P_S)$. Of course the dimension of V is at most $N(\delta-\gamma)$. Then we have (for all vectors with $\mathrm{supp}(v) \subseteq S$)

$$\big\|\tilde{A}P_Q^\perp v\big\| \ge \sigma_{\min}(\tilde{A}|_V)\,\|P_Q^\perp v\|\,, \tag{5.8}$$

where $\tilde{A}|_V$ is the restriction of $\tilde{A}$ to the subspace V. By invariance of the distribution of $\tilde{A}$ under rotation, $\sigma_{\min}(\tilde{A}|_V)$ is distributed as the minimum singular value of a gaussian matrix of dimensions $N\delta \times \dim(V)$. The latter is almost surely bounded away from 0 as $N \to \infty$, since $\dim(V) \le N(\delta-\gamma)$ (see for instance Theorem F.2). Large deviation estimates [LPRTJ05] imply that the probability that the minimum singular value is smaller than a constant $c_1(\gamma)$ is exponentially small. □

Finally a simple bound to control the norm of Ev.

Lemma 5.6. *There exists a constant $c = c(t) > 0$ such that, defining the event,*

$$\mathcal{E}_3 \equiv \big\{\|EP_Qv\| \ge c(t)\|P_Qv\|\,, \|EP_Q^\perp v\| \le c(t)^{-1}\|P_Q^\perp v\|, \;\; \textit{for all} \;\; v \in \mathbb{R}^N\big\}\,, \tag{5.9}$$

we have that $\mathcal{E}_3$ holds eventually almost surely as $N \to \infty$.

Proof. Without loss of generality take $v = Qa$ for $a \in \mathbb{R}^t$. By Lemma 5.1 we have $\|EP_Qv\|^2 = \|Ya\|^2 \ge \lambda_{\min}(Y^*Y)\|a\|^2$. Analogously $\|P_Qv\|^2 = \|Qa\|^2 \le \lambda_{\max}(Q^*Q)\|a\|^2$. The bound $\|EP_Qv\| \ge c(t)\|P_Qv\|$ follows then from Lemma 5.2.

The bound $\|EP_Q^\perp v\| \le c(t)^{-1}\|P_Q^\perp v\|$ is proved analogously. □

We can now prove Lemma 5.3 as promised.

Proof of Lemma 5.3. By Lemma 5.6 we can assume that event $\mathcal{E}_3$ holds, for some function $c = c(t)$ (without loss of generality $c < 1/2$). We will let $\mathcal{E}$ be the event

$$\mathcal{E} \equiv \Big\{ \min_{\|v\|=1,\,\mathrm{supp}(v)\subseteq S} \big\|Ev + P_M^\perp \tilde{A} P_Q^\perp v\big\| \le \alpha_2(t)\Big\}\,. \tag{5.10}$$

for $\alpha_2(t) > 0$ small enough.

Let us assume first that $\|P_Q^\perp v\| \le c^2/10$, whence

$$\begin{aligned}\|Ev - P_M \tilde{A} P_Q^\perp\| &\ge \|EP_Q v\| - \|EP_Q^\perp v\| - \|P_M \tilde{A} P_Q^\perp v\| \\ &\ge c\|P_Q v\| - (c^{-1} + \|\tilde{A}\|_2)\|P_Q^\perp v\| \\ &\ge \frac{c}{2} - \frac{c}{10} - \|\tilde{A}\|_2 \frac{c^2}{10} = \frac{2c}{5} - \|\tilde{A}\|_2 \frac{c^2}{10}\,,\end{aligned}$$

where the last inequality uses $\|P_Q v\| = \sqrt{1 - \|P_Q^\perp v\|^2} \ge 1/2$. Therefore, using Lemma 5.4, we get

$$\mathbb{P}\{\mathcal{E}|\mathfrak{S}_t\} \le \mathbb{P}\Big\{\frac{2c}{5} - \|\tilde{A}\|_2 \frac{c^2}{10} \le \sqrt{\frac{4\delta}{\gamma}}\alpha_2(t)\Big|\mathfrak{S}_t\Big\} + e^{-Na}\,,$$

and the thesis follows from large deviation bounds on the norm $\|\tilde{A}\|_2$ [Led01] by first taking c small enough, and then choosing $\alpha_2(t) < \frac{c}{5}\sqrt{\frac{\gamma}{4\delta}}$.

Next we assume $\|P_Q^\perp v\| \ge c^2/10$. Due to Lemma 5.4 and 5.5 we can assume that events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Therefore

$$\big\|Ev + P_M^\perp \tilde{A} P_Q^\perp v\big\| \ge \Big(\frac{\gamma}{4\delta}\Big)^{1/2} \|\tilde{A} P_Q^\perp v\| \ge \Big(\frac{\gamma}{4\delta}\Big)^{1/2} c_1(\gamma)\|P_Q^\perp v\|\,,$$

which proves our thesis. ☐

5.4 Proof of Lemma 3.5

The key step consists in establishing the following result, which will be instrumental in the proof of Lemma 4.3 as well (and whose proof is deferred to Appendix C.1).

Lemma 5.7. *Assume $\alpha > \alpha_{\min}(\delta)$ and let $\{\mathsf{R}_{s,t}\}$ be defined by the recursion (4.13) with initial condition (4.14). Then there exists constants B_1, $\mathsf{r}_1 > 0$ such that for all $t \ge 0$*

$$\begin{aligned}\big|\mathsf{R}_{t,t} - \tau_*^2\big| &\le \mathsf{B}_1\, e^{-\mathsf{r}_1 t}\,, &(5.11)\\ \big|\mathsf{R}_{t,t+1} - \tau_*^2\big| &\le \mathsf{B}_1\, e^{-\mathsf{r}_1 t}\,. &(5.12)\end{aligned}$$

It is also useful to prove the following fact.

Lemma 5.8. *For any $\alpha > 0$ and $T \ge 0$, the $T \times T$ matrix $R_{T+1} \equiv \{\mathsf{R}_{s,t}\}_{0\le s,t<T}$ is strictly positive definite.*

Proof. In proof of Theorem 4.2 we show that

$$\mathsf{R}_{s,t} = \lim_{N\to\infty} \langle h^{s+1}, h^{t+1}\rangle = \lim_{N\to\infty}\langle m^s, m^t\rangle\,,$$

almost surely. Hence, $R_{T+1} \stackrel{\text{a.s.}}{=} \delta \lim_{N\to\infty}(M^*_{T+1}M_{T+1}/N)$. Thus the result follows from Lemma 5.2. □

It is then relatively easy to deduce the following.

Lemma 5.9. *Assume $\alpha > \alpha_{\min}(\delta)$ and let $\{\mathsf{R}_{s,t}\}$ be defined by the recursion (4.13) with initial condition (4.14). Then there exists constants B_2, $\mathsf{r}_2 > 0$ such that for all $t_1, t_2 \geq t \geq 0$*

$$\big|\mathsf{R}_{t_1,t_2} - \tau_*^2\big| \leq \mathsf{B}_2\, e^{-\mathsf{r}_2\, t}\,. \tag{5.13}$$

Proof. By triangular inequality and Eq. (5.11), we have

$$\big|\mathsf{R}_{t_1,t_2} - \tau_*^2\big| \leq \frac{1}{2}\big|\mathsf{R}_{t_1,t_1} - 2\mathsf{R}_{t_1,t_2} + \mathsf{R}_{t_2,t_2}\big| + \mathsf{B}_1\, e^{-\mathsf{r}_1\, t}\,. \tag{5.14}$$

By Lemma 5.8 there exist gaussian random variables $Z_0, Z_1, Z_2, \dots$ on the same probability space with $\mathbb{E}\{Z_t\} = 0$ and $\mathbb{E}\{Z_tZ_s\} = \mathsf{R}_{t,s}$ (in fact in proof of Theorem 4.2 we show that $\{Z_i\}_{T\geq i\geq 0}$ is the weak limit of the empirical distribution of $\{h^{i+1}\}_{T\geq i\geq 0}$). Then (assuming, without loss of generality, $t_2 > t_1$) we have

$$\begin{aligned}
\big|\mathsf{R}_{t_1,t_1} - 2\mathsf{R}_{t_1,t_2} + \mathsf{R}_{t_2,t_2}\big| &= \mathbb{E}\{(Z_{t_1} - Z_{t_2})^2\}\\
&= \sum_{i,j=t_1}^{t_2-1} \mathbb{E}\{(Z_{i+1} - Z_i)(Z_{j+1} - Z_j)\}\\
&\leq \Big[\sum_{i=t_1}^{t_2-1} \mathbb{E}\{(Z_{i+1} - Z_i)^2\}^{1/2}\Big]^2\\
&\leq 4\mathsf{B}_1\Big[\sum_{i=t_1}^{\infty} e^{-\mathsf{r}_1 i/2}\Big]^2\\
&\leq \frac{4\mathsf{B}_1}{(1 - e^{-\mathsf{r}_1/2})^2}\, e^{-\mathsf{r}_1 t_1}\,,
\end{aligned}$$

which, together with Eq. (5.14) proves our claim. □

We are now in position to prove Lemma 3.5.

Proof of Lemma 3.5. We will show that, under the assumptions of the Lemma, $\lim_{N\to\infty}|S_{t_2}(\gamma) \setminus S_{t_1}(\gamma)|/N \leq \xi$ almost surely, which implies our claim. Indeed, by Theorem 4.2 we have

$$\begin{aligned}
\lim_{N\to\infty}\frac{1}{N}\,|S_{t_2}(\gamma) \setminus S_{t_1}(\gamma)| &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{\big\{|v_i^{t_2}|\geq 1-\gamma,\ |v_i^{t_1}|<1-\gamma\big\}}\\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{\big\{|x^{t_2-1}+A^*z^{t_2-1}-x^{t_2}|\geq(1-\gamma)\theta_{t_2-1},\ |x^{t_1-1}+A^*z^{t_1-1}-x^{t_1}|<(1-\gamma)\theta_{t_2-1}\big\}}\\
&= \mathbb{P}\big\{|X_0 + Z_{t_2-1}| \geq (1-\gamma)\theta_{t_2-1},\ |X_0 + Z_{t_1-1}| < (1-\gamma)\theta_{t_1-1}\big\} \equiv P_{t_1,t_2}\,,
\end{aligned}$$

where (Z_{t_1}, Z_{t_2}) are jointly normal with $\mathbb{E}\{Z_{t_1}^2\} = \mathsf{R}_{t_1,t_1}$, $\mathbb{E}\{Z_{t_1}Z_{t_2}\} = \mathsf{R}_{t_1,t_2}$, $\mathbb{E}\{Z_{t_2}^2\} = \mathsf{R}_{t_2,t_2}$. (Notice that, although the function $\mathbb{I}\{\cdots\}$ is discontinuous, the random vector $(X_0 + Z_{t_1-1}, X_0 + Z_{t_2-1})$ admits a density and hence Theorem 4.2 applies by weak convergence of the empirical distribution of $\{(x_i^{t_1-1} + (A^*z^{t_1-1})_i\,,\ x_i^{t_2-1} + (A^*z^{t_2-1})_i)\}_{1\le i\le N}$.)

Let $a \equiv (1-\gamma)\alpha\tau_*$. By Proposition 1.3, for any $\varepsilon > 0$ and all t_* large enough we have $|(1-\gamma)\theta_{t_i-1} - a| \le \varepsilon$ for $i \in \{1, 2\}$. Then

$$\begin{aligned}
P_{t_1,t_2} &\le \mathbb{P}\big\{|X_0 + Z_{t_2-1}| \ge a - \varepsilon,\ |X_0 + Z_{t_1-1}| < a + \varepsilon\big\} \\
&\le \mathbb{P}\big\{|Z_{t_1-1} - Z_{t_2-1}| \ge 2\varepsilon\big\} + \mathbb{P}\{a - 3\varepsilon \le |X_0 + Z_{t_1-1}| \le a + \varepsilon\big\} \\
&\le \frac{1}{4\varepsilon^2}[\mathsf{R}_{t_1-1,t_1-1} - 2\mathsf{R}_{t_1-1,t_2-1} + \mathsf{R}_{t_2-1,t_2-1}] + \frac{4\varepsilon}{\sqrt{2\pi\mathsf{R}_{t_1-1,t_1-1}}} \\
&\le \frac{1}{\varepsilon^2}\mathsf{B}_2\, e^{-\mathsf{r}_2 t_*} + \frac{\varepsilon}{\tau_*}\,,
\end{aligned}$$

where the last inequality follows by Lemma 5.9. By taking $\varepsilon = e^{-\mathsf{r}_2\, t_*/3}$ we finally get (for some constant C) $P_{t_1,t_2} \le C\, e^{-\mathsf{r}_2 t_*}$, which implies our claim. □

Acknowledgement

It is a pleasure to thank David Donoho and Arian Maleki for many stimulating exchanges. We are also indebted with José Bento who collaborated in preparing Figures 4 to 7.

An earlier version of this paper stated some auxiliary lemmas in terms of convergence *in probability*. We rectified this to convergence *almost sure* as for the main theorems (with virtually no change in the proofs). We are grateful to Edgar Dobriban and Weijie Su for pointing out this inconsistency.

This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211.

A Properties of the state evolution recursion

A.1 Proof of Proposition 1.3

It is a straightforward calculus exercise to compute the partial derivatives

$$\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \theta) = \frac{1}{\delta}\,\mathbb{E}\Big\{\Phi\Big(\frac{X_0 - \theta}{\tau}\Big) + \Phi\Big(\frac{-X_0 - \theta}{\tau}\Big)\Big\} - \frac{1}{\delta}\,\mathbb{E}\Big\{\frac{X_0}{\tau}\,\phi\Big(\frac{X_0 - \theta}{\tau}\Big) - \frac{X_0}{\tau}\,\phi\Big(\frac{-X_0 - \theta}{\tau}\Big)\Big\}\,, \tag{A.1}$$

$$\frac{\partial \mathsf{F}}{\partial \theta}(\tau^2, \theta) = \frac{2\theta}{\delta}\,\mathbb{E}\Big\{\Phi\Big(\frac{X_0 - \theta}{\tau}\Big) + \Phi\Big(\frac{-X_0 - \theta}{\tau}\Big)\Big\} - \frac{2\tau}{\delta}\,\mathbb{E}\Big\{\phi\Big(\frac{X_0 - \theta}{\tau}\Big) + \phi\Big(\frac{-X_0 - \theta}{\tau}\Big)\Big\}\,. \tag{A.2}$$

From these formulae we obtain the total derivative

$$\begin{aligned}
\delta\,\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau^2, \alpha\tau) \;&=\; (1+\alpha^2)\,\mathbb{E}\Big\{\Phi\Big(\frac{X_0 - \alpha\tau}{\tau}\Big) + \Phi\Big(\frac{-X_0 - \alpha\tau}{\tau}\Big)\Big\} \\
&\quad -\mathbb{E}\Big\{\Big(\frac{X_0 + \alpha\tau}{\tau}\Big)\,\phi\Big(\frac{X_0 - \alpha\tau}{\tau}\Big) - \Big(\frac{X_0 - \alpha\tau}{\tau}\Big)\,\phi\Big(\frac{-X_0 - \alpha\tau}{\tau}\Big)\Big\}\,.
\end{aligned} \tag{A.3}$$

Differentiating once more

$$\delta\frac{\mathrm{d}^2\mathsf{F}}{\mathrm{d}(\tau^2)^2}(\tau^2,\alpha\tau) = -\frac{1}{2\tau^2}\mathbb{E}\Big\{\Big(\frac{X_0}{\tau}\Big)^3\Big[\phi\Big(\frac{X_0-\alpha\tau}{\tau}\Big)-\phi\Big(\frac{-X_0-\alpha\tau}{\tau}\Big)\Big]\Big\}\,.$$

Now we have

$$u^3[\phi(u-\alpha)-\phi(-u-\alpha)]\ge 0\,, \tag{A.4}$$

with the inequality being strict whenever $\alpha>0$, $u\neq 0$. It follows that $\tau^2\mapsto\mathsf{F}(\tau^2,\alpha\tau)$ is concave, and strictly concave provided $\alpha>0$ and X_0 is not identically 0.

From Eq. (A.3) we obtain

$$\lim_{\tau^2\to\infty}\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau^2,\alpha\tau) = \frac{2}{\delta}\big\{(1+\alpha^2)\Phi(-\alpha)-\alpha\,\phi(\alpha)\big\}\,, \tag{A.5}$$

which is strictly positive for all $\alpha\ge 0$. To see this, let $f(\alpha)\equiv(1+\alpha^2)\Phi(-\alpha)-\alpha\,\phi(\alpha)$, and notice that $f'(\alpha)=2\alpha\Phi(-\alpha)-2\phi(\alpha)<0$, and $f(\infty)=0$.

Since $\tau^2\mapsto\mathsf{F}(\tau^2,\alpha\tau)$ is concave, and strictly increasing for τ^2 large enough, it also follows that it is increasing everywhere.

Notice that $\alpha\mapsto f(\alpha)$ is strictly decreasing with $f(0)=1/2$. Hence, for $\alpha>\alpha_{\min}(\delta)$, we have $\mathsf{F}(\tau^2,\alpha\tau)>\tau^2$ for τ^2 small enough and $\mathsf{F}(\tau^2,\alpha\tau)<\tau^2$ for τ^2 large enough. Therefore the fixed point equation admits at least one solution. It follows from the concavity of $\tau^2\mapsto\mathsf{F}(\tau^2,\alpha\tau)$ that the solution is unique and that the sequence of iterates τ_t^2 converge to τ_*. □

A.2 Proof of Proposition 1.4

As a first step, we claim that $\alpha\mapsto\tau_*^2(\alpha)$ is continuously differentiable on $(0,\infty)$. Indeed this is defined as the unique solution of

$$\tau_*^2 = \mathsf{F}(\tau_*^2,\alpha\tau_*)\,. \tag{A.6}$$

Since $(\tau^2,\alpha)\mapsto\mathsf{F}(\tau_*^2,\alpha\tau_*)$ is continuously differentiable and $0\le\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau_*^2,\alpha\tau_*)<1$ (the second inequality being a consequence of concavity plus $\lim_{\tau^2\to\infty}\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau^2,\alpha\tau)<1$, both shown in the proof of Proposition 1.3), the claim follows from the implicit function theorem applied to the mapping $(\tau^2,\alpha)\mapsto[\tau^2-F(\tau^2,\alpha)]$.

Next notice that $\tau_*^2(\alpha)\to+\infty$ as $\alpha\downarrow\alpha_{\min}(\delta)$. Indeed, introducing the notation $\mathsf{F}'_\infty\equiv\lim_{\tau^2\to\infty}\frac{\mathrm{d}\mathsf{F}}{\mathrm{d}\tau^2}(\tau^2,\alpha\tau)$, we have, again by concavity,

$$\tau_*^2\ge\mathsf{F}(0,0)+\mathsf{F}'_\infty\tau_*^2\,,$$

i.e. $\tau_*^2\ge\mathsf{F}(0,0)/(1-\mathsf{F}'_\infty)$. Now $\mathsf{F}(0,0)\ge\sigma^2$, while $\mathsf{F}'_\infty\uparrow 1$ as $\alpha\downarrow\alpha_{\min}(\delta)$ (shown in the proof of Proposition 1.3), whence the claim follows.

Finally $\tau_*^2(\alpha)\to\sigma^2+\mathbb{E}\{X_0^2\}/\delta$ as $\alpha\to\infty$. Indeed for any fixed $\tau^2>0$ we have $\mathsf{F}(\tau^2,\alpha\tau)\to\sigma^2+\mathbb{E}\{X_0^2\}/\delta$ as $\alpha\to\infty$ whence the claim follows by uniqueness of τ_*.

Next consider the function $(\alpha,\tau^2)\mapsto g(\alpha,\tau^2)$ defined by

$$g(\alpha,\tau^2)\equiv\alpha\tau\Big\{1-\frac{1}{\delta}\,\mathbb{P}\{|X_0+\tau\,Z|\ge\alpha\tau\}\Big\}\,.$$

Notice that $\lambda(\alpha) = g(\alpha, {\tau_*}^2(\alpha))$. Since g is continuously differentiable, it follows that $\alpha \mapsto \lambda(\alpha)$ is continuously differentiable as well.

Next consider $\alpha \downarrow \alpha_{\min}$, and let $l(\alpha) \equiv 1 - \frac{1}{\delta}\,\mathbb{P}\{|X_0 + \tau_*\, Z| \ge \alpha\tau_*\}$. Since $\tau_* \to +\infty$ in this limit, we have

$$l_* \equiv \lim_{\alpha\to\alpha_{\min}+} l(\alpha) = 1 - \frac{1}{\delta}\,\mathbb{P}\{|Z| \ge \alpha_{\min}\} = 1 - \frac{2}{\delta}\,\Phi(-\alpha_{\min})\,.$$

Using the characterization of $\alpha_{\min}$ in Eq. (1.14) (and the well known inequality $\alpha\Phi(-\alpha) \le \phi(\alpha)$ valid for all $\alpha > 0$), it is immediate to show that $l_* < 0$. Therefore

$$\lim_{\alpha\to\alpha_{\min}+} \lambda(\alpha) = l_* \lim_{\alpha\to\alpha_{\min}+} \alpha\tau_*(\alpha) = -\infty\,.$$

Finally let us consider the limit $\alpha \to \infty$. Since $\tau_*(\alpha)$ remains bounded, we have $\lim_{\alpha\to\infty}\mathbb{P}\{|X_0 + \tau_*\, Z| \ge \alpha\tau_*\} = 0$ whence

$$\lim_{\alpha\to\infty} \lambda(\alpha) = \lim_{\alpha\to\infty} \alpha\tau_*(\alpha) = \infty\,.$$

□

A.3 Proof of Corollary 1.7

By Proposition 1.4, it is sufficient to prove that, for any $\lambda > 0$ there exists a unique $\alpha > \alpha_{\min}$ such that $\lambda(\alpha) = \lambda$. Assume by contradiction that there are two distinct such values α_1, α_2.

Notice that in this case, the function $\alpha(\lambda)$ is not defined uniquely and we can apply Theorem 1.5 to both choices $\alpha(\lambda) = \alpha_1$ and $\alpha(\lambda) = \alpha_2$. Using the test function $\psi(x, y) = (x - y)^2$ we deduce that

$$\lim_{N\to\infty} \frac{1}{N}\|\widehat{x} - x_0\|^2 = \mathbb{E}\big\{[\eta(X_0 + \tau_* Z\,;\,\alpha\tau_*) - X_0]^2\big\} = \delta(\tau_*^2 - \sigma^2)\,.$$

Since the left hand side does not depend on the choice of α, it follows that $\tau_*(\alpha_1) = \tau_*(\alpha_2)$.

Next apply Theorem 1.5 to the function $\psi(x, y) = |x|$. We get

$$\lim_{N\to\infty} \frac{1}{N}\|\widehat{x}\|_1 = \mathbb{E}\big\{|\eta(X_0 + \tau_* Z\,;\,\alpha\tau_*)|\big\}\,.$$

For fixed τ_*, $\theta \mapsto \mathbb{E}\big\{|\eta(X_0 + \tau_* Z\,;\,\theta)|\big\}$ is strictly decreasing in θ. It follows that $\alpha_1\tau_*(\alpha_1) = \alpha_2\tau_*(\alpha_2)$. Since we already proved that $\tau_*(\alpha_1) = \tau_*(\alpha_2)$, we conclude $\alpha_1 = \alpha_2$. □

B Proof of Theorem 4.2

First note that using representation (4.2) we have $x^t + A^* z^t = x_0 - h^{t+1}$. Furthermore, using Lemma F.3(b) we have almost surely

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi\big(x_{0,i} - h_i^{s+1}, x_{0,i} - h_i^{t+1}, x_{0,i}\big) &= \mathbb{E}\Big\{\psi\big(X_0 - \tilde{Z}_s, X_0 - \tilde{Z}_t, X_0\big)\Big\}\\ &= \mathbb{E}\Big\{\psi\big(X_0 + \tilde{Z}_s, X_0 + \tilde{Z}_t, X_0\big)\Big\}\end{aligned}$$

for gaussian variables $\tilde{Z}_s$, $\tilde{Z}_t$ that have zero mean and are independent of X_0. Define for all $s \geq 0$ and $t \geq 0$,

$$\tilde{\mathsf{R}}_{t,s} \equiv \lim_{N\to\infty} \langle h^{t+1}, h^{s+1} \rangle = \mathbb{E}\{\tilde{Z}_t \tilde{Z}_s\}\,. \tag{B.1}$$

Therefore, all we need to show is that for all $s, t \geq 0$: $\mathsf{R}_{t,s}$ and $\tilde{\mathsf{R}}_{t,s}$ are equal. We prove this by induction on $\max(s,t)$.

- For $s = t = 0$ we have using Lemma F.3(b) almost surely

$$\tilde{\mathsf{R}}_{0,0} \equiv \lim_{N\to\infty} \langle h^1, h^1 \rangle = \tau_0^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\{X_0^2\}\,,$$

  that is equal to $\mathsf{R}_{0,0}$.

- *Induction hypothesis:* Assume that for all $s \leq k$ and $t \leq k$,

$$\mathsf{R}_{t,s} = \tilde{\mathsf{R}}_{t,s}\,. \tag{B.2}$$

- Then we prove Eq. (B.2) for $t = k+1$ (case $s = k+1$ is similar). First assume $s = 0$ and $t = k+1$ in which using Lemma F.3(c) we have almost surely

$$\begin{aligned}
\tilde{\mathsf{R}}_{k+1,0} &= \lim_{N\to\infty} \langle h^{k+2}, h^1 \rangle = \lim_{n\to\infty} \langle m^{k+1}, m^0 \rangle \\
&= \lim_{n\to\infty} \langle b^{k+1} - w, b^0 - w \rangle = \sigma^2 + \frac{1}{\delta} \lim_{N\to\infty} \langle q^{k+1}, q^0 \rangle \\
&= \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(X_0 - \tilde{Z}_k;\theta_k) - X_0][-X_0]\right\}\,, \\
&= \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta(X_0 + \tilde{Z}_k;\theta_k) - X_0][-X_0]\right\}\,,
\end{aligned}$$

  where the last equality uses $q^0 = -x_0$ and Lemma F.3(b) for the pseudo-Lipschitz function $(h_i^{k+1}, x_{0,i}) \mapsto [\eta(x_{0,i} - h_i^{k+1};\theta_k) - x_{0,i}][-x_{0,i}]$. Here $X_0 \sim p_{X_0}$ and $\tilde{Z}_k$ are independent and the latter is mean zero gaussian with $\mathbb{E}\{\tilde{Z}_k^2\} = \tilde{\mathsf{R}}_{k,k}$. But using the induction hypothesis, $\tilde{\mathsf{R}}_{k,k} = \mathsf{R}_{k,k}$ holds. Hence, we can apply Eq. (4.14) to obtain $\tilde{\mathsf{R}}_{t,0} = \mathsf{R}_{t,0}$.

  Similarly, for the case $t = k+1$ and $s > 0$, using Lemma F.3(b)(c) we have almost surely

$$\begin{aligned}
\tilde{\mathsf{R}}_{k+1,s} &= \lim_{N\to\infty} \langle h^{k+2}, h^{s+1} \rangle = \lim_{n\to\infty} \langle m^{k+1}, m^s \rangle \\
&= \lim_{n\to\infty} \langle b^{k+1} - w, b^s - w \rangle = \sigma^2 + \frac{1}{\delta} \lim_{N\to\infty} \langle q^{k+1}, q^s \rangle \\
&= \sigma^2 + \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + \tilde{Z}_k;\theta_k) - X_0][\eta(X_0 + \tilde{Z}_{s-1};\theta_{s-1}) - X_0]\}\,,
\end{aligned}$$

  for $X_0 \sim p_{X_0}$ independent of zero mean gaussian variables $\tilde{Z}_k$ and $\tilde{Z}_{s-1}$ that satisfy

$$\mathsf{R}_{k,s-1} = \mathbb{E}\{\tilde{Z}_k \tilde{Z}_{s-1}\}\,, \quad \mathsf{R}_{k,k} = \mathbb{E}\{\tilde{Z}_k^2\}\,, \quad \mathsf{R}_{s-1,s-1} = \mathbb{E}\{\tilde{Z}_{s-1}^2\}\,,$$

  using the induction hypothesis. Hence the result follows.

C Proof of Lemma 4.3

The proof of Lemma 4.3 relies on Lemma 5.7 which we will prove in the first subsection.

C.1 Proof of Lemma 5.7

Before proving Lemma 5.7, we state and prove the following property of gaussian random variables.

Lemma C.1. *Let Z_1 and Z_2 be jointly gaussian random variables with $\mathbb{E}(Z_1^2) = \mathbb{E}(Z_2^2) = 1$ and $\mathbb{E}(Z_1 Z_2) = c \geq 0$. Let I be a measurable subset of the real line. Then $\mathbb{P}(Z_1 \in I,\ Z_2 \in I)$ is an increasing function of $c \in [0,1]$.*

Proof. Let $\{X_s\}_{s\in\mathbb{R}}$ be the standard Ornstein-Uhlenbeck process. Then (Z_1, Z_2) is distributed as (X_0, X_t) for t satisfying $c = e^{-2t}$. Hence

$$\mathbb{P}(Z_1 \in I,\ Z_2 \in I) = \mathbb{E}[f(X_0)f(X_t)]\,, \tag{C.1}$$

for f the indicator function of I. Since the Ornstein-Uhlenbeck process is reversible with respect to the standard gaussian measure $\mu_{\rm G}$, we have

$$\mathbb{E}[f(X_0)f(X_t)] = \sum_{\ell=0}^{\infty} e^{-\lambda_\ell t}\,(\psi_\ell, f)_{\mu_{\rm G}}^2 = \sum_{\ell=0}^{\infty} c^{\frac{\lambda_\ell}{2}}\,(\psi_\ell, f)_{\mu_{\rm G}}^2 \tag{C.2}$$

with $0 \leq \lambda_0 \leq \lambda_1 \leq \dots$ the eigenvalues of its generator, $\{\psi_\ell\}_{\ell\geq 0}$ the corresponding eigenvectors and $(\,\cdot\,,\,\cdot\,)_{\mu_{\rm G}}$ the scalar product in $L^2(\mu_{\rm G})$. The thesis follows. ☐

We now pass to the proof of Lemma 5.7.

Proof of Lemma 5.7. It is convenient to change coordinates and define

$$y_{t,1} \equiv \mathsf{R}_{t-1,t-1} = \tau_{t-1}^2\,, \quad y_{t,2} \equiv \mathsf{R}_{t,t} = \tau_t^2\,, \quad y_{t,3} \equiv \mathsf{R}_{t-1,t-1} - 2\mathsf{R}_{t,t-1} + \mathsf{R}_{t,t}\,. \tag{C.3}$$

The vector $y_t = (y_{t,1}, y_{t,2}, y_{t,3})$ belongs to $\mathbb{R}_+^3$ by Lemma 5.8. Using Eq. (4.13), it is immediate to see that this is updated according to the mapping

$$
\begin{aligned}
y_{t+1} &= \mathsf{G}(y_t)\,, \\
\mathsf{G}_1(y_t) &\equiv y_{t,2}\,, && \text{(C.4)}\\
\mathsf{G}_2(y_t) &\equiv \sigma^2 + \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + Z_t; \alpha\sqrt{y_{t,2}}) - X_0]^2\}\,, && \text{(C.5)}\\
\mathsf{G}_3(y_t) &\equiv \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + Z_t; \alpha\sqrt{y_{t,2}}) - \eta(X_0 + Z_{t-1}; \alpha\sqrt{y_{t,1}})]^2\}\,. && \text{(C.6)}
\end{aligned}
$$

where (Z_t, Z_{t-1}) are jointly gaussian with zero mean and covariance determined by $\mathbb{E}\{Z_t^2\} = y_{t,2}$, $\mathbb{E}\{Z_{t-1}^2\} = y_{t,1}$, $\mathbb{E}\{(Z_t - Z_{t-1})^2\} = y_{t,3}$. This mapping is defined for $y_{t,3} \leq 2(y_{t,1} + y_{t,2})$.

Next we will show that by induction on t that the stronger inequality $y_{t,3} < (y_{t,1} + y_{t,2})$ holds for all t. We have indeed

$$y_{t+1,1} + y_{t+1,2} - y_{t+1,3} = 2\sigma^2 + \frac{2}{\delta}\,\mathbb{E}\{\eta(X_0 + Z_t; \alpha\sqrt{y_{t,2}})\,\eta(X_0 + Z_{t-1}; \alpha\sqrt{y_{t,1}})\}\,.$$

Since $\mathbb{E}\{Z_t Z_{t-1}\} = (y_{t,1} + y_{t,2} - y_{t,3})/2$ and $x \mapsto \eta(x;\theta)$ is monotone, we deduce that $y_{t,3} < (y_{t,1} + y_{t,2})$ implies that Z_t, Z_{t-1} are positively correlated. Therefore $\mathbb{E}\{\eta(X_0 + Z_t; \alpha\sqrt{y_{t,2}})\, \eta(X_0 + Z_{t-1}; \alpha\sqrt{y_{t,1}})\} \geq 0$, which in turn yields $y_{t+1,3} < (y_{t+1,1} + y_{t+1,2})$.

The initial condition implied by Eq. (4.14) is

$$\begin{aligned} y_{1,1} &= \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\{X_0^2\}\,, \\ y_{1,2} &= \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\{[\eta(X_0 + Z_0;\theta_0) - X_0]^2\}\,, \\ y_{1,3} &= \frac{1}{\delta}\,\mathbb{E}\{\eta(X_0 + Z_0;\theta_0)^2\}\,, \end{aligned}$$

It is easy to check that these satisfy $y_{1,3} < y_{1,1} + y_{1,2}$. (This follows from $\mathbb{E}\{X_0[X_0 - \eta(X_0 + Z_0;\theta_0)]\} > 0$ because $x_0 \mapsto x_0 - \mathbb{E}_Z\eta(x_0 + Z_0;\theta_0)$ is monotone increasing.) We can hereafter therefore assume $y_{t,3} < y_{t,1} + y_{t,2}$ for all t.

We will consider the above iteration for arbitrary initialization y_0 (satisfying $y_{0,3} < y_{0,1} + y_{0,2}$) and will show the following three facts:

Fact *(i)***.** As $t \to \infty$, $y_{t,1}, y_{t,2} \to \tau_*^2$. Further the convergence is monotone.

Fact *(ii)***.** If $y_{0,1} = y_{0,2} = \tau_*^2$ and $y_{0,3} \leq 2\tau_*^2$, then $y_{t,1} = y_{t,2} = \tau_*^2$ for all t and $y_{t,3} \to 0$.

Fact *(iii)***.** The jacobian $J = J_{\mathsf{G}}(y_*)$ of G at $y_* = (\tau_*^2, \tau_*^2, 0)$ has spectral radius $\sigma(J) < 1$.

By simple compactness arguments, Facts *(i)* and *(ii)* imply $y_t \to y_*$ as $t \to \infty$. (Notice that $y_{t,3}$ remains bounded since $y_{t,3} \leq (y_{t,1} + y_{t,2})$ and by the convergence of $y_{t,1}, y_{t,2}$.) Fact *(iii)* implies that convergence is exponentially fast.

Proof of Fact (i). Notice that $y_{t,2}$ evolves independently by $y_{t+1,2} = \mathsf{G}_2(y_t) = \mathsf{F}(y_{2,t}, \alpha\sqrt{y_{2,t}})$, with $\mathsf{F}(\,\cdot\,,\,\cdot\,)$ the state evolution mapping introduced in Eq. (1.6). It follows from Proposition 1.3 that $y_{t,2} \to \tau_*^2$ monotonically for any initial condition. Since $y_{t+1,1} = y_{t,2}$, the same happens for $y_{t,1}$.

Proof of Fact (ii). Consider the function $\mathsf{G}_*(x) = \mathsf{G}_3(\tau_*^2, \tau_*^2, x)$. This is defined for $x \in [0, 4\tau_*^2]$ but since $y_{t,3} < y_{t,1} + y_{t,2}$ we will only consider $\mathsf{G}_* : [0, 2\tau_*^2] \to \mathbb{R}_+$. Obviously $\mathsf{G}_*(0) = 0$. Further G_* can be represented as follows in terms of the independent random variables Z, $W \sim \mathsf{N}(0,1)$:

$$\mathsf{G}_*(x) = \frac{1}{\delta}\mathbb{E}\{[\eta(X_0 + \sqrt{\tau_*^2 - x/4}Z + (\sqrt{x}/2)W; \alpha\tau_*) - \eta(X_0 + \sqrt{\tau_*^2 - x/4}Z - (\sqrt{x}/2)W; \alpha\tau_*)]^2\} \tag{C.7}$$

A straightforward calculation yields

$$\mathsf{G}_*'(x) = \frac{1}{\delta}\mathbb{E}\{\eta'(X_0 + Z_t; \alpha\tau_*)\eta'(X_0 + Z_{t-1}; \alpha\tau_*)\} = \frac{1}{\delta}\mathbb{P}\{|X_0 + Z_t| \geq \alpha\tau_*,\ |X_0 + Z_{t-1}| \geq \alpha\tau_*\}\,,$$

where $Z_{t-1} = \sqrt{\tau_*^2 - x^2/4}Z + (x/2)W$, $Z_t = \sqrt{\tau_*^2 - x^2/4}Z - (x/2)W$. In particular, by Lemma C.1, $x \mapsto \mathsf{G}_*(x)$ is strictly increasing (notice that the covariance of Z_{t-1} and Z_t is $\tau_*^2 - (x/2)$ which is decreasing in x). Further

$$\mathsf{G}_*'(0) = \frac{1}{\delta}\mathbb{E}\{\eta'(X_0 + \tau_*\, Z; \alpha\tau_*)\}\,.$$

Hence, since $\lambda > 0$ using Eq. (1.15) we have $\mathsf{G}'(0) < 1$. Finally, by Lemma C.1, $x \mapsto \mathsf{G}'(x)$ is decreasing in $[0, 2\tau_*)$. It follows that $y_{t,3} \le \mathsf{G}'(0)^t y_{0,3} \to 0$ as claimed.

Proof of Fact (iii). From the definition of G, we have the following expression for the Jacobian

$$J_{\mathsf{G}}(y_*) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \mathsf{F}'(\tau_*^2) & 0 \\ a & \mathsf{G}_*'(0) & b \end{pmatrix}$$

where with an abuse of notation we let $\mathsf{F}'(\tau_*^2) \equiv \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathsf{F}(\tau^2, \alpha\tau)\Big|_{\tau^2=\tau_*^2}$. Computing the eigenvalues of the above matrix, we get

$$\sigma(J) = \max\left\{\, \mathsf{F}'(\tau_*^2)\,,\, \mathsf{G}_*'(0)\,\right\}.$$

Since $\mathsf{G}_*'(0) < 1$ as proved above, and $\mathsf{F}(\tau_*^2) < 1$ as per Proposition 1.3, the claim follows. □

C.2 Lemma 5.7 implies Lemma 4.3

Using representations (4.4) and (4.3) (i.e., $b^t = w - z^t$ and $q^t = x_0 - x^t$) and Lemma F.3(c) we obtain,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\|z^{t+1} - z^t\|_2^2 &= \lim_{n\to\infty} \frac{1}{n}\|b^{t+1} - b^t\|_2^2 \\ &\overset{\text{a.s.}}{=} \frac{1}{\delta} \lim_{N\to\infty} \frac{1}{N}\|q^{t+1} - q^t\|_2^2 \\ &= \frac{1}{\delta} \lim_{N\to\infty} \frac{1}{N}\|x^{t+1} - x^t\|_2^2\,, \end{aligned}$$

where the last equality uses $q^t = x^t - x_0$. Therefore, it is sufficient to prove the thesis for $\|x^{t+1} - x^t\|_2$. By state evolution, Theorem 4.2, we have

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N}\|x^{t+1} - x^t\|_2^2 &= \mathbb{E}\big\{\big[\eta(X_0 + Z_t;\theta_t) - \eta(X_0 + Z_{t-1};\theta_{t-1})\big]^2\big\} \\ &\le 2(\theta_t - \theta_{t-1})^2 + 2\,\mathbb{E}\{(Z_t - Z_{t-1})^2\} = 2(\theta_t - \theta_{t-1})^2 + 2(\mathsf{R}_{t,t} - 2\mathsf{R}_{t,t-1} + \mathsf{R}_{t-1,t-1})\,. \end{aligned}$$

The first term vanishes as $t \to \infty$ because $\theta_t = \alpha\tau_t \to \alpha\tau_*$ by Proposition 1.3. The second term instead vanishes since $\mathsf{R}_{t,t} \to \tau_*$, $\mathsf{R}_{t,t-1} \to \tau_*$ by Lemma 5.7.

D Proof of Lemma 5.2

First note that the upper bound on $\lambda_{\max}(R/N)$ is trivial since using representations (4.7), (4.8), $q^t = f_t(h^t, x_0)$, $m^t = g_t(b^t, w)$ and Lemma F.3(c)(d) all entries of the matrix R/N are bounded as $N \to \infty$ and the matrix has fixed dimensions. Hence, we only focus on the lower-bound for $\lambda_{\min}(R/N)$.

The result for $R = M^*M$ and $R = Q^*Q$ follows directly from Lemma F.3(g) and Lemma 8 of [BM11].

For $R = Y^*Y$ and $R = X^*X$ the proof is by induction on t.

- For $t = 1$ we have $Y_t = b^0$ and $X_t = h^1 + \xi_0 q^0 = h^1 - x_0$. Using Lemma F.3(b)(c) we obtain almost surely

$$\lim_{N\to\infty} \frac{Y_t^* Y_t}{N} = \delta \lim_{n\to\infty} \langle b^0, b^0 \rangle = \lim_{N\to\infty} \langle q^0, q^0 \rangle = \mathbb{E}\{X_0^2\}\,,$$
$$\lim_{N\to\infty} \frac{X_t^* X_t}{N} = \lim_{N\to\infty} \langle h^1 - x^0, h^1 - x^0 \rangle = \mathbb{E}\{(\tau_0 Z_0 + X_0)^2\} = \sigma^2 + \frac{\delta+1}{\delta}\mathbb{E}\{X_0^2\}\,,$$

where both are positive by the assumption $\mathbb{P}\{X_0 \neq 0\} > 0$.

- *Induction hypothesis:* Assume that for all $t \leq k$ there exist positive constants $c_X(t)$ and $c_Y(t)$ such that as $N \to \infty$

$$c_Y(t) \leq \lambda_{\min}(\frac{Y_t^* Y_t}{N})\,, \tag{D.1}$$
$$c_X(t) \leq \lambda_{\min}(\frac{X_t^* X_t}{N})\,. \tag{D.2}$$

- Now we prove Eq. (D.1) for $t = k+1$ (proof of (D.2) is similar). We will prove that there is a positive constant c such that as $N \to \infty$, for any vector $\vec{a}_t \in \mathbb{R}^t$:

$$\langle Y_t\, \vec{a}_t, Y_t\, \vec{a}_t \rangle \geq c \|\vec{a}_t\|_2^2\,.$$

First write $\vec{a}_t = (a_1, \dots, a_t)$ and denote its first $t-1$ coordinates with $\vec{a}_{t-1}$. Next, we consider the conditional distribution $A|_{\mathfrak{S}_{t-1}}$. Using Eqs. (4.9) and (4.10) we obtain (since $Y_t = AQ_t$)

$$\begin{aligned} Y_t\, \vec{a}_t|_{\mathfrak{S}_{t-1}} &\stackrel{\text{d}}{=} A|_{\mathfrak{S}_{t-1}}(Q_{t-1}\, \vec{a}_{t-1} + a_t q^{t-1}) \\ &= E_{t-1}(Q_{t-1}\, \vec{a}_{t-1} + a_t q^{t-1}) + a_t P^{\perp}_{M_{t-1}} \tilde{A} q_{\perp}^{t-1}\,. \end{aligned}$$

Hence, conditional on $\mathfrak{S}_{t-1}$ we have, almost surely

$$\lim_{N\to\infty} \langle Y_t\, \vec{a}_t, Y_t\, \vec{a}_t \rangle = \lim_{N\to\infty} \frac{1}{N} \|Y_{t-1}\, \vec{a}_{t-1} + a_t E_{t-1} q^{t-1}\|^2 + a_t^2 \lim_{N\to\infty} \langle q_{\perp}^{t-1}, q_{\perp}^{t-1} \rangle\,. \tag{D.3}$$

Here we used the fact that $\tilde{A}$ is a random matrix with i.i.d. $\mathsf{N}(0, 1/n)$ entries independent of $\mathfrak{S}_{t-1}$ (cf. Lemma F.4) which implies that almost surely

- $\lim_{N\to\infty} \langle P^{\perp}_{M_{t-1}} \tilde{A} q_{\perp}^{t-1}, P^{\perp}_{M_{t-1}} \tilde{A} q_{\perp}^{t-1} \rangle = \lim_{N\to\infty} \langle q_{\perp}^{t-1}, q_{\perp}^{t-1} \rangle$,

- $\lim_{N\to\infty} \langle P^{\perp}_{M_{t-1}} \tilde{A} q_{\perp}^{t-1}, Y_{t-1}\, \vec{a}_{t-1} + a_t b^{t-1} + a_t \lambda_{t-1} m^{t-2} \rangle = 0$.

From Lemma F.3(g) we know that $\lim_{N\to\infty} \langle q_{\perp}^{t-1}, q_{\perp}^{t-1} \rangle$ is larger than a positive constant ς_t. Hence, from representation (D.3) and induction hypothesis (D.1)

$$\lim_{N\to\infty} \langle Y_t\, \vec{a}_t, Y_t\, \vec{a}_t \rangle \geq \lim_{N\to\infty} \left[\sqrt{c_Y(t-1)} \|\vec{a}_{t-1}\| - \frac{|a_t|}{\sqrt{N}} \|b^{t-1} + \lambda_{t-1} m^{t-2}\| \right]^2 + a_t^2 \varsigma_t\,.$$

To simplify the notation let $c'_t \equiv \lim_{N\to\infty} N^{-1/2} \|b^{t-1} + \lambda_{t-1} m^{t-2}\|$. Now if $c'_t |a_t| \leq \sqrt{c_Y(t-1)} \|\vec{a}_{t-1}\|/2$ then

$$\lim_{N\to\infty} \langle Y_t\, \vec{a}_t, Y_t\, \vec{a}_t \rangle \geq \frac{c_Y(t-1)}{4} \|\vec{a}_{t-1}\|^2 + a_t^2 \varsigma_t \geq \min\left(\frac{c_Y(t-1)}{4}, \varsigma_t \right) \|\vec{a}_t\|_2^2\,, \tag{D.4}$$

which proves the result. Otherwise, we obtain the inequality

$$\lim_{N\to\infty} \langle Y_t\,\vec{a}_t, Y_t\,\vec{a}_t\rangle \geq a_t^2 \varsigma_t \geq \left(\frac{\varsigma_t\, c_Y(t-1)}{4(c_t')^2 + c_Y(t-1)}\right) \|\vec{a}_t\|_2^2\,,$$

that completes the induction argument.

E A concentration estimate

The following proposition follows from standard concentration-of-measure arguments.

Proposition E.1. *Let $V \subseteq \mathbb{R}^m$ a uniformly random linear space of dimension d. For $\lambda \in (0,1)$, let P_λ denote the orthogonal projector on the first $m\lambda$ coordinates of $\mathbb{R}^m$. Define $Z(\lambda) \equiv \sup\{\|P_\lambda v\| \,:\, v \in V,\ \|v\| = 1\}$. Then, for any $\varepsilon > 0$ there exists $c(\varepsilon) > 0$ such that, for all m large enough (and d fixed)*

$$\mathbb{P}\{|Z(\kappa) - \sqrt{\lambda}| \geq \varepsilon\} \leq e^{-m\,c(\varepsilon)}\,. \tag{E.1}$$

Proof. Let $Q \in \mathbb{R}^{m\times d}$ be a uniformly random orthogonal matrix. Its image is a uniformly random subspace of $\mathbb{R}^m$ whence the following equivalent characterization of $Z(\lambda)$ is obtained

$$Z(\lambda) \stackrel{\mathrm{d}}{=} \sup\{\|P_\lambda Q u\| \,:\, u \in S^d\}$$

where $S^d \equiv \{x \in \mathbb{R}^d \,:\, \|x\| = 1\}$ is the d-dimensional sphere, and $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution.

Let $N_d(\varepsilon/2)$ be a $(\varepsilon/2)$-net in S_d, i.e. a subset of vectors $\{u^1, \dots, u^M\} \in S^d$ such that, for any $u \in S^d$, there exists $i \in \{1, \dots, M\}$ such that $\|u - u^i\| \leq \varepsilon/2$. It follows from a standard counting argument [Led01] that there exists an $(\varepsilon/2)$-net of size $|N_d(\varepsilon/2)| \equiv M \leq (100/\varepsilon)^d$. Define

$$Z_{\varepsilon/2}(\lambda) \equiv \sup\{\|P_\lambda Q u\| \,:\, u \in N_d(\varepsilon/2)\}\,.$$

Since $u \mapsto P_\lambda Q u$ is Lipschitz with modulus 1, we have

$$\begin{aligned}\mathbb{P}\{|Z(\kappa) - \sqrt{\lambda}| \geq \varepsilon\} &\leq \mathbb{P}\{|Z_{\varepsilon/2}(\kappa) - \sqrt{\lambda}| \geq \varepsilon/2\} \\ &\leq \sum_{i=1}^{M} \mathbb{P}\{|\|P_\lambda Q u^i\| - \sqrt{\lambda}| \geq \varepsilon/2\}\,.\end{aligned}$$

But for each i, Qu^i is a uniformly random vector with norm 1 in $\mathbb{R}^m$. By concentration of measure in S^m [Led01], there exists a function $c(\varepsilon) > 0$ such that, for $x \in S^m$ uniformly random

$$\mathbb{P}\big\{\big|\|P_\lambda x\| - \sqrt{\lambda}\big| \geq \varepsilon/2\big\} \leq e^{-m\,c(\varepsilon)}\,.$$

Therefore we get

$$\mathbb{P}\{|Z(\kappa) - \sqrt{\lambda}| \geq \varepsilon\} \leq |N_d(\varepsilon/2)| e^{-m\,c(\varepsilon)} \leq \Big(\frac{100}{\varepsilon}\Big)^d e^{-m\,c(\varepsilon)}$$

which is smaller than $e^{-mc(\varepsilon)/2}$ for all m large enough. ☐

F Useful reference material

In this appendix we collect a few known results that are used several times in our proof. We also provide some pointers to the literature.

F.1 Equivalence of ℓ^2 and ℓ^1 norm on random vector spaces

In our proof we make use of the following well-known result of Kashin in the theory of diameters of smooth functions [Kas77].

Theorem F.1 (Kashin 1977)**.** *For any positive number υ there exist a universal constant c_υ such that for any $n \geq 1$, with probability at least $1-2^{-n}$, for a uniformly random subspace $V_{n,\upsilon}$ of dimension $\lfloor n(1-\upsilon)\rfloor$,*

$$\forall\; x \in V_{n,\upsilon}: \qquad c_\upsilon \|x\|_2 \leq \frac{1}{\sqrt{n}}\|x\|_1\,.$$

F.2 Singular values of random matrices

We will repeatedly make use of limit behavior of extreme singular values of random matrices. A very general result was proved in [BY93] (see also [BS05]).

Theorem F.2 ([BY93])**.** *Let $A \in \mathbb{R}^{n\times N}$ be a matrix with i.i.d. entries such that $\mathbb{E}\{A_{ij}\} = 0$, $\mathbb{E}\{A_{ij}^2\} = 1/n$, and $n = N\delta$. Let $\sigma_{\max}(A)$ be the largest singular value of A, and $\hat{\sigma}_{\min}(A)$ be its smallest non-zero singular value. Then*

$$\lim_{N\to\infty} \sigma_{\max}(A) \overset{\text{a.s.}}{=} \frac{1}{\sqrt{\delta}} + 1\,, \tag{F.1}$$

$$\lim_{N\to\infty} \hat{\sigma}_{\min}(A) \overset{\text{a.s.}}{=} \frac{1}{\sqrt{\delta}} - 1\,. \tag{F.2}$$

We will also use the following fact that follows from the standard singular value decomposition

$$\min\big\{\|Ax\|_2\;:\; x\in \ker(A)^{\perp},\; \|x\|=1\big\} = \sigma_{\min}(A)\,. \tag{F.3}$$

F.3 Two Lemmas from [BM11]

Our proof uses the results of [BM11]. We state copy here the crucial technical lemma in that paper. Notations refer to the general algorithm in Eq. (4.1). General state evolution defines quantities $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 0}$ via

$$\tau_t^2 = \mathbb{E}\big\{g_t(\sigma_t Z, W)^2\big\}\,, \qquad \sigma_t^2 = \frac{1}{\delta}\,\mathbb{E}\big\{f_t(\tau_{t-1}Z, X_0)^2\big\}\,, \tag{F.4}$$

where $W \sim p_W$ and $X_0 \sim p_{X_0}$ are independent of $Z \sim \mathsf{N}(0,1)$

Lemma F.3. *Let $\{q_0(N)\}_{N\geq 0}$ and $\{A(N)\}_{N\geq 0}$ be, respectively, a sequence of deterministic initial conditions and a sequence of matrices $A \in \mathbb{R}^{n\times N}$ indexed by N with i.i.d. entries $A_{ij} \sim \mathsf{N}(0,1/n)$. Assume $n/N \to \delta \in (0,\infty)$. Consider deterministic sequences of vectors $\{x_0(N), w(N)\}_{N\geq 0}$, whose empirical distributions converge weakly to probability measures p_{X_0} and p_W on $\mathbb{R}$ with bounded $(2k-2)^{th}$ moment, and assume:*

(i) $\lim_{N\to\infty}\mathbb{E}_{\hat{p}_{x_0(N)}}(X_0^{2k-2}) = \mathbb{E}_{p_{X_0}}(X_0^{2k-2}) < \infty$.

(ii) $\lim_{N\to\infty}\mathbb{E}_{\hat{p}_{w(N)}}(W^{2k-2}) = \mathbb{E}_{p_W}(W^{2k-2}) < \infty$.

(iii) $\lim_{N\to\infty}\mathbb{E}_{\hat{p}_{q_0(N)}}(X^{2k-2}) < \infty$.

Let $\{\sigma_t, \tau_t\}_{t\geq 0}$ *be defined uniquely by the recursion* (F.4) *with initialization* $\sigma_0^2 = \delta^{-1}\lim_{n\to\infty}\langle q^0, q^0\rangle$*. Then the following hold for all* $t \in \mathbb{N}\cup\{0\}$

(*a*)

$$h^{t+1}|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1}\alpha_i h^{i+1} + \tilde{A}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(1)\,, \tag{F.5}$$

$$b^t|_{\mathfrak{S}_{t,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1}\beta_i b^i + \tilde{A} q_\perp^t + \tilde{M}_t\vec{o}_t(1)\,, \tag{F.6}$$

where $\tilde{A}$ *is an independent copy of* A *and the matrix* $\tilde{Q}_t$ *(*$\tilde{M}_t$*) is such that its columns form an orthogonal basis for the column space of* Q_t *(*M_t*) and* $\tilde{Q}_t^*\tilde{Q}_t = N\,\mathbf{I}_{t\times t}$ *(*$\tilde{M}_t^*\tilde{M}_t = n\,\mathbf{I}_{t\times t}$*).*

(*b*) *For all pseudo-Lipschitz functions* $\phi_h, \phi_b : \mathbb{R}^{t+2}\to\mathbb{R}$ *of order* k

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi_h(h_i^1,\ldots,h_i^{t+1},x_{0,i}) \stackrel{\text{a.s.}}{=} \mathbb{E}\big\{\phi_h(\tau_0 Z_0,\ldots,\tau_t Z_t, X_0)\big\}\,, \tag{F.7}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\phi_b(b_i^0,\ldots,b_i^t,w_i) \stackrel{\text{a.s.}}{=} \mathbb{E}\big\{\phi_b(\sigma_0\hat{Z}_0,\ldots,\sigma_t\hat{Z}_t, W)\big\}\,, \tag{F.8}$$

where $(Z_0,\ldots,Z_t)$ *and* $(\hat{Z}_0,\ldots,\hat{Z}_t)$ *are two zero-mean gaussian vectors independent of* X_0*,* W*, with* $Z_i, \hat{Z}_i \sim \mathsf{N}(0,1)$*.*

(*c*) *For all* $0 \leq r, s \leq t$ *the following equations hold and all limits exist, are bounded and have degenerate distribution (i.e. they are constant random variables):*

$$\lim_{N\to\infty}\langle h^{r+1}, h^{s+1}\rangle \stackrel{\text{a.s.}}{=} \lim_{n\to\infty}\langle m^r, m^s\rangle\,, \tag{F.9}$$

$$\lim_{n\to\infty}\langle b^r, b^s\rangle \stackrel{\text{a.s.}}{=} \frac{1}{\delta}\lim_{N\to\infty}\langle q^r, q^s\rangle\,. \tag{F.10}$$

(*d*) *For all* $0 \leq r, s \leq t$*, and for any Lipschitz function* $\varphi : \mathbb{R}^2 \to \mathbb{R}$ *, the following equations hold and all limits exist, are bounded and have degenerate distribution (i.e. they are constant random variables):*

$$\lim_{N\to\infty}\langle h^{r+1}, \varphi(h^{s+1}, x_0)\rangle \stackrel{\text{a.s.}}{=} \lim_{N\to\infty}\langle h^{r+1}, h^{s+1}\rangle\langle\varphi'(h^{s+1}, x_0)\rangle, \tag{F.11}$$

$$\lim_{n\to\infty}\langle b^r, \varphi(b^s, w)\rangle \stackrel{\text{a.s.}}{=} \lim_{n\to\infty}\langle b^r, b^s\rangle\langle\varphi'(b^s, w)\rangle\,. \tag{F.12}$$

Here φ' *denotes derivative with respect to the first coordinate of* φ*.*

(e) *For $\ell = k-1$, the following hold almost surely*

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}(h_i^{t+1})^{2\ell} < \infty\,, \tag{F.13}$$

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(b_i^{t})^{2\ell} < \infty. \tag{F.14}$$

(f) *For all $0 \le r \le t$:*

$$\lim_{N\to\infty} \frac{1}{N}\langle h^{r+1}, q^0\rangle \overset{\text{a.s.}}{=} 0\,. \tag{F.15}$$

(g) *For all $0 \le r \le t$ and $0 \le s \le t-1$ the following limits exist, and there exist strictly positive constants ρ_r and ς_s (independent of N, n) such that almost surely*

$$\lim_{N\to\infty} \langle q_\perp^r, q_\perp^r\rangle > \rho_r\,, \tag{F.16}$$

$$\lim_{n\to\infty} \langle m_\perp^s, m_\perp^s\rangle > \varsigma_s\,. \tag{F.17}$$

It is also useful to recall some simple properties of gaussian random matrices.

Lemma F.4. *For any deterministic $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^n$ with $\|u\| = \|v\| = 1$ and a gaussian matrix $\tilde{A}$ distributed as A we have*

(a) $v^*\tilde{A}u \overset{\text{d}}{=} Z/\sqrt{n}$ *where* $Z \sim \mathsf{N}(0,1)$.

(b) $\lim_{n\to\infty}\|\tilde{A}u\|^2 = 1$ *almost surely.*

(c) *Consider, for $d \le n$, a d-dimensional subspace W of $\mathbb{R}^n$, an orthogonal basis $w_1, \dots, w_d$ of W with $\|w_i\|^2 = n$ for $i = 1, \dots, d$, and the orthogonal projection P_W onto W. Then for $D = [w_1|\dots|w_d]$, we have $P_W Au \overset{\text{d}}{=} Dx$ with $x \in \mathbb{R}^d$ that satisfies: $\lim_{n\to\infty}\|x\| \overset{\text{a.s.}}{=} 0$ (the limit being taken with d fixed). Note that x is $\vec{o}_d(1)$ as well.*

References

[AG07] G. Andrew and J. Gau, *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: asymptotic results and real world examples*, Neural Information Processing Systems (NIPS), 2010.

[BBM11] M. Bayati, J .Bento, and A. Montanari, *The LASSO risk: asymptotic results and real world examples*, in preparation, 2011.

[BM11] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.

[BS05] Z. Bai and J. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer, 2005.

[BT09] A. Beck and M. Teboulle, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM J. Imaging Sciences **2** (2009), 183–202.

[BY93] Z. D. Bai and Y. Q. Yin, *Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix*, The Annals of Probability **21** (1993), 1275–1294.

[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.

[DJ94] D. L. Donoho and I. M. Johnstone, *Minimax risk over l_p balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.

[DJ98] ———, *Minimax estimation via wavelet shrinkage*, Annals of Statistics **26** (1998), 879–921.

[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.

[DMM10] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, Preprint, 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.

[Joh06] I. Johnstone, *High Dimensional Statistical Inference and Random Matrices*, Proc. International Congress of Mathematicians (Madrid), 2006.

[Kas77] B. Kashin, *Diameters of Some Finite-Dimensional Sets and Classes of Smooth Functions*, Math. USSR Izv. **11** (1977), 317–333.

[KM10] S. Korada and A. Montanari, *Applications of Lindeberg Principle in Communications and Statistical Learning*, preprint available in 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.

[Led01] M. Ledoux, *The concentration of measure phenomenon*, American Mathematical Society, Berlin, 2001.

[LPRTJ05] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Advances in Mathematics **195** (2005), 491–523.

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

[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.

[Tel99] E. Telatar, *Capacity of Multi-antenna Gaussian Channels*, European Transactions on Telecommunications **10** (1999), 585–595.

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

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

Mohsen Bayati is an assistant professor of operations and information technology at Stanford university Graduate School of Business. Mohsen received his PhD in Electrical Engineering from Stanford University in 2007. His dissertation was on machine learning and modeling aspects of large-scale networks. During the summers of 2005 and 2006 he interned at IBM Research and Microsoft Research respectively. He was a Postdoctoral Researcher with Microsoft Research from 2007 to 2009 working mainly on applications of machine learning and optimization methods in healthcare and online advertising. In particular, he focused on hospital readmissions. He has been a Postdoctoral Scholar at Stanford University from 2009 to 2011 with a research focus in high-dimensional statistical data-mining.

Andrea Montanari is an associate professor in the Departments of Electrical Engineering and of Statistics, Stanford University. He received the Laurea degree in physics in 1997, and the Ph.D. degree in theoretical physics in 2001, both from Scuola Normale Superiore, Pisa, Italy. He has been a Postdoctoral Fellow with the Laboratoire de Physique Théorique of Ecole Normale Supérieure (LPTENS), Paris, France, and the Mathematical Sciences Research Institute, Berkeley, CA. Since 2002, he has been Chargé de Recherche (a research position with Centre National de la Recherche Scientifique, CNRS) at LPTENS. In September 2006, he joined the faculty of Stanford University. Dr. Montanari was coawarded the ACM SIGMETRICS Best Paper Award in 2008. He received the CNRS Bronze Medal for Theoretical Physics in 2006 and the National Science Foundation CAREER award in 2008. His research focuses on algorithms on graphs, graphical models, statistical inference and estimation.