The dynamics of message passing on dense graphs, with applications to compressed sensing

Mohsen Bayati* and Andrea Montanari*,†

Abstract

'Approximate message passing' algorithms proved to be extremely effective in reconstructing sparse signals from a small number of incoherent linear measurements. Extensive numerical experiments further showed that their dynamics is accurately tracked by a simple one-dimensional iteration termed *state evolution*. In this paper we provide the first rigorous foundation to state evolution. We prove that indeed it holds asymptotically in the large system limit for sensing matrices with independent and identically distributed gaussian entries.

While our focus is on message passing algorithms for compressed sensing, the analysis extends beyond this setting, to a general class of algorithms on dense graphs. In this context, state evolution plays the role that density evolution has for sparse graphs.

The proof technique is fundamentally different from the standard approach to density evolution, in that it copes with large number of short loops in the underlying factor graph. It relies instead on a conditioning technique recently developed by Erwin Bolthausen in the context of spin glass theory.

1 Introduction and main results

Given an $n \times N$ matrix A, the compressed sensing reconstruction problem requires to reconstruct a sparse vector $x_0 \in \mathbb{R}^N$ from a (small) vector of linear observations $y = Ax_0 + w \in \mathbb{R}^n$. Here w is noise vector and A is assumed to be known. Recently [DMM09] suggested the following first order *approximate message-passing (AMP)* algorithm for reconstructing x_0 given A, y. Start with an initial guess $x^0 = 0$ and proceed by

$$\begin{aligned} x^{t+1} &= \eta_t(A^* z^t + x^t), \\ z^t &= y - Ax^t + \frac{1}{\delta} z^{t-1} \left\langle \eta'_{t-1}(A^* z^{t-1} + x^{t-1}) \right\rangle , \end{aligned} \tag{1.1}$$

for an appropriate sequence of non-linear functions $\{\eta_t\}_{t\geq 0}$. (Here by convention any variable with negative index is assumed to be 0.) The algorithm succeeds if x^t converges to a good approximation of x_0 (cf. [DMM09] for details).

Throughout this paper, the matrix A is normalized in such a way that its columns have ℓ_2 norm (approximately) equal to 1. Given a vector $v \in \mathbb{R}^N$ we write $f(x)$ for the vector obtained by applying f componentwise. Further, $\delta = n/N$, $\langle v \rangle \equiv N^{-1} \sum_{i=1}^{N} v_i$ and A^* is the transpose of matrix A.

Three findings were presented in [DMM09]:

*Department of Electrical Engineering, Stanford University

†Department of Statistics, Stanford University

(1) For random or pseudo-random matrices A, the behavior of AMP algorithm is accurately described by a formalism called 'state evolution' (SE);

(2) The sparsity-undersampling tradeoff of AMP as derived from SE coincides, for an appropriate choice of the functions η_t, with the one of (much more complex) convex optimization approaches;

(3) As a consequence of (1) and (2), SE allows to re-derive reconstruction phase boundaries earlier determined via random polytope geometry [DT05, DT09].

These findings were based on heuristic arguments and extensive numerical simulations. In this paper we provide the first rigorous support to finding (1), by proving that SE holds in the large system limit, for random sensing matrices A with gaussian entries. Implications on points (2) and (3) above are quite strong and will be reported in a forthcoming paper.

Even more interestingly, state evolution provides sharp predictions that cannot be derived from random polytope geometry. A prominent example is the noise sensitivity of LASSO, which is investigated in [DMM10b].

Note that AMP is an approximation to the following message-passing algorithm. For all $i, j \in [N]$ and $a, b \in [n]$ (here and below $[N] \equiv \{1, 2, \dots, N\}$) start with messages $x^0_{j\to a} = 0$ and proceed by

$$z^t_{a\to i} = y_a - \sum_{j\in[N]\setminus i} A_{aj} x^t_{j\to a}\,, \tag{1.2}$$

$$x^{t+1}_{i\to a} = \eta_t\left(\sum_{b\in[n]\setminus a} A_{bi} z^t_{b\to i}\right).$$

As argued in [DMM10a], AMP accurately approximates message passing in the large system limit. An important tool for the analysis of message passing algorithms is provided by density evolution [RU08]. Density evolution is known to hold asymptotically for sequences of sparse graphs that are locally tree-like. The factor graph underlying the algorithm (1.2) is dense: indeed it is the complete bipartite graph. State evolution plays the role of density evolution for dense graphs, and can be regarded (in a very precise sense) as the limit of density evolution for dense graphs.

For the sake of concreteness, we will focus in this Section on the algorithm (1.1), and will keep to the compressed sensing language. Nevertheless our analysis applies to a much larger family of message passing algorithms on dense graphs, for instance the multi-user detection algorithm studied in [Kab03, NS05, MT06]. Applications to such algorithms are discussed in Section 2. Section 3 describes an even more general formulation, as well as the proof of our theorems. Finally, Section 4 describes a generalization to the case of symmetric matrices A that is directly related to the work of Erwin Bolthausen [Bol09] .

It is important to mention that the algorithms (1.1) and (1.2) are completely different from gaussian belief propagation (BP). The gaussian assumption refers indeed to the distribution of the matrix entries, not to the variables to be inferred. More generally, none of the existing rigorous results for BP seem to be applicable here.

It is truly remarkable that density evolution (in its special incarnation, SE) holds for dense graphs. This upsets a very popular piece of wisdom: 'density evolution (and message passing) works *because* the graph is locally tree-like, and does not work on graphs with many short loops.'

1.1 Main result

We begin with some missing definitions for algorithm (1.1). We assume

$$y = Ax_0 + w\,, \tag{1.3}$$

with $w \in \mathbb{R}^n$ a vector with i.i.d. entries with mean 0 and variance σ^2. Further, let $\{\eta_t\}_{t\geq 0}$ be a sequence of scalar functions $\eta_t : \mathbb{R} \to \mathbb{R}$ almost everywhere differentiable with bounded derivative. Define the sequence of vectors $\{x^t\}_{t\geq 0}$, $x^t \in \mathbb{R}^N$, $\{z^t\}_{t\geq 0}$, $z^t \in \mathbb{R}^n$, through Eqs. (1.1).

Next, let us define formally state evolution. Given a probability distribution p_{X_0} and let $\tau_0^2 \equiv \sigma^2 + \mathbb{E}\{X_0^2\}/\delta$, and define recursively for $t \geq 0$,

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\{[\eta_t(X_0 + \tau_t Z) - X_0]^2\right\}\,, \tag{1.4}$$

with $X_0 \sim p_{X_0}$ and $Z \sim \mathsf{N}(0,1)$ independent from X_0. We will use the term *state evolution* to refer both to the recursion (1.4) (or its more general version introduced in Section 3.1) and to the sequence $\{\tau_t\}_{t\geq 0}$ that it defines.

Let us denote the empirical distribution[1] of a vector $x_0 \in \mathbb{R}^N$ by $\hat{p}_{x_0}$. Further, we say a function $\phi : \mathbb{R}^m \to \mathbb{R}$ is *pseudo-Lipschitz* of order k and denote it by $\phi \in \mathrm{PL}(k)$ if there exists a constant $L > 0$ such that, for all $x, y \in \mathbb{R}^m$:

$$|\phi(x) - \phi(y)| \leq L(1 + \|x\|^{k-1} + \|y\|^{k-1})\,\|x - y\|\,. \tag{1.5}$$

Notice that when $\phi \in \mathrm{PL}(k)$, the following two properties follow:

(i) There is a constant L' such that for all $x \in \mathbb{R}^m$, $|\phi(x)| \leq L'(1 + \|x\|^k)$.

(ii) ϕ is locally Lipschitz, that is for any $M > 0$ there exist a constant $L_M < \infty$ such that for all $x, y \in [-M, M]^m$,

$$|\phi(x) - \phi(y)| \leq L_M\|x - y\|.$$

Further, $L_M \leq C(1 + M^{k-1})$ for some constant C.

In the following we shall use generically L for Lipschitz constants entering bounds of this type. It is understood (and will not be mentioned explicitly) that the constant must be properly adjusted at various passages.

Theorem 1. *Let $\{A(N)\}_{N\geq 0}$ be a sequence of sensing matrices $A \in \mathbb{R}^{n\times N}$ indexed by N, with iid entries $A_{ij} \sim \mathsf{N}(0, 1/n)$, and assume $n/N \to \delta \in (0,\infty)$. Consider further a sequence of signals $\{x_0(N)\}_{N\geq 0}$, whose empirical distributions converge weakly to a probability measure p_{X_0} on $\mathbb{R}$ with bounded $(2k-2)^{th}$ moment, and assume $\mathbb{E}_{\hat{p}_{x_0(N)}}(X_0^{2k-2}) \to \mathbb{E}_{p_{X_0}}(X_0^{2k-2})$ as $N \to \infty$ for some $k \geq 2$. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of order k and all $t \geq 0$, almost surely*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi(x_i^{t+1}, x_{0,i}) = \mathbb{E}\Big[\psi\big(\eta_t(X_0 + \tau_t Z), X_0\big)\Big] \tag{1.6}$$

with $X_0 \sim p_{X_0}$ and $Z \sim \mathsf{N}(0,1)$ independent.

[1] The probability distribution that puts a point mass $1/N$ at each of the N entries of the vector.

Up to a trivial change of variables, this is a formalization of the findings of [DMM09] (cf. in particular Eqs. [7], [8] and Finding 2 in that paper).

As an immediate consequence of the above theorem we have the following *decoupling principle* implying that a typical (finite) subset of the coordinates of x^t are asymptotically independent.

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

Then for all $t > 0$, almost surely

$$\lim_{N\to\infty} \mathsf{E}\,\psi(x^t_{I(1)}, \dots, x^t_{I(\ell)}, x_{0,I(1)}, \dots, x_{0,I(\ell)}) = \mathbb{E}\Big\{\psi\big(\widehat{X}_1, \dots, \widehat{X}_\ell, X_{0,1}, \dots, X_{0,\ell}\big)\Big\}\,, \tag{1.7}$$

where $\widehat{X}_i \equiv \eta_{t-1}(X_{0,i} + \tau_{t-1} Z_i)$ for $X_{0,i} \sim p_{X_0}$ and $Z_i \sim \mathsf{N}(0,1)$, $i = 1, \dots, \ell$ mutually independent.

For the proof of this corollary we refer to Section 3.10.

2 Examples

In this section we discuss in greater detail some of the applications of Theorem 1 to specific problems. To be definite, it is convenient to keep in mind a specific observable for applying Theorem 1. If we choose the test function $\psi(x,y) = (x-y)^2$, We get

$$\lim_{N\to\infty} \frac{1}{N}\|x^t - x_0\|^2 = (\tau_t^2 - \sigma^2)\delta\,. \tag{2.1}$$

Therefore state evolution allows to predict the mean square error of the iterative algorithm (1.1). More generally, state evolution can be used to estimate ℓ_p distances for $p \leq k$ through

$$\lim_{N\to\infty} \frac{1}{N}\|x^t - x_0\|_p^p = \mathbb{E}\big\{[\eta_{t-1}(X_0 + \tau_{t-1}Z) - X_0]^p\big\}\,. \tag{2.2}$$

2.1 Linear estimation

As a warm-up example consider the case in which the *a priori* distribution of x_0 is gaussian, namely its entries are i.i.d. $\mathsf{N}(0, a^2)$. It is a consequence of state evolution that the optimal AMP algorithm makes use of linear scalar estimators

$$\eta_t(x) = \lambda_t\, x\,. \tag{2.3}$$

Obviously such functions are almost everywhere differentiable, with bounded derivative, for any λ_t finite. The AMP algorithm (1.1) becomes

$$\begin{aligned} x^{t+1} &= \lambda_t(A^* z^t + x^t), \\ z^t &= y - Ax^t + (\lambda_{t-1}/\delta)\, z^{t-1}\,. \end{aligned} \tag{2.4}$$

State evolution reads

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}(1-\lambda_t)^2 a^2 + \frac{1}{\delta}\lambda_t^2\tau_t^2\,. \tag{2.5}$$

Since $(A^* z_t + x^t)_i - x_{0,i}$ is asymptotically gaussian noise with mean 0 and variance τ_t, the optimal choice of λ_t is

$$\lambda_t = \frac{a^2}{a^2 + \tau_t^2} \, . \tag{2.6}$$

Notice that this also minimizes the right hand side of Eq. (2.5). Under this choice, the recursion (2.5) yields

$$\tau_{t+1} = \sigma^2 + \frac{1}{\delta} \frac{a^2 \tau_t^2}{a^2 + \tau_t^2} \, . \tag{2.7}$$

The right hand side is a concave function of τ_t^2, and is easy to show that $\tau_t \to \tau_\infty$ exponentially fast, where, for $c = (1-\delta)/\delta$.

$$\tau_\infty^2 = \frac{1}{2}\Big\{(\sigma^2 + ca^2) + \sqrt{(\sigma^2 + ca^2)^2 + 4\sigma^2 a^2}\Big\} \, . \tag{2.8}$$

The mean square error of the resulting algorithm is estimated via Eq. (2.1). In particular, under the optimal choice of λ_t, the latter converges to $(\tau_\infty^2 - \sigma^2)\delta$ with τ_∞ given as above.

The asymptotic mean square error of optimal (MMSE) linear estimation can be computed using random matrix theory [TH99, VS99]. Remarkably, this coincides with the value $(\tau_\infty^2 - \sigma^2)\delta$ predicted here.

2.2 Compressed sensing via soft thresholding

In this case the vector x_0 is ℓ sparse (i.e. it has at most ℓ nonvanishing entries). Assuming that the empirical distribution of x_0 converges to the probability measure p_{X_0}, it is also natural to assume $\ell/N \to \varepsilon$ as $N \to \infty$ with

$$\mathbb{P}\{X_0 \neq 0\} = \varepsilon \, . \tag{2.9}$$

(Indeed Theorem 1 accommodates for a more general behavior, since $\hat{p}_{x_0(N)}$ is only required to converge weakly.)

In [DMM09], the authors proposed an algorithm of the form (1.1) with $\eta_t(x) = \eta(x;\theta_t)$ a sequence of soft-threshold functions

$$\eta(x;\theta) = \begin{cases} (x-\theta) & \text{if } x > \theta, \\ 0 & \text{if } -\theta \le x \le \theta, \\ (x+\theta) & \text{if } x < -\theta. \end{cases} \tag{2.10}$$

This choice is optimal in minimax sense. The function $x \mapsto \eta(x;\theta)$ is non-linear but nevertheless almost everywhere differentiable with bounded derivative. Therefore Theorem 1 applies to this case, and allows to predict the asymptotic mean square error using Eqs. (1.4), (1.6).

2.3 Multi-User Detection

The model (1.3) is used to describe the input-output relation in code division multiple access (CDMA) channel. The matrix A contains the users signatures. The entries $x_{0,i}$ belong to the signal constellation used by the system. For the sake of simplicity, let us consider the case of antipodal signaling, i.e. $x_{0,i} \in \{+1, -1\}$ uniformly at random. Following [Kab03, NS05, MT06] we take

$$\eta_t(x) = \tanh\left\{x/\tau_t^2\right\} . \tag{2.11}$$

The rationale for this choice is that it gives the conditional expectation of a uniformly random signal $X_{0,i} \in \{+1, -1\}$, given the observation $X_{0,i} + \tau_t Z_i = x$ for $Z_i \sim \mathsf{N}(0, 1)$ gaussian noise. The algorithm (1.1) reads in this case

$$x^{t+1} = \tanh\left\{\frac{1}{\tau_t^2}(A^* z^t + x^t)\right\}, \tag{2.12}$$
$$z^t = y - Ax^t + \frac{z^{t-1}}{\delta\tau_t^2}\Big\{1 - \big\langle \tanh^2\big[(A^* z^t + x^t)/\tau_t^2\big]\big\rangle\Big\}.$$

State evolution yields

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big\{\big[\tanh\big(\tau_t^{-2} + \tau_t^{-1} X\big) - 1\big]^2\Big\}. \tag{2.13}$$

This state evolution recursion was proved in [MT06] for properly chosen sparse signature matrices A. Theorem 1 provides the first generalization to the more relevant case of dense signatures.

3 Proof

The proof is based on a conditioning technique developed by Erwin Bolthausen for the analysis of the so-called TAP equations in spin glass theory [Bol09]. Related ideas can also be found in [Don06].

First we introduce some new notations and state and prove a more general result than Theorem 1.

3.1 A general statement

We describe now a more general recursion than in (1.1). In the next section we show that the AMP algorithm (1.1) can be regarded as a special cases of the recursion defined here.

We will say that a function $F : \mathbb{R}^2 \to \mathbb{R}$ is *almost smooth* if it is continuously differentiable with bounded derivatives in a measurable domain $\mathcal{C}_F$ with the following property: for each $y \in \mathbb{R}$, the set $\mathcal{C}_F(y) \equiv \{x \in \mathbb{R} : (x, y) \notin \mathcal{C}_F\} \subseteq \mathbb{R}$ has zero Lebesgue measure. Notice that, in particular, F must be almost everywhere differentiable in $\mathbb{R}^2$.

The algorithm is defined by two sequences of function $\{f_t\}_{t\geq 0}$, $\{g_t\}_{t\geq 0}$, where for each $t \geq 0$, $f_t : \mathbb{R}^2 \to \mathbb{R}$ and $g_t : \mathbb{R}^2 \to \mathbb{R}$ are assumed to be *almost smooth*. As before, given $a, b \in \mathbb{R}^K$, we write $f_t(a, b)$ for the vector obtained by applying componentwise f_t to a, b. When b is clear from the context, we will just write, with an abuse of notation, $f_t(a)$. We will use analogous notations for g_t.

Given $w \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^N$, define the sequence of vectors $h^t, q^t \in \mathbb{R}^N$ and $z^t, m^t \in \mathbb{R}^n$, by fixing initial condition q^0, and defining b^t for $t \geq 0$, m^t for $t \geq 0$, h^t for $t \geq 1$, and q^t, for $t \geq 1$, through

$$\begin{aligned} h^{t+1} &= A^* m^t - \xi_t\, q^t, \qquad && m^t = g_t(b^t, w), \\ b^t &= A\, q^t - \lambda_t m^{t-1}, \qquad && q^t = f_t(h^t, x_0), \end{aligned} \tag{3.1}$$

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).

Assume that the limit $\sigma_0^2 = \delta^{-1}\lim_{n\to\infty}\langle q^0, q^0\rangle$ exists, for a sequence of initial conditions of increasing dimensions. State evolution defines the quantities τ_t^2, for $t \geq 0$, and σ_t^2, for $t \geq 1$, via

$$\tau_t^2 = \mathbb{E}\big\{g_t(\sigma_t Z, W)^2\big\}, \qquad \sigma_t^2 = \frac{1}{\delta}\,\mathbb{E}\big\{f_t(\tau_{t-1} Z, X_0)^2\big\}, \tag{3.2}$$

with $W \sim p_W$, $X_0 \sim p_{X_0}$ are independent of $Z \sim \mathsf{N}(0, 1)$. We have the following general result.

Theorem 2. *Let $\{A(N)\}_{N\geq 0}, \{q_0(N)\}_{N\geq 0}$ be, respectively, a sequence of matrices $A \in \mathbb{R}^{n\times N}$ indexed by N with iid entries $A_{ij} \sim \mathsf{N}(0, 1/n)$, and of initial conditions. Assume $n/N \to \delta \in (0,\infty)$. Consider 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) $\lim_{N\to\infty} \mathbb{E}_{\hat{p}_{q_0(N)}}(X^{2k-2}) < \infty$.

Then, for any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \to \mathbb{R}$ of order at most k and all $t \geq 0$, almost surely

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi(h_i^{t+1}, x_{0,i}) = \mathbb{E}\Big[\psi\big(\tau_t Z, X_0\big)\Big]\,, \tag{3.3}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(b_i^{t}, w_i) = \mathbb{E}\Big[\psi\big(\sigma_t Z, W\big)\Big]\,, \tag{3.4}$$

with $X_0 \sim p_{X_0}$, $W \sim p_W$, and $Z \sim \mathsf{N}(0,1)$ independent from X_0, W.

3.2 Example: AMP and Theorem 1

As already mentioned, the AMP algorithm (1.1) is a special case of the recursion (3.1). The reduction is obtained by defining

$$h^{t+1} = x_0 - (A^* z^t + x^t)\,, \tag{3.5}$$
$$q^t = x^t - x_0\,, \tag{3.6}$$
$$b^t = w - z^t\,, \tag{3.7}$$
$$m^t = -z^t\,. \tag{3.8}$$

The functions f_t and g_t are given by

$$f_t(s, x_0) = \eta_{t-1}(x_0 - s) - x_0\,, \qquad g_t(s, w) = s - w\,. \tag{3.9}$$

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

Note 1. *(a) Although the recursions (3.1) and (1.1) are equivalent mathematically, only the latter can be used as an algorithm. Indeed the recursion (3.1) tracks the difference of the current estimates x^t from x_0, and is initialized using x^0 itself. The recursion (3.1) is only relevant for mathematical analysis.*

(b) Due to symmetry, for each t, all coordinates of the vector h^t have the same distribution (similarly for b^t, q^t and m^t).

3.3 Proof of Theorem 1

First note that (3.2) reduces to

$$\tau_t^2 = \sigma^2 + \sigma_t^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big[\Big(\eta_{t-1}(X_0 + \tau_{t-1}Z) - X_0\Big)^2\Big],$$

with $\tau_0^2 = \sigma^2 + \delta^{-1}\mathbb{E}(X_0^2)$. This follows from

$$\sigma_0^2 = \delta^{-1} \lim_{N\to\infty} \langle q^0, q^0\rangle = \delta^{-1}\mathbb{E}_{p_{X_0}}(X_0^2)$$

or $\tau_0^2 = \sigma^2 + \delta^{-1}\mathbb{E}(X_0^2)$. Also, by definition $x^{t+1} = \eta_t(A^* b^t + x^t) = \eta_t(x_0 - h^{t+1})$. Therefore, applying Theorem 2 to the function $(h_i^t, x_{0,i}) \mapsto \psi(\eta_{t-1}(x_{0,i} - h_i^t), x_{0,i})$ we obtain

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi(x_i^t, x_{0,i}) \overset{\text{a.s.}}{=} \mathbb{E}\left\{\psi\big(\eta_{t-1}(X_0 - \tau_{t-1}Z), X_0\big)\right\}$$

with $Z \sim \mathsf{N}(0,1)$ independent of X_0, which yields the claim as $Z \overset{\text{d}}{=} -Z$. Note that since η has bounded derivative, $(h_i^t, x_{0,i}) \mapsto \psi(\eta_{t-1}(x_{0,i} - h_i^t), x_{0,i})$ is also pseudo-Lipschitz of order at most k.

3.4 Definitions

When the update equation for h^{t+1} in (3.1) is used, all values of $b^0, \dots, b^t$, $m^0, \dots, m^t$, $h^1, \dots, h^t$ and $q^0, \dots, q^t$ have been previously calculated. Hence, we can consider the distribution of h^{t+1} when it is conditioned on all these known variables and x_0, w. In particular, define $\mathfrak{S}_{t_1,t_2}$ to be the σ-algebra generated by $b^0, \dots, b^{t_1-1}$, $m^0, \dots, m^{t_1-1}$, $h^1, \dots, h^{t_2}$, and $q^0, \dots, q^{t_2}$. The basic idea of the proof is to compute the conditional distribution $b^t|_{\mathfrak{S}_{t,t}}$ and $h^{t+1}|_{\mathfrak{S}_{t+1,t}}$. This is done by characterizing the conditional distribution of the matrix A given this filtration.

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:

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

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

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

We also introduce the notation $m_\parallel^t$ for the projection of m^t onto the column space of M_t and define $m_\perp^t = m^t - m_\parallel^t$. Similarly, define $q_\parallel^t, q_\perp^t$ to be the parallel and orthogonal projections of q^t onto column space of Q_t.

Recall that D^* denote the transpose of the matrix D. For vectors $u, v \in \mathbb{R}^m$ we define $\langle u\rangle = \sum_{i=1}^m u_i/m$. We will also often use the scalar product $\langle u, v\rangle = \sum_{i=1}^m u_i v_i/m$.

Given two random variable X, Y, and a σ-algebra $\mathfrak{S}$, the notations $X|_{\mathfrak{S}} \overset{\text{d}}{=} Y$ means that for any integrable function ϕ and for any random variable Z measurable on $\mathfrak{S}$, $\mathbb{E}\{\phi(X)Z\} = \mathbb{E}\{\phi(Y)Z\}$. In word we will say that X is distributed as (or is equal in distribution to) Y *conditional on* $\mathfrak{S}$. In case $\mathfrak{S}$ is the trivial σ algebra we simply write $X \overset{\text{d}}{=} Y$ (i.e. X and Y are equal in distribution). For random variables X, Y the notation $X \overset{\text{a.s.}}{=} Y$ means that X and Y are equal almost surely.

Finally the large system limit will be denoted either as $\lim_{N\to\infty}$ or as $n \to \infty$. It is understood that either of the two dimensions can index the sequence of problems under considerations, and that $n/N \to \delta$.

3.5 Main technical Lemma

We prove the following more general result.

Lemma 1. *Let, $\{A(N)\}$, $\{q_0(N)\}_N$, $\{x_0(N)\}_N$ and $\{w(N)\}_N$ be sequences as in Theorem 2, with $n/N \to \delta \in (0,\infty)$ and let $\{\sigma_t, \tau_t\}_{t\geq 0}$ be the sequence of satisfying* (3.2)*. Then the following hold for all $t \in \mathbb{N}\cup\{0\}$*

(*a*)

$$h^{t+1}|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1} \alpha_i h^{i+1} + \tilde{A}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(1)\,, \tag{3.10}$$

$$b^t|_{\mathfrak{S}_{t,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1} \beta_i b^i + \tilde{A} q_\perp^t + \tilde{M}_t\vec{o}_t(1)\,, \tag{3.11}$$

where $\tilde{A}$ is an independent copy of A and coefficients α_i, β_j satisfy $m_\parallel^t = \sum_{i=0}^{t-1}\alpha_i m^i$ and $q_\parallel^t = \sum_{i=0}^{t-1}\beta_i q^i$. 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 $\tilde{Q}_t^*\tilde{Q}_t = N\,\mathbf{I}_t$ ($\tilde{M}_t^*\tilde{M}_t = n\,\mathbf{I}_t$). Here $\vec{o}_t(1) \in \mathbb{R}^t$ is a finite dimensional random vector that converges to 0 almost surely as $N\to\infty$.

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

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi_h(h_i^1,\ldots,h_i^{t+1},x_{0,i}) \stackrel{\text{a.s.}}{=} \mathbb{E}\big[\phi_h(\tau_0 Z_0,\ldots,\tau_t Z_t, X_0)\big]\,, \tag{3.12}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \phi_b(b_i^0,\ldots,b_i^t,w_i) \stackrel{\text{a.s.}}{=} \mathbb{E}\big[\phi_b(\sigma_0 \hat{Z}_0,\ldots,\sigma_t \hat{Z}_t, W)\big]\,, \tag{3.13}$$

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 non-random:*

$$\lim_{N\to\infty}\langle h^{r+1}, h^{s+1}\rangle \stackrel{\text{a.s.}}{=} \lim_{n\to\infty}\langle m^r, m^s\rangle\,, \tag{3.14}$$

$$\lim_{n\to\infty}\langle b^r, b^s\rangle \stackrel{\text{a.s.}}{=} \frac{1}{\delta}\lim_{N\to\infty}\langle q^r, q^s\rangle\,. \tag{3.15}$$

(*d*) *For all $0\leq r,s\leq t$, and for any function $\varphi:\mathbb{R}^2\to\mathbb{R}$ almost everywhere continuously differentiable with bounded derivative, the following equations hold and all limits exist, are bounded and non-random.*

$$\lim_{N\to\infty}\langle h^{r+1}, \varphi(h^{s+1},x_0)\rangle \stackrel{\text{a.s.}}{=} \lim_{N\to\infty}\langle h^{r+1}, h^{s+1}\rangle\langle\varphi'(h^{s+1},x_0)\rangle, \tag{3.16}$$

$$\lim_{n\to\infty}\langle b^r, \varphi(b^s,w)\rangle \stackrel{\text{a.s.}}{=} \lim_{n\to\infty}\langle b^r, b^s\rangle\langle\varphi'(b^s,w)\rangle \tag{3.17}$$

where φ' represents derivative with respect to the first coordinate.

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

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} (h_i^{t+1})^{2\ell} < \infty\,, \tag{3.18}$$

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (b_i^{t})^{2\ell} < \infty. \tag{3.19}$$

Note 2. (a) *In the following we will use the notations $\vec{\alpha} = (\alpha_0, \dots, \alpha_{t-1})$ and $\vec{\beta} = (\beta_0, \dots, \beta_{t-1})$*

(b) *Equations (3.16) and (3.17) have the form of Stein's lemma [Ste72] (Lemma 3 in our Section 3.7).*

Proof of Theorem 2. The result follows from applying Lemma 1(b) to the functions $\phi_h(y_0, \dots, y_t, x_{0,i}) = \psi(y_t, x_{0,i})$ and $\phi_b(y_0, \dots, y_t, w_i) = \psi(y_t, w_i)$. □

3.6 Useful probability facts

Before embarking in the actual proof, it is convenient to summarize a few facts that will be used repeatedly.

We will use the following strong law of large numbers (SLLN) which follows from [HT97, Theorem 2.1].

Theorem 3 (SLLN, [HT97])**.** *Let $\{X_{n,i} : 1 \le i \le n, n \ge 1\}$ be a triangular array of random variables with $(X_{n,1}, \dots, X_{n,n})$ mutually independent with mean equal zero for each n and $\mathbb{E}|X_{n,i}|^{2+\kappa} \le C$ for some $\kappa > 0$, $C < \infty$. Then $\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \to 0$ a.s. for $n \to \infty$.*

Next, we present a standard property of Gaussian matrices without proof.

Lemma 2. *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) *$v^*\tilde{A}u \stackrel{\mathrm{d}}{=} \frac{Z}{\sqrt{n}}$ where $Z \sim \mathsf{N}(0,1)$.*

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

(c) *Consider, for each $n \ge d$, 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 to W. Then for $D = [w_1|\dots|w_d]$, we have $P_W Au \stackrel{\mathrm{d}}{=} Dx$ with $x \in \mathbb{R}^d$ that satisfies: $\lim_{n\to\infty} \|x\| \stackrel{\text{a.s.}}{=} 0$ (the limit being taken with d fixed). In this paper, we denote such vector x by $\vec{o}_d(1)$ as well.*

Lemma 3 (Stein's Lemma [Ste72])**.** *For jointly gaussian random variables Z_1, Z_2 and any function $\varphi : \mathbb{R} \to \mathbb{R}$ where $\mathbb{E}[\varphi'(Z_1)]$ and $\mathbb{E}[Z_1\varphi(Z_2)]$ exists, the following holds*

$$\mathbb{E}[Z_1\varphi(Z_2)] = \mathrm{Cov}(Z_1, Z_2)\mathbb{E}[\varphi'(Z_2)].$$

We will apply the following law of large numbers to the sequence $\{x_0(N), w(N)\}_N$. Its proof can be found in Appendix A.1.

Lemma 4. *Let $k \geq 2$ and consider sequence of vectors $\{v(N)\}_{N\geq 0}$, whose empirical distribution converges weakly to probability measure p_V on $\mathbb{R}$ with bounded k^{th} moment, and assume $\mathbb{E}_{\hat{p}_{v(N)}}(V^k) \to \mathbb{E}_{p_V}(V^k)$ as $N \to \infty$. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order at most k, almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi(v_i) = \mathbb{E}\big[\psi(V)\big] \tag{3.20}$$

Finally, a lemma on almost smooth functions (whose proof is in Appendix A.2).

Lemma 5. *Let $F : \mathbb{R}^2 \to \mathbb{R}$ be almost smooth and denote by $F'(x, y)$ its derivative with respect to the first argument at $(x, y) \in \mathbb{R}^2$. Assume (X_n, Y_n) be a sequence of random vectors in $\mathbb{R}^2$ converging in distribution to the random vector (X, Y) as $n \to \infty$. Assume further that X, Y are independent and that the distribution of X is absolutely continuous with respect to the Lebesgue measure. Then*

$$\lim_{n\to\infty} \mathbb{E}\{F'(X_n, Y_n)\} = \mathbb{E}\{F'(X, Y)\}\,. \tag{3.21}$$

3.7 Conditional distributions

In order to calculate $b^t|_{\mathfrak{S}_{t,t}}, h^{t+1}|_{\mathfrak{S}_{t+1,t}}$ we will characterize the conditional distributions $A|_{\mathfrak{S}_{t,t}}, A|_{\mathfrak{S}_{t+1,t}}$.

Lemma 6. *Let $(t_1, t_2) = (t, t)$ or $(t_1, t_2) = (t+1, t)$. Then the conditional distribution of the random matrix A given the σ-algebra $\mathfrak{S}_{t_1,t_2}$, satisfies*

$$A|_{\mathfrak{S}_{t_1,t_2}} \stackrel{\mathrm{d}}{=} E_{t_1,t_2} + \mathcal{P}_{t_1,t_2}(\tilde{A}). \tag{3.22}$$

Here $\tilde{A} \stackrel{\mathrm{d}}{=} A$ is a random matrix independent of $\mathfrak{S}_{t_1,t_2}$ and $E_{t_1,t_2} = \mathbb{E}(A|\mathfrak{S}_{t_1,t_2})$ is given by

$$E_{t_1,t_2} = Y_{t_1}(Q^*_{t_1}Q_{t_1})^{-1}Q^*_{t_1} + M_{t_2}(M^*_{t_2}M_{t_2})^{-1}X^*_{t_2} - M_{t_2}(M^*_{t_2}M_{t_2})^{-1}M^*_{t_2}Y_{t_1}(Q^*_{t_1}Q_{t_1})^{-1}Q^*_{t_1}\,. \tag{3.23}$$

Further, $\mathcal{P}_{t_1,t_2}$ is the orthogonal projector onto subspace $\mathrm{V}_{t_1,t_2} = \{A|AQ_{t_1} = 0, A^*M_{t_2} = 0\}$, defined by

$$\mathcal{P}_{t_1,t_2}(\tilde{A}) = P^{\perp}_{M_{t_2}}\tilde{A}P^{\perp}_{Q_{t_1}}\,.$$

Here $P^{\perp}_{M_{t_2}} = I - P_{M_{t_2}}$, $P^{\perp}_{Q_{t_1}} = I - P_{Q_{t_1}}$, and $P_{Q_{t_1}}$, $P_{M_{t_2}}$ are orthogonal projector onto column spaces of Q_{t_1} and M_{t_2} respectively.

Recall the following well-known formula.

Lemma 7. *Let $z \in \mathbb{R}^n$ be a random vector of iid $\mathsf{N}(0, \alpha)$ variables and let $\mathrm{D} \in \mathbb{R}^{m\times n}$ be a linear operator. Then for any constant vector $b \in \mathbb{R}^m$ the distribution of z conditioned on $\mathrm{D}z = b$ satisfies:*

$$z|_{\mathrm{D}z=b} \stackrel{\mathrm{d}}{=} \mathrm{D}^*(\mathrm{D}\mathrm{D}^*)^{-1}b + \mathrm{P}_{\{\mathrm{D}z=0\}}(\tilde{z})$$

where $\mathrm{P}_{\{\mathrm{D}z=0\}}$ is the orthogonal projection onto the subspace $\{\mathrm{D}z = 0\}$ and $\tilde{z}$ is a random vector of iid $\mathsf{N}(0, \alpha)$. Moreover, $\mathrm{D}^*(\mathrm{D}\mathrm{D}^*)^{-1}b = \arg\min_z \big\{\|z\|^2 | \mathrm{D}z = b\big\}$.

Proof. The result is trivial if $\mathrm{D} = [I_{m\times m}|0_{m\times(n-m)}]$ (with $I_{m\times m} \in \mathbb{R}^{m\times m}$ is the identity matrix). For general D, it follows by invariance of the gaussian distribution under rotations. Finally, using a least square calculation, it is simple to see that $\mathrm{D}^*(\mathrm{D}\mathrm{D}^*)^{-1}b = \arg\min_z\{\|z\|^2|\mathrm{D}z = b\}$. □

Lemma 6 follows from applying Lemma 7 to the operator D that maps A to (AQ, M^*A). Note that we can assume, without loss of generality f, g to be non-constant as a function of their first argument. If this is the case, it is easy to see that, for finite values of t, the matrices $M_t^*M_t$ and $Q_t^*Q_t$ are non-singular almost surely, and hence the above expressions are well defined. Proof of Lemma 6 appears in Section 3.8.

Lemma 8. *The following holds*

$$E_{t+1,t}^* m^t \overset{\text{a.s.}}{=} X_t(M_t^*M_t)^{-1}M_t^*m_{\|}^t + Q_{t+1}(Q_{t+1}^*Q_{t+1})^{-1}Y_{t+1}^*m_{\perp}^t, \tag{3.24}$$

$$E_{t,t}q^t \overset{\text{a.s.}}{=} Y_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}X_t^*q_{\perp}^t. \tag{3.25}$$

Proof. Write $m^t = m_{\|}^t + m_{\perp}^t$. Using (3.23) and the fact that $M_t^*m_{\perp}^t = 0$, we obtain

$$E_{t+1,t}^* m_{\perp}^t = Q_{t+1}(Q_{t+1}^*Q_{t+1})^{-1}Y_{t+1}^*m_{\perp}^t.$$

On the other hand let $m_{\|}^t = \sum_{i=0}^{t-1}\alpha_i m^i = M_t\vec{\alpha}$. Then using $A^*M_t = X_t$, (3.22), and $[\mathcal{P}_{t+1,t}(\tilde{A})]^*m_{\|}^t = 0$ we have, *conditionally on* $\mathfrak{S}_{t+1,t}$,

$$E_{t+1,t}^*m_{\|}^t \overset{\text{d}}{=} A^*m_{\|}^t \overset{\text{d}}{=} A^*M_t\vec{\alpha} \overset{\text{d}}{=} X_t\vec{\alpha} \overset{\text{d}}{=} X_t(M_t^*M_t)^{-1}M_t^*M_t\vec{\alpha} \overset{\text{d}}{=} X_t(M_t^*M_t)^{-1}M_t^*m_{\|}^t.$$

Since the first and last term are measurable on $\mathfrak{S}_{t+1,t}$, they must be equal almost surely, and Eq. (3.24) follows.

Similarly, use $q^t = q_{\|}^t + q_{\perp}^t$, $q_{\|}^t = Q_t\vec{\beta}$ and $Q_t^*q_{\perp}^t = 0$ to obtain (3.25). □

3.8 Proof of Lemma 6

Conditioning on $\mathfrak{S}_{t_1,t_2}$ is equivalent to conditioning on the linear constraints $AQ_{t_1} = Y_{t_1}$ and $A^*M_{t_2} = X_{t_2}$. To simplify the notation we will drop all sub-indices t_1, t_2. The expression (3.23) for the conditional expectation $E = \mathbb{E}(A|\mathfrak{S}_{t_1,t_2})$ follows from Lemma 7 and the following calculation for

$$E = \arg\min_A \left\{ \|A\|_F^2 \middle| AQ = Y, A^*M = X \right\}.$$

We use Lagrange multipliers' method to obtain this minimum. Consider the Lagrangian

$$L(A, \Theta, \Gamma) = \|A\|_F^2 + \big(\Theta, (Y - AQ)\big) + \big(\Gamma, (X - A^*M)\big),$$

with $\Theta \in \mathbb{R}^{n\times t_1}$, $\Gamma \in \mathbb{R}^{N\times t_2}$ and $(A, B) \equiv \mathrm{Tr}(AB^*)$ the usual scalar product among matrices. Imposing the stationarity conditions yields

$$2A = \Theta Q^* + M\Gamma^* \tag{3.26}$$

Equation (3.26) does not have a unique solution for the parameters Θ and Γ. In fact if Θ_0, Γ_0 are a solution then for any $t_2 \times t_1$ matrix R the new parameters $\Theta_R = \Theta_0 + MR$ and $\Gamma_R = \Gamma_0 - QR^*$ satisfy $\Theta_RQ^* + M\Gamma_R^* = \Theta_0Q^* + M\Gamma_0^* = 2A$. In particular for $R_1 = \Gamma_0^*Q(Q^*Q)^{-1}$ we have $Q^*\Gamma_{R_1} = 0$. Multiplying (3.26) by Q from right (using $\Theta_{R_1}, \Gamma_{R_1}$) we have $2Y = \Theta_{R_1}Q^*Q$ or $\Theta_{R_1} = 2Y(Q^*Q)^{-1}$. Now multiplying (3.26) by M^* from left we obtain $2X^* = 2M^*Y(Q^*Q)^{-1}Q^* + M^*M\Gamma_{R_1}^*$ which leads to $\Gamma_{R_1}^* = 2(M^*M)^{-1}\big[X^* - M^*Y(Q^*Q)^{-1}Q^*\big]$. From these we see that $E = \mathbb{E}(A|\mathfrak{S}_{t_1,t_2})$ satisfies

$$E = Y(Q^*Q)^{-1}Q^* + M(M^*M)^{-1}\big[X - M^*Y(Q^*Q)^{-1}Q^*\big].$$

Now we are left to prove $\mathcal{P}_{t_1,t_2}(\tilde{A}) = P_M^{\perp}\tilde{A}P_Q^{\perp}$. We need to show that the operator $\mathcal{F} : A \mapsto P_M^{\perp}AP_Q^{\perp}$ satisfies

(a) $\mathcal{F} \circ \mathcal{F} = \mathcal{F}$.

(b) $\mathcal{F}(A) \in V = \{A | AQ_{t_1} = 0, A^* M_{t_2} = 0\}$.

(c) $\mathcal{F}(A) = A$ for $A \in V$

(d) $\mathcal{F}$ is symmetric. That is for all matrices A, B: $(\mathcal{F}(A), B) = (A, \mathcal{F}(B))$.

(a) is correct since
$$\mathcal{F} \circ \mathcal{F}(A) = P_M^\perp P_M^\perp A P_Q^\perp P_Q^\perp = P_M^\perp A P_Q^\perp .$$
(b) is correct since by definition of $\mathcal{F}(A)Q = P_M^\perp A P_Q^\perp Q = 0$ and similarly $\mathcal{F}(A)^* M = 0$.
(c) follows because
$$\mathcal{F}(A) = A - P_M A - A P_Q + P_M A P_Q,$$
and each of the last three term vanishes either because $AQ = 0$ or because $A^* M = 0$.
(d) is correct because

$$\begin{aligned}(\mathcal{F}(A), B) &= \mathrm{Tr}\left(P_M^\perp A P_Q^\perp B^*\right) \\ &= \mathrm{Tr}\left(A P_Q^\perp B^* P_M^\perp\right) = (A, \mathcal{F}(B)) .\end{aligned}$$

3.9 Proof of Lemma 1

The proof is by induction on t. Let $\mathcal{H}_{t+1}$ be the property that (3.10), (3.12), (3.14), (3.16), and (3.18) are correct. Similarly, let $\mathcal{B}_t$ be the property that (3.11), (3.13), (3.15), (3.17), and (3.18) hold. The inductive proof consists of the following four main steps.

1. $\mathcal{B}_0$ holds.
2. $\mathcal{H}_1$ holds.
3. If $\mathcal{B}_r$, $\mathcal{H}_s$ hold for all $r < t$ and $s \le t$ then $\mathcal{B}_t$ holds.
4. If $\mathcal{B}_r$, $\mathcal{H}_s$ hold for all $r \le t$ and $s \le t$ then $\mathcal{H}_{t+1}$ holds.

For each of these steps we will have to prove five properties that we will denote by (a), (b), (c), (d), (e) according to their appearance in Lemma 1.

3.9.1 Step 1: $\mathcal{B}_0$

Note that $b^0 = Aq^0$.

(a) $\mathfrak{S}_{0,0}$ is the trivial σ-algebra. Also $q^0 = q^0_\perp$ since Q_0 is an empty matrix. Hence

$$b^0|_{\mathfrak{S}_{0,0}} = A q^0_\perp .$$

(b) Let $\phi_b : \mathbb{R}^2 \to \mathbb{R}$ be a pseudo-Lipschitz function of order at most k. Hence, $|\phi_b(x)| \le L(1 + \|x\|^k)$. Given q^0, w, the random variable $\sum_{i=1}^n \phi_b([Aq^0]_i, w_i)/n$ is a sum of independent random variables. By Lemma 2 $[Aq^0]_i \stackrel{\mathrm{d}}{=} Z\|q^0\|/\sqrt{n}$ for $Z \sim \mathsf{N}(0,1)$. And using

$$\lim_{n\to\infty} \langle q^0, q^0 \rangle = \delta \sigma_0^2 < \infty,$$

for all $p \geq 2$ there exist a constant C_p such that $\mathbb{E}|(Aq_0)_i|^p = \langle q^0, q^0 \rangle^{p/2} \mathbb{E}|Z|^p < C_p$. Therefore,

$$\mathbb{E}\big\{ |\phi_b\left([Aq^0]_i, w_i\right)|^3 \big\} \leq L^3 \big[1 + \mathbb{E}\{|(Aq^0)_i|^{3k}\} + |w_i|^{3k} \big] \leq C$$

for a constant C. Now, we can apply Theorem 3 to get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [\phi_b(b_i^0, w_i) - \mathbb{E}_A \phi_b(b_i^0, w_i)] \overset{\text{a.s.}}{=} 0.$$

Hence, using Lemma 4 for $v = w$ and for $\psi(w_i) = \mathbb{E}_Z[\phi_b(\|q^0\|Z/\sqrt{n}, w_i)]$ we get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_A[\phi_b(b_i^0, w_i)] \overset{\text{a.s.}}{=} \mathbb{E}\big[\phi_b(\sigma_0 Z, W)\big].$$

(c) Using Lemma 2,

$$\lim_{n\to\infty} \langle b^0, b^0 \rangle = \lim_{n\to\infty} \frac{\|Aq^0\|^2}{n} \overset{\text{a.s.}}{=} \lim_{N\to\infty} \frac{\langle q^0, q^0 \rangle}{\delta} = \sigma_0^2.$$

(d) Using $\mathcal{B}_0(b)$, and $\phi(x, w_i) = x\varphi(x, w_i)$ we obtain $\lim_{n\to\infty} \langle b^0, \varphi(b^0, w)\rangle \overset{\text{a.s.}}{=} \mathbb{E}(\sigma_0 \hat{Z} \varphi(\sigma_0 \hat{Z}, W))$, which is equal to $\sigma_0^2 \mathbb{E}[\varphi'(\sigma_0 \hat{Z}, W)]$ using Lemma 3. On the other hand, in proof of (b) we showed $\lim_{n\to\infty} \langle b^0, b^0 \rangle \overset{\text{a.s.}}{=} \sigma_0^2$.

By part (b), the empirical distribution of (b^0, w) (i.e. the probability distribution on $\mathbb{R}^2$ that puts mass $1/n$ on each point (b_i^0, w_i), $i \in [n]$) converges weakly to the distribution of $(\sigma_0 \hat{Z}, W)$. Using Lemma 5, we get $\lim_{n\to\infty} \langle \varphi'(b^0, w) \rangle \overset{\text{a.s.}}{=} \mathbb{E}[\varphi'(\sigma_0 \hat{Z}, W)]$.

(e) Similar to (b), conditioning on q^0, the term $\sum_{i=1}^{n} ([Aq^0]_i)^{2\ell}/n$ is sum of independent gaussians and $\mathbb{E}\{|[Aq^0]_i|^p\} = \langle q^0, q^0 \rangle^{p/2} \mathbb{E}(Z^p) < C$ for a constant C. Therefore, by Theorem 3, we get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big[([Aq^0]_i)^{2\ell} - \mathbb{E}_A\{([Aq^0]_i)^{2\ell}\} \Big] \overset{\text{a.s.}}{=} 0.$$

But, $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_A\{([Aq^0]_i)^{2\ell}\} = \langle q^0, q^0 \rangle^{\ell} \mathbb{E}_Z(Z^{2\ell}) < \infty$.

3.9.2 Step 2: $\mathcal{H}_1$

Note that $h^1 = A^* m^0 - \xi_0 q^0$.

(a) $\mathfrak{S}_{1,0}$ is generated by b^0 and m^0. Also $m^0 = m_\perp^0$ since M_0 is an empty matrix. Applying Lemma 6 we have

$$A|_{\mathfrak{S}_{1,0}} \overset{\text{d}}{=} b^0 \|q^0\|^{-2} (q^0)^* + \tilde{A} P_{q_0}^{\perp}.$$

Hence

$$h^1|_{\mathfrak{S}_{1,0}} \overset{\text{d}}{=} P_{q_0}^{\perp} \tilde{A}^* m^0 + \delta \frac{\langle b_0, m_0 \rangle}{\langle q_0, q_0 \rangle} q^0 - \xi_0 q^0.$$

But using $\mathcal{B}_0(d)$ we have

$$\lim_{n\to\infty} \langle b^0, m^0 \rangle = \lim_{n\to\infty} \langle b^0, g_0(b^0, w) \rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \langle b^0, b^0 \rangle \langle g_0'(b^0, w) \rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \xi_0 \frac{\langle q^0, q^0 \rangle}{\delta}.$$

Further by $\mathcal{B}_0(b)$, applied to the function $\phi_b(x,w) = g_0(x,w)^2$ we obtain

$$\lim_{n\to\infty} \langle m^0, m^0 \rangle \overset{\text{a.s.}}{=} \mathbb{E}[g_0(\sigma_0 Z, W)^2] = \tau_0^2. \tag{3.27}$$

Therefore

$$P_{q_0}^{\perp}\tilde{A}^* m^0 = \tilde{A}^* m^0 - P_{q_0}\tilde{A}^* m^0 = \tilde{A}^* m^0 + \vec{o}_1(1)\tilde{q}^0\,,$$

where the last estimate follows from Lemma 2(c) and (3.27). Hence,

$$h^1|_{\mathfrak{S}_{1,0}} \overset{\text{d}}{=} \tilde{A}^* m^0 + \vec{o}_1(1) q^0.$$

(c) Using Lemma 2 and Eq. (3.9.2), we get

$$\lim_{N\to\infty} \langle h^1, h^1 \rangle|_{\mathfrak{S}_{1,0}} = \lim_{N\to\infty} \frac{\|\tilde{A}^* m_0\|^2}{N} \overset{\text{a.s.}}{=} \lim_{N\to\infty} \langle m^0, m^0 \rangle \overset{\text{a.s.}}{=} \tau_0^2.$$

(e) First note that, conditioning on $\mathfrak{S}_{1,0}$,

$$\frac{1}{N}\sum_{i=1}^{N}(h_i^1)^{2\ell} = \frac{1}{N}\sum_{i=1}^{N}([\tilde{A}^* m^0]_i + \vec{o}_1(1) q_i^0)^{2\ell} \leq \frac{2^{2\ell}}{2}\frac{1}{N}\sum_{i=1}^{N}\left\{([\tilde{A}^* m^0]_i)^{2\ell} + \vec{o}_1(1)(q_i^0)^{2\ell}\right\}.$$

By assumption, $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(q_i^0)^{2\ell} < \infty$ and finiteness of $\frac{1}{N}\sum_{i=1}^{N}([\tilde{A}^* m^0]_i)^{2\ell}$ can be established similar to $\mathcal{B}_0(e)$ for the sum of functions of independent gaussians $\sum_{i=1}^{N}([\tilde{A}^* m^0]_i)^{2\ell}/N$.

(b) This proof uses again Eq. (3.9.2) and is very similar to proof of $\mathcal{B}_0(b)$. First we should remove the error term $\vec{o}_1(1)q^0$. In other words we need to show

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}[\phi_h([\tilde{A}^* m^0]_i + \vec{o}_1(1)q_i^0, x_{0,i}) - \phi_h([\tilde{A}^* m^0]_i, x_{0,i})] \overset{\text{a.s.}}{=} 0.$$

To simplify the notation let $a_i = ([\tilde{A}^* m^0]_i + \vec{o}_1(1)q_i^0, x_{0,i})$ and $c_i = ([\tilde{A}^* m^0]_i, x_{0,i})$. Now, using the pseudo-Lipschitz property of ϕ_h:

$$|\phi_h(a_i) - \phi_h(c_i)| \leq L\{1 + \max(\|a_i\|^{k-1}, \|c_i\|^{k-1})\}\,|q_i^0|\vec{o}_1(1).$$

Using Cauchy-Schwartz inequality,

$$\frac{1}{N}\sum_{i=1}^{N}|\phi_h(a_i) - \phi_h(c_i)| \leq L\max\left(\frac{1}{N}\sum_{i=1}^{N}\|a_i\|^{2k-2}, \frac{1}{N}\sum_{i=1}^{N}\|c_i\|^{2k-2}\right)^{1/2}\langle q^0, q^0\rangle^{1/2}\, o_N(1).$$

Hence, we only need to show $\frac{1}{N}\sum_{i=1}^{N}\|a_i\|^{2k-2} < \infty$ and $\frac{1}{N}\sum_{i=1}^{N}\|c_i\|^{2k-2} < \infty$ as $N\to\infty$. But

$$\frac{1}{N}\sum_{i=1}^{N}\|a_i\|^{2k-2} = O(\frac{1}{N}\sum_{i=1}^{N}|h_i^1|^{2k-2} + \frac{1}{N}\sum_{i=1}^{N}|x_{0,i}|^{2k-2})$$

which are bounded by part (e) and the original assumption on x_0. Similar argument plus the fact that $\frac{1}{N}\sum_{i=1}^{N}\|q_i^0\|^{2k-2} < \infty$ yield $\frac{1}{N}\sum_{i=1}^{N}\|c_i\|^{2k-2} < \infty$.

Thus, from here we consider $\tilde{h}^1|_{\mathfrak{S}_{1,0}} \stackrel{\mathrm{d}}{=} \tilde{A}^* m^0$ and will follow the steps taken in $\mathcal{B}_0(b)$. Conditioning on $\mathfrak{S}_{1,0}$, we can apply Theorem 3 to get

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} [\phi_h(\tilde{h}_i^1, x_{0,i}) - \mathbb{E}_{\tilde{A}} \phi_h(\tilde{h}_i^1, x_{0,i})] \stackrel{\text{a.s.}}{=} 0,$$

and use Lemma 4 for $\psi(v_i) = \mathbb{E}_{\tilde{A}} \phi_h(\tilde{h}_i^1, v_i)]$, obtaining

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{n} \mathbb{E}_{\tilde{A}}[\phi_h(h_i^1, x_{0,i})] = \lim_{N\to\infty} \mathbb{E}_Z[\phi_h(\frac{\|m^0\|}{\sqrt{n}} Z, X_0)] \stackrel{\text{a.s.}}{=} \mathbb{E}\big[\phi_h(\tau_0 Z, X_0)\big].$$

The last equality used $\mathcal{B}_0(c)$ and definition of τ_0.

(d) Using $\mathcal{H}_1(b)$ for $\phi_h(x, x_{0,i}) = x\varphi(x, x_{0,i})$ we obtain $\lim_{N\to\infty}\langle h^1, \varphi(h^1, x_0)\rangle \stackrel{\text{a.s.}}{=} \mathbb{E}(\tau_0 Z \varphi(\tau_0 Z, X_0))$, which is equal to $\tau_0^2 \mathbb{E}[\varphi'(\tau_0 Z, X_0)]$ using Lemma 3. On the other hand, in proof of (b) we showed $\lim_{N\to\infty}\langle h^1, h^1\rangle \stackrel{\text{a.s.}}{=} \tau_0^2$.

By part (b) the empirical distribution of (h^1, x_0) (i.e. the probability distribution on $\mathbb{R}^2$ that puts mass $1/B$ on each point $(h_i^1, x_{0,i})$, $i \in [N]$) converges weakly to $(\tau_0 Z, Z_0)$. By applying Lemma 5 to the almost smooth function φ, we get $\lim_{N\to\infty}\langle \varphi'(h^1, x_0)\rangle \stackrel{\text{a.s.}}{=} \mathbb{E}[\varphi'(\tau_0 Z, X_0)]$.

3.9.3 Step 3: $\mathcal{B}_t$

This part is analogous to step 1 albeit more complex.

(a) Note that

$$Y_t = B_t + [0|M_{t-1}]\Lambda_t, \quad X_t = H_t + Q_t \Xi_t, \tag{3.28}$$

where $B_t = [b^0|\cdots|b^{t-1}]$, $\Lambda_t = \mathrm{diag}(\lambda_0, \ldots, \lambda_{t-1})$, $\Xi_t = \mathrm{diag}(\xi_0, \ldots, \xi_{t-1})$ and $H_t = [h^1|\cdots|h^t]$.

Lemma 9. *The following holds*

(a) $h^{t+1}|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} H_t(M_t^* M_t)^{-1} M_t^* m_{\|}^t + P_{Q_{t+1}}^{\perp} \tilde{A}^* P_{M_t}^{\perp} m^t + Q_t \vec{o}_t(1)$.

(b) $b^t|_{\mathfrak{S}_{t,t}} \stackrel{\mathrm{d}}{=} B_t(Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + P_{M_t}^{\perp} \tilde{A} P_{Q_t}^{\perp} q^t + M_t \vec{o}_t(1)$.

Proof. In light of Lemmas 6 and 8 we have

$$\begin{aligned} h^{t+1}|_{\mathfrak{S}_{t+1,t}} &\stackrel{\mathrm{d}}{=} X_t(M_t^* M_t)^{-1} M_t^* m_{\|}^t + Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1} Y_{t+1}^* m_{\perp}^t + P_{Q_{t+1}}^{\perp} \tilde{A}^* P_{M_t}^{\perp} m^t - \xi_t q^t, \\ b^t|_{\mathfrak{S}_{t,t}} &\stackrel{\mathrm{d}}{=} Y_t(Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + M_t(M_t^* M_t)^{-1} X_t^* q_{\perp}^t + P_{M_t}^{\perp} \tilde{A} P_{Q_t}^{\perp} q^t - \lambda_t m^{t-1}. \end{aligned}$$

Now using equations (3.28), we only need to show

$$\begin{aligned} Q_t \Xi_t (M_t^* M_t)^{-1} M_t^* m_{\|}^t + Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1} Y_{t+1}^* m_{\perp}^t - \xi_t q^t &= Q_t \vec{o}_t(1), \\ [0|M_{t-1}]\Lambda_t (Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + M_t(M_t^* M_t)^{-1} X_t^* q_{\perp}^t - \lambda_t m^{t-1} &= M_t \vec{o}_t(1). \end{aligned}$$

Recall that $m^t_\parallel = M_t\vec{\alpha}$ and $q^t_\parallel = Q_t\vec{\beta}$. On the other hand $Y^*_{t+1}m^t_\perp = B^*_{t+1}m^t_\perp$ because $M^*_t m^t_\perp = 0$. Similarly, $X^*_t q^t_\perp = H^*_t q^t_\perp$. Hence we need to show

$$Q_t\Xi_t\vec{\alpha} + Q_{t+1}(Q^*_{t+1}Q_{t+1})^{-1}B^*_{t+1}m^t_\perp - \xi_t q^t = Q_t\vec{o}_t(1) \tag{3.29}$$

$$[0|M_{t-1}]\Lambda_t\vec{\beta} + M_t(M^*_tM_t)^{-1}H^*_tq^t_\perp - \lambda_t m^{t-1} = M_t\vec{o}_t(1). \tag{3.30}$$

Here is our strategy to prove (3.30) (proof of (3.29) is similar). The left hand side is a linear combination of vectors $m^0, \ldots, m^{t-1}$. For any $\ell = 1, \ldots, t$ we will prove that the coefficient of $m^{\ell-1} \in \mathbb{R}^n$ converges to 0. This coefficient in the left hand side is equal to

$$\left[(M^*_tM_t)^{-1}H^*_tq^t_\perp\right]_\ell - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}} = \sum_{r=1}^{t}\left[(\frac{M^*_tM_t}{n})^{-1}\right]_{\ell,r}\frac{\langle h^r, q^t - \sum_{s=0}^{t-1}\beta_s q^s\rangle}{\delta} - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}}.$$

To simplify the notation denote the matrix $M^*_tM_t/n$ by G. Therefore,

$$\lim_{N\to\infty}\text{ Coefficient of } m^{\ell-1} = \lim_{N\to\infty}\left\{\sum_{r=1}^{t}(G^{-1})_{\ell,r}\langle h^r, q^t - \sum_{s=0}^{t-1}\beta_s q^s\rangle\frac{1}{\delta} - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}}\right\}.$$

But using the induction hypothesis $\mathcal{H}_t(d)$ for $\varphi = f_1, \ldots, f_t$, the term $\langle h^r, q^t - \sum_{s=0}^{t-1}\beta_s q^s\rangle/\delta$ is almost surely equal to the limit of $\langle h^r, h^t\rangle\lambda_t - \sum_{s=0}^{t-1}\beta_s\langle h^r, h^s\rangle\lambda_s$. This can be modified, using the induction hypothesis $\mathcal{H}_t(c)$, to $\langle m^{r-1}, m^{t-1}\rangle\lambda_t - \sum_{s=0}^{t-1}\beta_s\langle m^{r-1}, m^{s-1}\rangle\lambda_s$ almost surely, which can be written as $G_{r,t}\lambda_t - \sum_{s=0}^{t-1}\beta_s G_{r,s}\lambda_s$. Hence,

$$\begin{aligned}\lim_{N\to\infty}\text{ Coefficient of } m^{\ell-1} &\overset{\text{a.s.}}{=} \lim_{N\to\infty}\left\{\sum_{r=1}^{t}(G^{-1})_{\ell,r}[G_{r,t}\lambda_t - \sum_{s=0}^{t-1}\beta_s G_{r,s}\lambda_s] - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}}\right\}\\ &\overset{\text{a.s.}}{=} \lim_{N\to\infty}\left\{\lambda_t\mathbb{I}_{t=\ell} - \sum_{s=0}^{t-1}\beta_s\lambda_s\mathbb{I}_{\ell=s} - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}}\right\}\\ &\overset{\text{a.s.}}{=} 0\end{aligned}$$

Equation (3.29) is proved analogously, using $\xi_t = \langle g'(b^t, w)\rangle$. ☐

The proofs of Eq. (3.11) follows immediately since the last lemma yields

$$b^t|_{\mathfrak{S}_{t,t}} \overset{\text{d}}{=} \sum_{i=0}^{t-1}\beta_i b^i + \tilde{A}q^t_\perp - M_t(M^*_tM_t)^{-1}M^*_t\tilde{A}q^t_\perp + M_t\vec{o}_t(1)\,.$$

Now, using Lemma 2(c), as $n, N \to \infty$,

$$M_t(M^*_tM_t)^{-1}M^*_t\tilde{A}q^t_\perp \overset{\text{d}}{=} \tilde{M}_t\vec{o}_t(1)\,,$$

which finishes the proof since $\tilde{M}_t\vec{o}_t(1) + M_t\vec{o}_t(1) = \tilde{M}_t\vec{o}_t(1)$.

(c) For $r, s < t$ we can use induction hypothesis. For $s = t, r < t$, using Lemma 9 for b^t that was proved above, we get

$$\langle b^t, b^r\rangle|_{\mathfrak{S}_{t,t}} \overset{\text{d}}{=} \sum_{i=0}^{t-1}\beta_i\langle b^i, b^r\rangle + \langle P^\perp_{M_t}\tilde{A}q^t_\perp, b^r\rangle + \sum_{i=0}^{t-1}o(1)\langle m^i, b^r\rangle\,,$$

Note that, by induction hypothesis $\mathcal{B}_{t-1}(d)$ applied to $\varphi = g_{t-1}$, and using the bound $\mathcal{B}_{t-1}(e)$ to control $\langle b^i, b^r\rangle$, we deduce that each term $\langle m^i, b^r\rangle$ has a finite limit. Thus,

$$\lim_{n\to\infty} \sum_{i=0}^{t-1} o(1)\langle m^i, b^r\rangle \overset{\text{a.s.}}{=} 0.$$

We can use Lemma 2 for $\langle P_{M_t}^{\perp} \tilde{A} q_{\perp}^t, b^r\rangle = \langle \tilde{A} q_{\perp}^t, P_{M_t}^{\perp} b^r\rangle$ (recalling that $\tilde{A}$ is independent of $q_{\perp}^t, P_{M_t}^{\perp} b^r$) to obtain

$$\langle \tilde{A} q_{\perp}^t, P_{M_t}^{\perp} b^r\rangle \overset{\text{d}}{=} \frac{\|q_{\perp}^t\| \|P_{M_t}^{\perp} b^r\|}{N} \frac{Z}{\sqrt{N}} \overset{\text{a.s.}\to}{} 0$$

where the last estimate uses the induction hypothesis $\mathcal{B}_r(c)$ and $\mathcal{H}_t(c)$ which imply, almost surely, for some constant C, $\langle P_{M_t}^{\perp} b^r, P_{M_t}^{\perp} b^r\rangle \le \langle b^r, b^r\rangle < C$ and $\langle q_{\perp}^t, q_{\perp}^t\rangle \le \langle q^t, q^t\rangle < C$ for all N large enough. Finally, using the induction hypothesis $\mathcal{B}_r(c)$ or $\mathcal{B}_i(c)$ for each term of the form $\langle b^i, b^r\rangle$ we have

$$\begin{aligned}
\lim_{n\to\infty} \langle b^t, b^r\rangle &\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{\delta} \sum_{i=0}^{t-1} \beta_i \langle q^i, q^r\rangle \\
&\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{\delta} \langle q_{\|}^t, q^r\rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{\delta} \langle q^t, q^r\rangle .
\end{aligned}$$

Last line uses the definition of β_i and $q_{\perp}^t \perp q^r$.

For the case of $r = s = t$, similarly, we have

$$\langle b^t, b^t\rangle|_{\mathfrak{S}_{t,t}} \overset{\text{d}}{=} \sum_{i,j=0}^{t-1} \beta_i \beta_j \langle b^i, b^j\rangle + \langle P_{M_t}^{\perp} \tilde{A} q_{\perp}^t, P_{M_t}^{\perp} \tilde{A} q_{\perp}^t\rangle + o(1).$$

Note that we used similar argument (Lemma 2 and $\mathcal{B}_{t-1}(c)$) to show the contribution of all products of the form $\langle M_t \vec{o}_t(1), \cdot\rangle$ and $\langle P_{M_t}^{\perp} \tilde{A} q_{\perp}^t, b^i\rangle$ a.s. tend to 0. Moreover, using Lemma 2,

$$\begin{aligned}
\lim_{n\to\infty} \langle P_{M_t}^{\perp} \tilde{A} q_{\perp}^t, P_{M_t}^{\perp} \tilde{A} q_{\perp}^t\rangle &= \lim_{n\to\infty} \left[\langle \tilde{A} q_{\perp}^t, \tilde{A} q_{\perp}^t\rangle - \langle P_{M_t} \tilde{A} q_{\perp}^t, P_{M_t} \tilde{A} q_{\perp}^t\rangle \right] \\
&\overset{\text{a.s.}}{=} \lim_{n\to\infty} \left[\frac{\langle q_{\perp}^t, q_{\perp}^t\rangle}{\delta} - o(1) \right]
\end{aligned}$$

Now, using induction hypothesis,

$$\begin{aligned}
\lim_{n\to\infty} \langle b^t, b^t\rangle|_{\mathfrak{S}_{t,t}} &\overset{\text{a.s.}}{=} \lim_{n\to\infty} \sum_{i,j=0}^{t-1} \beta_i \beta_j \frac{\langle q^i, q^j\rangle}{\delta} + \lim_{n\to\infty} \frac{\langle q_{\perp}^t, q_{\perp}^t\rangle}{\delta} \\
&\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{\langle q_{\|}^t, q_{\|}^t\rangle}{\delta} + \lim_{n\to\infty} \frac{\langle q_{\perp}^t, q_{\perp}^t\rangle}{\delta} \\
&\overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{\langle q^t, q^t\rangle}{\delta}.
\end{aligned}$$

(e) Conditioning on $\mathfrak{S}_{t,t}$ and using Eq. (3.11) (proved at point (a) above), almost surely,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{n} (b_i^t)^{2\ell} &= O\left(\frac{1}{n} \sum_{i=1}^{n} (\sum_{r=0}^{t-1} \beta_r b_i^r)^{2\ell} + \frac{1}{n} \sum_{i=1}^{n} ([\tilde{A} q_{\perp}^t]_i)^{2\ell} + o(1) \frac{1}{n} \sum_{r=0}^{t-1} \sum_{i=1}^{n} ([\tilde{m}^r]_i)^{2\ell} \right) \\
&= O\left(\frac{1}{n} \sum_{i=1}^{n} ((q_{\|}^r)_i)^{2\ell} + \frac{1}{n} \sum_{i=1}^{n} ([\tilde{A} q_{\perp}^t]_i)^{2\ell} + o(1) \sum_{r=0}^{t-1} \frac{1}{n} \sum_{i=1}^{n} ([\tilde{m}^r]_i)^{2\ell} \right)
\end{aligned}$$

The term $\frac{1}{n}\sum_{i=1}^{n}([\tilde{A}q_\perp^t]_i)^{2\ell}$ has a finite limit using the same proof as in $\mathcal{B}_0(b)$ and the fact that $\lim_{n\to\infty}\langle q_\perp^t, q_\perp^t\rangle \le \lim_{n\to\infty}\langle q^t, q^t\rangle < \infty$ almost surely. For the term $\frac{1}{n}\sum_{i=1}^{n}((q_\|^r)_i)^{2\ell}$ we use

$$((q_\|^r)_i)^2 \le (q_i^r)^2 = f_t(h_i^r, x_{i,0})^2 \le C\left((h_i^r)^2 + x_{i,0}^2\right) + f(0,0)^2,$$

that follows from the bounded derivative assumption on f_t. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}((q_\|^r)_i)^{2\ell} \le C\frac{1}{n}\sum_{i=1}^{n}(h_i^r)^{2\ell} + C\frac{1}{n}\sum_{i=1}^{n}x_{i,0}^{2\ell} + f(0,0)^{2\ell},$$

which has a finite limit almost surely, using the induction hypothesis $\mathcal{H}_r(e)$ and the assumption on x_0. Similarly, using the bounded derivative assumption on g_t we can show the finiteness of the third term $\frac{1}{n}\sum_{i=1}^{n}([\tilde{m}^r]_i)^{2\ell}$.

(b) Using part (a) we can write

$$\phi_b(b_i^0,\ldots,b_i^t,w_i)|_{\mathfrak{S}_{t,t}} \stackrel{\mathrm{d}}{=} \phi_b\left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r + \tilde{A}q_\perp^t + \tilde{M}_t\vec{o}_t(1)\right]_i, w_i\right).$$

Similar to the proof of $\mathcal{H}_0(b)$ we can drop the error term $M_t\vec{o}_t(1)$. Indeed defining

$$\begin{aligned} a_i &= (b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r + \tilde{A}q_\perp^t + \tilde{M}_t\vec{o}_t(1)\right]_i, w_i)\ ,\\ c_i &= \left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r + \tilde{A}q_\perp^t\right]_i, w_i\right). \end{aligned}$$

By the pseudo-Lipschitz assumption

$$|\phi_b(a_i) - \phi_b(c_i)| \le L\Big\{1 + \max\left(\|a_i\|^{k-1}, \|c_i\|^{k-1}\right)\Big\}\Big|\sum_{r=0}^{t-1}\tilde{m}_i^r\Big|\, o(1).$$

Therefore, using Cauchy-Schwartz inequality twice we have

$$\frac{|\sum_{i=1}^{n}\phi_b(a_i) - \sum_{i=1}^{n}\phi_b(c_i)|}{n} \le L\big[\max(\sum_{i=1}^{n}\frac{\|a_i\|^{2k-2}}{n}, \frac{\sum_{i=1}^{n}\|c_i\|^{2k-2}}{n})\big]^{\frac{1}{2}}\big[\sum_{r=0}^{t-1}t^{\frac{1}{2}}\langle\tilde{m}^r,\tilde{m}^r\rangle\big]^{\frac{1}{2}}o(1). \tag{3.31}$$

Also note that

$$\frac{\sum_{i=1}^{n}\|a_i\|^{2\ell}}{n} \le (t+1)^\ell\sum_{r=0}^{t}\frac{1}{n}\sum_{i=1}^{n}(b_i^r)^{2\ell} + \frac{1}{n}\sum_{i=1}^{n}(w_i)^{2\ell}$$

which is finite almost surely using the induction hypothesis and $\mathcal{B}_t(e)$ proved above and the assumption on w. Similarly, $n^{-1}\sum_{i=1}^{n}\|c_i\|^{2\ell}$ and $\sum_{r=0}^{t-1}\langle\tilde{m}^r,\tilde{m}^r\rangle$ are finite. Hence for any finite t, (3.31) vanishes almost surely when n goes to ∞.

Now given, $b^0,\ldots,b^{t-1}$, consider the random variables

$$\tilde{X}_{i,n} = \phi\left(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\beta_r b_i^r + (\tilde{A}q_\perp^t)_i, w_i\right)$$

and $X_{i,n} \equiv \tilde{X}_{i,n} - \mathbb{E}_{\tilde{A}} \tilde{X}_{i,n}$. Proceeding as in Step 1, and using the pseudo-Lipschitz property of ϕ, it is easy to check the conditions of Theorem 3. We therefore get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ \phi_b \left(b_i^0, \dots, b_i^{t-1}, \left[\sum_{r=0}^{t-1} \beta_r b^r + \tilde{A} q_\perp^t \right]_i, w_i \right) - \mathbb{E}_{\tilde{A}} \phi_b \left(b_i^0, \dots, b_i^{t-1}, \left[\sum_{r=0}^{t-1} \beta_r b^r + \tilde{A} q_\perp^t \right]_i, w_i \right) \Bigg\} \overset{\text{a.s.}}{=} 0. \quad (3.32)$$

Note that $[\tilde{A} q_\perp^t]_i$ is a gaussian random variable with variance $\|q_\perp^t\|^2/n$. Hence we can use induction hypothesis $\mathcal{B}_{t-1}(b)$ for

$$\widehat{\phi}_b(b_i^0, \dots, b_i^{t-1}, w_i) = \mathbb{E}_Z \phi_b \left(b_i^0, \dots, b_i^{t-1}, \sum_{r=0}^{t-1} \beta_r b_i^r + \frac{\|q_\perp^t\| Z}{\sqrt{n}}, w_i \right)$$

where Z is an independent $\mathsf{N}(0,1)$ random variable, to show

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbb{E}_{\tilde{A}} \phi_b \left(b_i^0, \dots, b_i^{t-1}, \left[\sum_{r=0}^{t-1} \beta_r b^r + \tilde{A} q_\perp^t \right]_i, w_i \right)}{n} \overset{\text{a.s.}}{=} \mathbb{E}\, \mathbb{E}_Z \phi_b \left(\sigma_0 Z_0, \dots \sigma_{t-1} Z_{t-1}, \sum_{r=0}^{t-1} \beta_r \sigma_r Z_r + \frac{\|q_\perp^t\| Z}{\sqrt{n}}, W \right) \quad (3.33)$$

Note that $\sum_{r=0}^{t-1} \beta_r \sigma_r Z_r + n^{-1/2} \|q_\perp^t\| Z$ is gaussian. All that we need, is to show that the variance of this gaussian is σ_t^2. But using a combination of (3.32) and (3.33) for the pseudo-Lipschitz function $\phi_b(y_0, \dots, y_t, w_i) = y_t^2$,

$$\lim_{n\to\infty} \langle b^t, b^t \rangle \overset{\text{a.s.}}{=} \mathbb{E} \left\{ \left(\sum_{r=0}^{t-1} \beta_r \sigma_r Z_r + \frac{\|q_\perp^t\| Z}{\sqrt{n}} \right)^2 \right\}. \quad (3.34)$$

On the other hand in part (c) we proved $\lim_{n\to\infty} \langle b^t, b^t \rangle \overset{\text{a.s.}}{=} \lim_{n\to\infty} \delta^{-1} \langle f(h^t, x_0), f(h^t, x_0) \rangle$. By induction hypothesis $\mathcal{H}_t$(b) for the pseudo-Lipschitz function $\phi_h(y_0, \dots, y_t, x_{0,i}) = f(y_t, x_{0,i})^2$ we get $\lim_{n\to\infty} \delta^{-1} \langle f(h^t, x_0), f(h^t, x_0) \rangle \overset{\text{a.s.}}{=} \delta^{-1} \mathbb{E}(f(\tau_{t-1} Z, X_0)^2)$. So by definition (3.2), both sides of (3.34) are equal to σ_t^2.

(d) Very similar to the proof of $\mathcal{B}_0(d)$, using part (b) for the pseudo-Lipschitz function $\phi_b : \mathbb{R}^{t+2} \to \mathbb{R}$ that is given by $\phi_b(y_0, \dots, y_t, w_i) = y_t \varphi(y_s, w_i)$ we can obtain

$$\lim_{n\to\infty} \langle b^t, \varphi(b^s, w) \rangle \overset{\text{a.s.}}{=} \mathbb{E}[\sigma_t \hat{Z}_t \varphi(\sigma_s \hat{Z}_s, W)]$$

for jointly gaussian $\hat{Z}_t, \hat{Z}_s$ with distribution $\mathsf{N}(0,1)$. Using Lemma 3, this is almost surely equal to $\mathrm{Cov}(\sigma_t \hat{Z}_t, \sigma_s \hat{Z}_s) \mathbb{E}(\varphi'(\sigma_s \hat{Z}_s, W))$. By another application of part (b) for $\phi_b(y_0, \dots, y_t, w_i) = y_s y_t$ transforms $\mathrm{Cov}(\sigma_t \hat{Z}_t, \sigma_s \hat{Z}_s)$ to $\lim_{n\to\infty} \langle b^t, b^s \rangle$. Similarly, $\mathbb{E}(\varphi'(\sigma_s \hat{Z}_s, W))$ can be transformed to $\lim_{n\to\infty} \langle \varphi'(b^t, w) \rangle$ almost surely. This finishes the proof of (d).

3.9.4 Step 4: $\mathcal{H}_{t+1}$

Due to symmetry, proof of this step is very similar to the proof of step 3 and we present only some differences.

(a) The proof of Eq. (3.10) follow since by Lemma 9(a) as for $\mathcal{B}_t(a)$

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

Now, using Lemma 2(c), as $n, N \to \infty$,

$$Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1} Q_{t+1}^* \tilde{A}^* m_\perp^t \stackrel{\mathrm{d}}{=} \tilde{Q}_{t+1} \vec{o}_t(1)$$

which finishes the proof since $\tilde{Q}_{t+1}\vec{o}_t(1) + Q_t\vec{o}_t(1) = \tilde{Q}_{t+1}\vec{o}_t(1)$.

(c) For $r, s < t$ we can use induction hypothesis. For $s = t, r < t$, very similar to the proof of $\mathcal{B}_t(a)$,

$$\langle h^{t+1}, b^{r+1}\rangle|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1} \alpha_i \langle h^{i+1}, h^{r+1}\rangle + \langle P_{Q_{t+1}}^\perp \tilde{A}^* m_\perp^t, h^{r+1}\rangle + \sum_{i=0}^{t-1} o(1)\langle q^i, h^{r+1}\rangle.$$

Now, by induction hypothesis $\mathcal{H}_t(d)$, for $\varphi = f$, each term $\langle q^i, h^{r+1}\rangle$ has a finite limit. Thus,

$$\lim_{N\to\infty} \sum_{i=0}^{t-1} o(1)\langle q^i, h^{r+1}\rangle \stackrel{\text{a.s.}}{=} 0.$$

We can use induction hypothesis $\mathcal{H}_{r+1}(c)$ or $\mathcal{H}_i(c)$ for each term of the form $\langle h^i, h^{r+1}\rangle$ and use Lemma 2 for $\langle \tilde{A}^* m_\perp^t, P_{Q_{t+1}}^\perp h^{r+1}\rangle$ to obtain

$$\begin{aligned}\lim_{N\to\infty} \langle h^{t+1}, h^{r+1}\rangle &\stackrel{\text{a.s.}}{=} \lim_{N\to\infty} \sum_{i=0}^{t-1} \alpha_i \langle m^i, m^r\rangle \\ &\stackrel{\text{a.s.}}{=} \lim_{N\to\infty} \langle m_\parallel^t, m^r\rangle \stackrel{\text{a.s.}}{=} \lim_{N\to\infty} \langle m^t, m^r\rangle\,.\end{aligned}$$

Last line uses the definition of α_i and $m_\perp^t \perp m^r$.

For the case of $r = s = t$, we have

$$\langle h^{t+1}, h^{t+1}\rangle|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \sum_{i,j=0}^{t-1} \alpha_i\alpha_j \langle h^{i+1}, h^{j+1}\rangle + \langle P_{Q_{t+1}}^\perp \tilde{A}^* m_\perp^t, P_{Q_{t+1}}^\perp \tilde{A}^* m_\perp^t\rangle + o(1).$$

Note that we used similar argument (Lemma 2 and $\mathcal{H}_t(c)$) to show the contribution of all products of the form $\langle Q_t\vec{o}_t(1), \cdot\rangle$ and $\langle P_{Q_{t+1}}^\perp \tilde{A}^* m_\perp^t, h^{i+1}\rangle$ a.s. tend to 0. Now, using induction hypothesis and Lemma 2

$$\begin{aligned}\lim_{N\to\infty} \langle h^{t+1}, h^{t+1}\rangle|_{\mathfrak{S}_{t+1,t}} &\stackrel{\text{a.s.}}{=} \lim_{n\to\infty} \sum_{i,j=0}^{t-1} \alpha_i\alpha_j\langle m^i, m^j\rangle + \lim_{N\to\infty} \frac{1}{N\delta}\|m_\perp^t\|^2 \\ &\stackrel{\text{a.s.}}{=} \lim_{n\to\infty} \langle m_\parallel^t, m_\parallel^t\rangle + \lim_{n\to\infty} \langle m_\perp^t, m_\perp^t\rangle \\ &\stackrel{\text{a.s.}}{=} \lim_{n\to\infty} \langle m^t, m^t\rangle.\end{aligned}$$

(e) This part is very similar to $\mathcal{B}_t(e)$.

(b) Using part (a) we can write

$$\phi_h(h_i^1, \dots, h_i^{t+1}, x_{0,i})|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \phi_h\left(h_i^1, \dots, h_i^t, \left[\sum_{r=0}^{t-1} \alpha_r h^{r+1} + \tilde{A}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(1)\right]_i, x_{0,i}\right).$$

Similar to proof of $\mathcal{B}_t(b)$ we can drop the error term $\tilde{Q}_{t+1}\vec{o}_{t+1}(1)$. Now given, $h^1, \dots, h^t$, consider the random variables

$$\tilde{X}_{i,N} = \phi_h\left(h_i^1, \dots, h_i^t, \sum_{r=0}^{t-1} \alpha_r h_i^{r+1} + (\tilde{A}^* m_\perp^t)_i, x_{0,i}\right)$$

and $X_{i,N} \equiv \tilde{X}_{i,N} - \mathbb{E}_{\tilde{A}}\tilde{X}_{i,N}$. Proceeding as in Step 2, and using the pseudo-Lipschitz property of ϕ_h, it is easy to check the conditions of Theorem 3. We therefore get

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left\{\phi_h\Big(h_i^1, \dots, h_i^t, \big[\sum_{r=0}^{t-1} \alpha_r h^{r+1} + \tilde{A}^* m_\perp^t\big]_i, x_{0,i}\Big)\right.$$
$$\left. - \mathbb{E}_{\tilde{A}}\phi_h\Big(h_i^1, \dots, h_i^t, \big[\sum_{r=0}^{t-1} \alpha_r b^{r+1} + \tilde{A}^* m_\perp^t\big]_i, x_{0,i}\Big)\right\} \stackrel{\text{a.s.}}{=} 0. \quad (3.35)$$

Note that $[\tilde{A}^* m_\perp^t]_i$ is a gaussian random variable with variance $\|m_\perp^t\|^2/n$. Hence we can use induction hypothesis $\mathcal{H}_t(b)$ for

$$\widehat{\phi}_h(h_i^1, \dots, h_i^t, x_{0,i}) = \mathbb{E}_Z\phi_h\left(h_i^1, \dots, h_i^t, \sum_{r=0}^{t-1} \alpha_r h_i^{r+1} + \frac{\|m_\perp^t\|Z}{\sqrt{n}}, x_{0,i}\right)$$

where Z is an independent $\mathsf{N}(0,1)$ random variable, to show

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} \mathbb{E}_{\tilde{A}}\phi_h\left(h_i^1, \dots, h_i^t, \left[\sum_{r=0}^{t-1} \alpha_r b^{r+1} + \tilde{A}^* m_\perp^t\right]_i, x_{i,0}\right)}{N}$$
$$\stackrel{\text{a.s.}}{=} \mathbb{E}\,\mathbb{E}_Z\phi_h\left(\tau_0 Z_0, \dots \tau_{t-1}Z_{t-1}, \sum_{r=0}^{t-1} \alpha_r \tau_r Z_r + \frac{\|m_\perp^t\|Z}{\sqrt{n}}, X_0\right) \quad (3.36)$$

Note that $\sum_{r=0}^{t-1} \alpha_r \tau_r Z_r + n^{-1/2}\|m_\perp^t\|Z$ is gaussian. All that we need, is to show that the variance of this gaussian is τ_t^2. But using combination of (3.35) and (3.36) for the pseudo-Lipschitz function $\phi_h(y_0, \dots, y_t, x_{0,i}) = y_t^2$,

$$\lim_{N\to\infty} \langle h^{t+1}, h^{t+1}\rangle \stackrel{\text{a.s.}}{=} \mathbb{E}\left\{\left(\sum_{r=0}^{t-1} \alpha_r \tau_r Z_r + \frac{\|m_\perp^t\|Z}{\sqrt{n}}\right)^2\right\}. \quad (3.37)$$

On the other hand in part (c) we proved $\lim_{N\to\infty}\langle h^{t+1}, h^{t+1}\rangle \stackrel{\text{a.s.}}{=} \lim_{N\to\infty}\langle g_t(b^t, w), g_t(b^t, w)\rangle$. By the induction hypothesis $\mathcal{B}_t$(b) for the pseudo-Lipschitz function $\phi_b(y_0, \dots, y_t, w) = g_t(y_t, w)^2$ we get $\lim_{n\to\infty}\langle g_t(b^t, w), g_t(b^t, w)\rangle \stackrel{\text{a.s.}}{=} \mathbb{E}(g_t(\sigma_t Z, W)^2)$. So by the definition (1.4), both sides of (3.37) are equal to τ_t^2.

(*d*) This is very similar to the proof of $\mathcal{B}_t(d)$. For the pseudo-Lipschitz function $\phi_h : \mathbb{R}^{t+2} \to \mathbb{R}$ that is given by $\phi_h(y_1, \dots, y_{t+1}, x_{0,i}) = y_{t+1}\varphi(y_{s+1}, x_{0,i})$ we can use part (a) to obtain

$$\lim_{N\to\infty} \langle h^{t+1}, \varphi(b^{s+1}, x_0)\rangle \overset{\text{a.s.}}{=} \mathbb{E}[\tau_t Z_t \varphi(\tau_s Z_s, X_0)]$$

for jointly gaussian Z_t, Z_s with distribution $\mathsf{N}(0,1)$. Using Lemma 3, this is almost surely equal to $\mathrm{Cov}(\tau_t Z_t, \tau_s Z_s)\mathbb{E}(\varphi'(\tau_s Z_s, X_0))$. And another application of part (b) for $\phi_h(y_1, \dots, y_{t+1}, x_{i,0}) = y_{s+1}y_{t+1}$ transforms $\mathrm{Cov}(\tau_t Z_t, \tau_s Z_s)$ to $\lim_{N\to\infty}\langle h^{t+1}, h^{s+1}\rangle$. Similarly, $\mathbb{E}(\varphi'(\tau_s Z_s, X_0))$ can be transformed to $\lim_{N\to\infty}\langle \varphi'(h^{t+1}, x_0)\rangle$ almost surely. This finishes the proof of (d).

3.10 Proof of Corollary 1

First notice that the statement to be proved is equivalent to the following claim. The joint distribution of $(x^t_{I(1)}, \dots, x^t_{I(\ell)}, x_{0,I(1)}, \dots, x_{0,I(\ell)})$, for $I(1), \dots, I(\ell) \in [N]$ uniformly random subset of distinct indices, converges weakly to to the distribution of $(\widehat{X}_1, \dots \widehat{X}_\ell, X_{0,1}, \dots, X_{0,\ell})$. By general theory of weak convergence, it is therefore sufficient to check Eq. (1.7) for functions of the form

$$\psi(x_1, \dots, x_\ell, y_1, \dots, y_\ell) = \psi_1(x_1, y_1)\cdots\psi_\ell(x_\ell, y_\ell)\,, \tag{3.38}$$

for $\psi_i : \mathbb{R}^2 \to \mathbb{R}$ Lipschitz and bounded. This case follows immediately from Theorem 1 once we notice that

$$\mathsf{E}\,\psi(x^t_{I(1)}, \dots, x^t_{I(\ell)}, x_{0,I(1)}, \dots, x_{0,I(\ell)}) = \prod_{s=1}^{\ell}\Big(\frac{1}{N}\sum_{i=1}^{N}\psi_s(x^t_i, x_{0,i})\Big) + O(1/N)\,. \tag{3.39}$$

4 Symmetric Case

Let $k \geq 2$. Let $G = A^* + A$ with $A \in \mathbb{R}^{N\times N}$ and of A are iid $\mathsf{N}(0, (2N)^{-1})$. Also let $f : \mathbb{R} \to \mathbb{R}$ be a function almost everywhere differentiable with bounded first derivative. Start with $m^0, m^1 \in \mathbb{R}^N$ where $m^0 = \vec{0}$ and m^1 is a fixed deterministic vector in $\mathbb{R}^N$ with $\sum_{i=1}^N (m_{1,i})^{2k-2} \leq Nc$ for a constant c, and proceed by the following iteration

$$\begin{aligned} h^{t+1} &= Gm^t - \lambda_t m^{t-1}, \\ m^t &= f(h^t) \end{aligned} \tag{4.1}$$

where $\lambda_t = \langle f'(h^t)\rangle$. Now let $\tau_1^2 = \lim_{N\to\infty}\langle m_1, m_1\rangle$, and define recursively for $t \geq 1$,

$$\tau_{t+1}^2 = \mathbb{E}\left\{[f(\tau_t Z)]^2\right\}\,, \tag{4.2}$$

with $Z \sim \mathsf{N}(0,1)$.

Theorem 4. *Let $\{A(N)\}_N$ be a sequence of matrices $A \in \mathbb{R}^{N\times N}$ indexed by N, with iid entries $A_{ij} \sim \mathsf{N}(0, 1/(2N)^{-1})$. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order at most k and all $t \in \mathbb{N}$, almost surely*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi(h_i^{t+1}) = \mathbb{E}\left[\psi(f(\tau_t Z))\right]\,. \tag{4.3}$$

Note 3. *This theorem was proved by Bolthausen in the case $f(x) = \tanh(\beta x + h)$ and $\langle m^1, m^1\rangle = \tau_*^2$, for τ_*^2 the fixed point of the recursion (4.2). The general proof is very similar to the one of Theorem 2, and exploits the same conditioning trick. We omit it to avoid repetitions.*

When we are calculating h^{t+1}, all values $h^1, \dots, h^t$ and hence $m^1, \dots, m^t$ are known to us. Denote the σ-algebra generated by all of these random variables by $\mathfrak{U}_t$. Moreover, use the following compact formulation for (4.2).

$$\underbrace{[h^2|h^3 + \lambda^2 m^1|\cdots|h^t + \lambda^{t-1}m^{t-2}]}_{Y_{t-1}} = G\,\underbrace{[m^1|\dots|m^{t-1}]}_{M_{t-1}},$$

The analogous of Lemma 1 is the following.

Lemma 10. *Let $\{A(N)\}_N$ be a sequence of sensing matrices as in Theorem 4. Then the following hold for all $t \in \mathbb{N}$*

(*a*)

$$h^{t+1}|_{\mathfrak{U}_t} \overset{\mathrm{d}}{=} \sum_{i=1}^{t-1} \alpha_i h^{i+1} + \tilde{G} m_\perp^t + \tilde{M}_{t-1}\vec{o}_t(1) \tag{4.4}$$

where $\tilde{G}$ is an independent copy of G and coefficients α_i satisfy $m_\|^t = \sum_{i=1}^{t-1}\alpha_i m^i$. The matrix $\tilde{M}_t$ is such that its columns form an orthogonal basis for the column space of M_t and $\tilde{M}_t^*\tilde{M}_t = n\,\mathbf{I}_t$. Further, $\vec{o}_t(1) \in \mathbb{R}^t$ is a finite dimensional random vector that converges to 0 almost surely as $N \to \infty$.

(*b*) *For any pseudo-Lipschitz function $\phi : \mathbb{R}^t \to \mathbb{R}$ of order at most k,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \phi(h_i^2, \dots, h_i^{t+1}) \overset{\mathrm{a.s.}}{=} \mathbb{E}\big[\phi(\tau_1 Z_1, \dots, \tau_t Z_t)\big] \tag{4.5}$$

where $Z_1, \dots, Z_t$ have $\mathsf{N}(0,1)$ distribution.

(*c*) *For all $1 \le r, s \le t$ the following equations hold and all limits exist, are bounded and non-random.*

$$\lim_{N\to\infty} \langle h^{r+1}, h^{s+1}\rangle \overset{\mathrm{a.s.}}{=} \lim_{N\to\infty} \langle m^r, m^s\rangle \tag{4.6}$$

(*d*) *For all $1 \le r, s \le t$, and for any almost everywhere differentiable function φ with bounded derivative, the following equations hold and all limits exist, are bounded and non-random.*

$$\lim_{N\to\infty} \langle h^{r+1}, \varphi(h^{s+1})\rangle \overset{\mathrm{a.s.}}{=} \lim_{N\to\infty} \langle h^{r+1}, h^{s+1}\rangle\langle\varphi'(h^{s+1})\rangle \tag{4.7}$$

(*e*) *For $\ell = k - 1$, almost surely $\lim_{N\to\infty}(h_i^{t+1})^{2\ell} < \infty$.*

Acknowledgement

We are deeply indebted with Erwin Bolthausen, who first presented the conditioning technique during the EURANDOM workshop on 'Order, disorder and double disorder' in September 2009. We are also grateful to Dave Donoho and Arian Maleki, for a number of insightful exchanges.

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

A Proof of two probability lemmas

In this Appendix we provide proofs of two probability lemmas stated in Section 3.6.

A.1 Proof of Lemma 4

Note that by definition of empirical measure, $N^{-1}\sum_{i=1}^{N}\psi(v_i) = \mathbb{E}_{\hat{p}_v}[\psi(v)]$. The proof uses a truncation technique. For a positive integer B define ψ_B by

$$\psi_B(x)\begin{cases} \psi(x) & |\psi(x)| \le B \\ B & \psi(x) > B \\ -B & \psi(x) < -B \end{cases}$$

and write $\psi(x) = \psi_B(x) + \tilde{\psi}_B(x)$. Since $\hat{p}_v$ converges weakly to p_V, for the bounded continuous function $\psi_B(x)$,

$$\lim_{N\to\infty} \mathbb{E}_{\hat{p}_v}[\psi_B(V)] = \mathbb{E}_{p_V}[\psi_B(V)]. \tag{A.1}$$

On the other hand, since ψ is pseudo-Lipschitz with order k we have $|\psi(x)| \le L(1+|x|^k)$. Therefore when $B > 1$,

$$|\tilde{\psi}_B(x)| \le L(1+|x|^k)\mathbb{I}_{\{|\psi|>B\}} \le L(1+|x|^k)\mathbb{I}_{\{|x|^k > \frac{B}{L}-1\}}.$$

From this we obtain

$$\begin{aligned}
\mathbb{E}_{p_V}[\psi_B(V)] - \lim\sup_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[L(1+|V|^k)\mathbb{I}_{\{|V|^k>\frac{B}{L}-1\}}] \\
\le \lim\inf_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[\psi(V)] \le \lim\sup_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[\psi(V)] \le \\
\mathbb{E}_{p_V}[\psi_B(V)] + \lim\sup_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[L(1+|V|^k)\mathbb{I}_{\{|V|^k>\frac{B}{L}\}}].
\end{aligned}$$

Now, by assumption $\lim_{N\to\infty}\mathbb{E}_{\hat{p}_{v(N)}}(|V|^k) = \mathbb{E}_{p_V}(|V|^k)$ we can write $|V|^k = |V|^k\mathbb{I}_{\{|V|^k>B/L-1\}} + |V|^k\mathbb{I}_{\{|V|^k\le B/L-1\}}$ and use the weak convergence of $\hat{p}_{v(N)}$ to p_V to get

$$\lim_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[L(1+|V|^k)\mathbb{I}_{\{|V|^k\le\frac{B}{L}-1\}}] = \mathbb{E}_{p_V}[L(1+|V|^k)\mathbb{I}_{\{|V|^k\le\frac{B}{L}-1\}}].$$

Therefore

$$\begin{aligned}
\lim\sup_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[L(1+|V|^k)\mathbb{I}_{\{|V|^k>\frac{B}{L}-1\}}] = \lim_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}[L(1+|V|^k)\mathbb{I}_{\{|V|^k>\frac{B}{L}-1\}}] = \\
= \mathbb{E}_{p_V}[L(1+|V|^k)\mathbb{I}_{\{V^k>\frac{B}{L}-1\}}].
\end{aligned}$$

Hence, all we need to show is that $\mathbb{E}_{p_V}[L|V|^k\mathbb{I}_{\{|V|^k>\frac{B}{L}-1\}}]$ converges to 0 as $B\to\infty$. But this follows using the bounded k^{th} moment of V and the dominated convergence theorem, when applied to the sequence of functions $L(1+|V|^k)\mathbb{I}_{\{|V|^k>B/L-1\}}| \le L(1+|V|^k)$, indexed by B.

A.2 Proof of Lemma 5

Recall that by Skorokhod's theorem, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a construction of the random variables $\{(X_n, Y_n)\}_{n\geq 1}$ and (X, Y) on this space, such that letting

$$A = \big\{\omega \in \Omega \,:\, (X_n(\omega), Y_n(\omega)) \to (X(\omega), Y(\omega))\big\}\,,$$

be the event that (X_n, Y_n) converges to (X, Y), we have $\mathbb{P}(A) = 1$. Let $\mathcal{C}_F \subseteq \mathbb{R}^2$ be the domain on which F is continuously differentiable. By Fubini's theorem $\mathcal{C}_F$ has measure 1 under the probability distribution of (X, Y). Hence if we let

$$B = \big\{\omega \in \Omega \,:\, (X(\omega), Y(\omega)) \in \mathcal{C}_F\big\}\,,$$

we have $\mathbb{P}(B) = 1$. On $A \cap B$, we also have $F'(X_n(\omega), Y_n(\omega)) \to F'(X(\omega), Y(\omega))$.

Letting $Z_n(\omega) \equiv F'(X_n(\omega), Y_n(\omega))$ (if $(X_n(\omega), Y_n(\omega)) \notin \mathcal{C}_F$ set $Z_n(\omega) = 0$) and $Z(\omega) \equiv F'(X(\omega), Y(\omega))$, we thus proved that

$$\mathbb{P}\big\{\lim_{n\to\infty} Z_n(\omega) = Z(\omega)\big\} = 1\,.$$

Since $|Z_n(\omega)| \leq C$ by assumption, the bounded convergence theorem implies $\mathbb{E}\{Z_n(\omega)\} \to \mathbb{E}\{Z(\omega)\}$ which proves our claim.

References

[Bol09] E. Bolthausen, *On the high-temperature phase of the Sherrington-Kirkpatrick model*, Seminar at EURANDOM, Eindhoven, September 2009.

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

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

[DMM10b] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, Preprint, 2010.

[Don06] D. Donoho, *For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution*, Communications on Pure and Applied Mathematics **59** (2006), 907–934.

[DT05] D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.

[DT09] ———, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of American Mathematical Society **22** (2009), 1–53.

[HT97] T. C. Hu and R. L. Taylor, *Strong law for arrays and for the bootstrap mean and variance*, Internat. J. Math. and Math. Sci. **20** (1997), 375–383.

[Kab03] Y. Kabashima, *A CDMA multiuser detection algorithm on the basis of belief propagation*, J. Phys. A **36** (2003), 11111–11121.

[MT06] A. Montanari and D. Tse, *Analysis of belief propagation for non-linear problems: the example of CDMA (or: how to prove Tanaka's formula)*, Proceedings of IEEE Inform. Theory Workshop (Punta de l'Este, Uruguay), 2006.

[NS05] J.P. Neirotti and D. Saad, *Improved message passing for inference in densely connected systems*, Europhys. Lett. **71** (2005), 866–872.

[RU08] T.J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.

[Ste72] C. Stein, *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 1972.

[TH99] D. N.C. Tse and S. V. Hanly, *Linear multiuser receivers: effective interference, effective bandwidth and user capacity*, IEEE Trans. on Inform. Theory **45** (1999), 641–657.

[VS99] S. Verdu and S. Shamai, *Spectral efficiency of CDMA with random spreading*, IEEE Trans. on Inform. Theory **45** (1999), 622–640.