The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing

Mohsen Bayati and Andrea Montanari

Dedicated to the memory of Ralf Koetter (1963–2009)

***Abstract*—"Approximate message passing" (AMP) algorithms have proved to be 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 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 a large number of short cycles in the underlying factor graph. It relies instead on a conditioning technique recently developed by Erwin Bolthausen in the context of spin glass theory.**

***Index Terms*—Compressed sensing, density evolution, message passing algorithms, random matrix theory, state evolution.**

I. 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 a noise vector and A is assumed to be known. Recently [10] 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 nonlinear 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. [10] for details).

Throughout this paper, the matrix A is normalized in such a way that its columns have ℓ_2 norm[1] concentrated around 1. Given a vector $x \in \mathbb{R}^N$ and a scalar function $f : \mathbb{R} \to \mathbb{R}$, 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 [10]:

1) For a large class of random matrices A, the behavior of the AMP algorithm is accurately described by a formalism called "state evolution" (SE). Extensive numerical experiments tested this claim on Gaussian, Radamacher, and partial Fourier matrices;
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 convex optimization approaches. Let us stress that standard convex optimization algorithms do not scale to large applications (e.g., to image processing), while the computational complexity of AMP is as low as the one of the simplest greedy algorithms;
3) As a byproduct of (1) and (2), SE allows to re-derive reconstruction phase boundaries earlier determined via random polytope geometry (see in particular [15], [16] and references therein).

These findings were based on heuristic arguments and 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) will be reported in a forthcoming paper.

Interestingly, state evolution gives access to sharp predictions that cannot be derived from random polytope geometry. A prominent example is the noise sensitivity of LASSO, which is investigated in [13].

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

$$\begin{aligned} z^t_{a\to i} &= y_a - \sum_{j\in[N]\setminus i} A_{aj} x^t_{j\to a} \\ x^{t+1}_{i\to a} &= \eta_t\left(\sum_{b\in[n]\setminus a} A_{bi} z^t_{b\to i} \right). \end{aligned} \tag{1.2}$$

[1]Recall that the ℓ_p norm of a vector v is $\|v\|_p \equiv (\sum_i |v_i|^p)^{1/p}$.

Manuscript received February 26, 2010; revised June 21, 2010; accepted August 05, 2010. Date of current version January 19, 2011. This work was partially supported by a Terman fellowship, the NSF CAREER Award CCF-0743978, and the NSF grant DMS-0806211. The material in this paper was presented at the IEEE International Symposium on Information Theory, Austin, TX, June 2010.

This paper is part of the special issue on "Facets of Coding Theory: From Algorithms to Networks," dedicated to the scientific legacy of Ralf Koetter.

M. Bayati is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: bayati@stanfordalumni.org).

A. Montanari is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: montanari@stanford.edu).

Communicated by G. D. Forney, Jr., Associate Editor for the special issue on "Facets of Coding Theory: From Algorithms to Networks."

Digital Object Identifier 10.1109/TIT.2010.2094817

As argued in [11], AMP accurately approximates message passing in the large system limit. We refer to Appendix A for an heuristic argument justifying the AMP update rules (1.1) starting from the algorithm (1.2). While this derivation is not necessary for the proofs of this paper, it can help the reader familiar with message passing algorithms to develop the correct intuition.

An important tool for the analysis of message passing algorithms is provided by density evolution [27]. 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 can be regarded (in a very precise sense) as the analogue 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 [20], [24], [25]. Applications to such algorithms are discussed in Section II. Section III describes an even more general formulation, as well as the proof of our theorems. Finally, Section IV describes a generalization to the case of symmetric matrices A that is directly related to the work of Erwin Bolthausen [4].

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 remarkable that density evolution (in its special incarnation, SE) holds for dense graphs. This is at odds with the standard argument used for justifying density evolution so far: "density evolution works *because* the graph is locally tree-like." To the best of our knowledge, the approach developed here is the first one that overcomes the limitations of the standard argument (a discussion of earlier literature is provided in Section I-D).

A. 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. In Section III-B, we will show that the i.i.d. assumption can be relaxed to existence of a weak limit for the empirical distribution of w with certain moment conditions. Further, let $\{\eta_t\}_{t\geq 0}$ be a sequence of scalar functions $\eta_t : \mathbb{R} \to \mathbb{R}$ which we assume to be Lipschitz continuous (and hence almost everywhere differentiable). 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 (1.1).

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

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

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 III-B) and to the sequence $\{\tau_t\}_{t\geq 0}$ that it defines.

Let us denote the empirical distribution[2] of a vector $x_0 \in \mathbb{R}^N$ by $\hat{p}_{x_0}$. Further, for $k > 1$ 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 $L_{M,m} < \infty$ such that for all $x, y \in [-M, M]^m$,

$$|\phi(x) - \phi(y)| \leq L_{M,m}\|x - y\|.$$

Further, $L_{M,m} \leq c[1 + (M\sqrt{m})^{k-1}]$ 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 i.i.d. 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 $k \geq 2$. Also, assume the noise w has iid entries with a distribution p_W that has bounded $(2k-2)^{th}$ moment. 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.

Up to a trivial change of variables, this is a formalization of the findings of [10] (cf. in particular [10, eqs. (1.7), (1.8), Finding 2]).

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 $J(1), \ldots, J(\ell) \in [N]$.

Then for all $t > 0$, almost surely

$$\begin{aligned}\lim_{N\to\infty} \mathsf{E}\, \psi(x^t_{J(1)}, \ldots, x^t_{J(\ell)}, x_{0,J(1)}, \ldots, x_{0,J(\ell)}) \\ = \mathbb{E}\Big\{\psi\big(\hat{X}_1, \ldots, \hat{X}_\ell, X_{0,1}, \ldots, X_{0,\ell}\big)\Big\},\end{aligned} \tag{1.7}$$

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

where $\hat{X}_i \equiv \eta_{t-1}(X_{0,i} + \tau_{t-1} Z_i)$ 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 III-J.

B. Universality

Our proof technique heavily relies on the assumption that $A(N)$ is Gaussian. Nevertheless, we expect the convergence expressed in Theorem 1 to be a fairly general result. In particular, we expect it to hold for matrices with i.i.d. entries with zero mean and variance $1/n$, under a suitable moment condition. This type of *universality* is quite common in random matrix theory, and several arguments suggest that it should hold in the present case. For instance, it is possible to prove that state evolution holds for this broader class of random matrices when $\eta_t(\cdot)$ is affine. Also, the heuristic argument discussed in the next section is clearly insensitive to the details of distribution of the entries.

Numerical evidence presented in [10] (we refer in particular to the online supplement) suggests that state evolution might hold for an even broader class of matrices. Determining the domain of such an universality class is an outstanding open problem.

C. State Evolution: The Basic Intuition

The state evolution recursion has a simple heuristic description, that is useful to present here since it clarifies the difficulties involved in the proof. In particular, this description brings up the key role played by the last term in the update equation for z^t, that we will call the 'Onsager term', following [10].

Consider again the recursion (1.1), but introduce the following three modifications: (i) Replace the random matrix A with a new independent copy $A(t)$ at each iteration t; (ii) Correspondingly replace the observation vector y with $y^t = A(t)x_0 + w$; (iii) Eliminate the last term in the update equation for z^t. We thus get the following dynamics:

$$x^{t+1} = \eta_t(A(t)^* z^t + x^t), \tag{1.8}$$

$$z^t = y^t - A(t)x^t \tag{1.9}$$

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

This recursion is most conveniently written by eliminating z^t

$$\begin{aligned} x^{t+1} &= \eta_t\big(A(t)^* y^t + (\mathsf{I} - A(t)^* A(t))x^t\big) \\ &= \eta_t\big(x_0 + A(t)^* w + B(t)(x^t - x_0)\big) \end{aligned} \tag{1.10}$$

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

Using the central limit theorem, it is easy to show that each entry of $B(t)$ is approximately normal, with zero mean and variance $1/n$. Further, distinct entries are approximately pairwise independent. Therefore, if we let $\hat{\tau}_t^2 = \lim_{N\to\infty} \|x^t - x_0\|^2/N$, we obtain that $B(t)(x^t - x_0)$ converges to a vector with i.i.d. normal entries with 0 mean and variance $N\hat{\tau}_t^2/n = \hat{\tau}_t^2/\delta$. Notice that this is true because $A(t)$ is independent of $\{A(s)\}_{1\le s\le t-1}$ and, in particular, of $(x^t - x^0)$.

Conditional on w, $A(t)^* w$ is a vector of i.i.d. normal entries with mean 0 and variance $(1/n)\|w\|^2$ which converges by the law of large numbers to σ^2. A slightly longer exercise shows that these entries are approximately independent from the ones of $B(t)(x^t - x_0)$. Summarizing, each entry of the vector in the argument of η_t in (1.10) converges to $X_0 + \tau_t Z$ with $Z \sim \mathsf{N}(0,1)$ independent of X_0, and

$$\begin{aligned} \tau_t^2 &= \sigma^2 + \frac{1}{\delta}\hat{\tau}_t^2, \\ \hat{\tau}_t^2 &= \lim_{N\to\infty} \frac{1}{N}\|x^t - x_0\|^2. \end{aligned} \tag{1.11}$$

On the other hand, by (1.10), each entry of $x^{t+1} - x_0$ converges to $\eta_t(X_0 + \tau_t\, Z) - X_0$, and therefore

$$\begin{aligned} \hat{\tau}_{t+1}^2 &= \lim_{N\to\infty} \frac{1}{N}\|x^{t+1} - x_0\|^2 \\ &= \mathbb{E}\big\{[\eta_t(X_0 + \tau_t\, Z) - X_0]^2\big\}. \end{aligned} \tag{1.12}$$

Using together (1.11) and (1.12) we finally obtain the state evolution recursion, (1.4).

We conclude that state evolution would hold if the matrix A was drawn independently from the same Gaussian distribution at each iteration. In the case of interest, A does not change across iterations, and the above argument falls apart because x^t and A are dependent. This dependency is nonnegligible even in the large system limit $N \to \infty$. This point can be clarified by considering the iteration

$$\begin{aligned} x^{t+1} &= \eta_t(A^* z^t + x^t) \\ z^t &= y - Ax^t \end{aligned} \tag{1.13}$$

with a matrix A constant across iterations. This iteration is the basis of several algorithms in compressed sensing, most notably the so-called 'iterative soft thresholding' [6]. Such algorithms have been the object of great interest because of the high computational cost of standard convex optimization methods in large scale applications.

Numerical studies of iterative soft thresholding [9], [10] show that its behavior is dramatically different from the one in (1.1) and in particular *state evolution does not hold for the iterative soft thresholding iteration* (1.13), even in the large system limit.

This is not a surprise: the correlations between A and x^t simply cannot be neglected. On the other hand, adding the Onsager term leads to an asymptotic cancellation of these correlations. As a consequence, state evolution holds for the AMP iteration (1.1) despite the fact that the matrix is kept constant.

D. Related Literature

As mentioned, the standard argument for justifying density evolution relies on the locally-tree like structure of the underlying graph. This argument was developed and systematically exploited for the analysis of low-density parity-check (LDPC) codes under iterative decoding [27]. In this context, density evolution provides an exact tool for computing asymptotic thresholds of code ensembles based on sparse graph constructions.

Optimization of these thresholds has been a major design principle in LDPC codes.

The locally tree-like property is a special case of local weak convergence. Local weak convergence of graph sequences was first defined and studied in probability theory by Benjamini and Schramm [5], and then greatly developed by David Aldous [2], in particular to study the so called 'random assignment problem' [1]. Loosely speaking, local weak convergence allows to treat sequences of graphs of increasing size, such that the neighborhood of a node converges to a well defined limit object.

The random assignment problem is defined as a distribution of random instances of the assignment problem on complete bipartite graphs. In particular, such graphs are not locally tree-like. Nevertheless, they admit a rather simple local weak limit (called the PWIT), which is a tree. The basic reason is that only a sparse subgraph of the complete bipartite graph is relevant for the minimum cost assignment, namely the one of edges with small cost. One concrete way to derive density evolution in this case is indeed to eliminate all the edges of cost larger than -say- Δ_n/n with Δ_n diverging slowly with the graph size n. The resulting graph is sparse and one can apply standard arguments (cf. [22] for an outline of this argument). A more sophisticated argument was presented in [28] which nevertheless uses the existence of a nontrivial local weak limit, and the fact that only a sparse subgraph is relevant [28, Lemma 4.1].

This reduction to a sparse graph, and, hence, to a limit tree, is impossible in the class of algorithms studied in our paper: the algorithm iteration cannot be approximated by an iteration on a sparse graph (at least not on an instance-by-instance basis). This corresponds to the fact that no (simple) local weak limit exists in our case. The underlying graph is the complete bipartite graph with vertex sets $[N] \equiv \{1, 2, \dots, N\}$ and $[n] \equiv \{1, 2, \dots, n\}$, and edge-weights A_{ai} for all $(a, i) \in [n] \times [N]$. If we choose a node $i \in [N]$ as root, its depth-1 neighborhood consists of $[n]$ node, each carrying a weight of order $1/\sqrt{n}$. Even this small neighborhood has no simple local weak limit.

This difference is analogous to the difference between mean-field spin glasses (e.g., the Sherrington-Kirkpatrick model) and the random assignment problem [30, Ch. 7]. As a consequence, our proof does not rely on local weak convergence, and has to deal directly with the intricacies of graphs with many short cycles.

The theorem proved in this paper is not only relevant for [10] but for a larger context as well. First of all, following the work by Tanaka [31], hundreds of papers have been published in information theory using the replica method to study multi-user detection problems. In its replica-symmetric version, the replica method typically predicts the system performances through the solution of a system of nonlinear equations, which coincide with the fixed point equations for state evolution. The present result provides a rigorous foundation to that line of work, along with the analysis of a concrete algorithm that achieves those performance. Further, [18] insisted on the role of a 'decoupling principle' that emerges from the replica method, and on the insight it provides. Corollary 1 indeed proves a specific form of this decoupling principle.

A more recent line of works uses the replica method to study typical performances of compressed sensing methods. Although nonrigorous and limited to asymptotic statements, the replica method has the advantage of providing sharp predictions. Standard techniques instead predict performances up to undetermined multiplicative constants. The determination of these constants can be of guidance for practical applications. This motivated several groups to publish results based on the replica method [17], [21], [26]. The present paper provides a rigorous foundation to this work as well.

II. 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 almost surely

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

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 $p \le k$ through

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

almost surely.

A. 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. $\mathsf{N}(0, v^2)$. It is a consequence of state evolution that the optimal AMP algorithm makes use of linear scalar estimators

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

Clearly, such functions are Lipschitz continuous, for any λ_t finite. The AMP algorithm (1.1) becomes

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

State evolution reads

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

Theorem 1 also shows that the empirical distribution of $\{(A^* z^t + x^t)_i - x_{0,i}\}_{i\in[N]}$ is asymptotically Gaussian with mean 0 and variance τ_t^2. Hence, the optimal choice of λ_t is

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

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

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

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

$$\tau_\infty^2 = \frac{1}{2}\Big\{(\sigma^2 + cv^2) + \sqrt{(\sigma^2 + cv^2)^2 + 4\sigma^2 v^2}\Big\}. \quad (2.8)$$

The mean square error of the resulting algorithm is estimated via (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, thus yielding

$$\lim_{t\to\infty}\lim_{N\to\infty} \frac{1}{N}\|x^t - x_0\|^2 = \frac{\delta}{2}\Big\{(-\sigma^2 + cv^2) + \sqrt{(\sigma^2 + cv^2)^2 + 4\sigma^2 v^2}\Big\}. \quad (2.9)$$

We recall that the asymptotic mean square error of optimal (MMSE) linear estimation was computed by Tse-Hanly and Verdú-Shamai in the case of random matrices A with i.i.d. entries [33], [34]. The motivation came from the analysis of multiuser receivers. The resulting MSE coincides with the value predicted in (2.9), thus showing that—in the linear case—the AMP algorithm is asymptotically equivalent to the MMSE estimator.

Notice that the computation of the MMSE in [33], [34] relied heavily on the Marcenko-Pastur law for the limit spectral law of Wishart matrices [23]. Conversely, any calculation of the MMSE as a function of the noise variance σ^2 gives access to the asymptotic Stieltjis transform of the spectral measure of A. This suggests that state evolution is a nontrivial result already in the case of linear $\eta_t(\cdot)$, since it can be used to derive the Marcenko-Pastur law in random matrix theory.

B. 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

$$\mathbb{P}\{X_0 \neq 0\} = \varepsilon. \quad (2.10)$$

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

In [10], the authors proposed an algorithm of the form (1.1) with $\eta_t(x) = \eta(x;\theta_t)$ a sequence of soft-threshold functions

$$\eta(x;\theta) = \begin{cases} (x-\theta) & \text{if } x > \theta \\ 0 & \text{if } -\theta \le x \le \theta \\ (x+\theta) & \text{if } x < -\theta. \end{cases} \quad (2.11)$$

The function $x \mapsto \eta(x;\theta)$ is nonlinear but nevertheless it is Lipschitz continuous. Therefore Theorem 1 applies to this case, and allows to predict the asymptotic mean square error using (1.4) and (1.6).

This choice of the nonlinearity η_t is close to the optimal in minimax sense. Indeed, a substantial literature (see, e.g., [7] and [8]) studies the problem of estimating the scalar X_0 from the noisy observation

$$Y = X_0 + Z \quad (2.12)$$

with $Z \sim \mathsf{N}(0, s^2)$. For an appropriate choice of the threshold $\theta = \theta(\varepsilon, s)$, and $\varepsilon \downarrow 0$ (very sparse sources), the soft thresholding estimator was proved to be minimax optimal, i.e., to achieve the minimum worst-case MSE over the class (2.10). State evolution allows to deduce that the choice (2.11) yields the best algorithm of the form (1.1) for estimating sparse vectors, over the worst-case vector x_0 [10].

It is argued in [10], [13], and proved in [3] in the case of Gaussian matrices, that the asymptotic MSE of AMP coincides with the one of a popular convex optimization estimation technique, known as the LASSO. The above argument is suggestive of a possible way to prove minimax optimality of the LASSO.

Finally, state evolution provides a systematic way of improving the choice of the nonlinearities η_t when the class of signal changes. The basic idea is to choose the function η_t that minimizes the RHS of (1.4) in minimax sense. This corresponds to constructing minimax MMSE estimators for the scalar problem (2.12). For instance, in the limit case in which the distribution of X_0 is known, the MMSE estimator is simply conditional expectation, which leads to the choice

$$\eta_t(x) = \mathbb{E}\{X_0 \mid X_0 + \tau_t Z = x\} \quad (2.13)$$

with $Z \sim \mathsf{N}(0,1)$. In other words, the very choice of the nonlinearities is dictated by the Gaussian convergence phenomenon described in Theorem 1.

C. 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. A frequently used setting for theoretical analysis consists in taking the large system limit with $n/N \to \delta$ giving the spreading factor, and in assuming that the signatures (and hence A) have i.i.d. components. The entries $x_{0,i}$ belong to the signal constellation used by the system. For the sake of simplicity, we consider the case of antipodal signaling, i.e., $x_{0,i} \in \{+1,-1\}$ uniformly at random. Other signal constellations can also be treated applying our Theorem 1. The hypothesis that x_0 is independent from A is also standard in this context and justified by the remark that the transmitted information is independent from the signatures. Further, the source-channel separation theorem naturally leads to the uniform distribution.

Following [20], [24], [25], we take

$$\eta_t(x) = \tanh\left\{\frac{x}{\tau_t^2}\right\}. \quad (2.14)$$

The rationale for this choice is that it gives the conditional expectation of a uniformly random signal $X_0 \in \{+1,-1\}$, given the observation $X_0 + \tau_t Z = x$ for $Z \sim \mathsf{N}(0,1)$ Gaussian noise. This is therefore a special case of the rule (2.13) and by the argument given there, it achieves minimal mean-square error within the class of algorithms (1.1).

The algorithm (1.1) reads in this case

$$\begin{aligned} x^{t+1} &= \tanh\left\{\frac{1}{\tau_t^2}(A^* z^t + x^t)\right\},\\ z^t &= y - Ax^t\\ &\quad + \frac{z^{t-1}}{\delta\tau_t^2}\left\{1 - \left\langle \tanh^2\left[\frac{(A^* z^t + x^t)}{\tau_t^2}\right]\right\rangle\right\}. \end{aligned} \tag{2.15}$$

State evolution yields

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big\{\big[\tanh\big(\tau_t^{-2} + \tau_t^{-1}Z\big) - 1\big]^2\Big\}. \tag{2.16}$$

This state evolution recursion was proved in [24] for properly chosen sparse signature matrices A. Theorem 1 provides the first generalization to the more relevant case of dense signatures.

As mentioned in Section I-D, Tanaka used the replica method to compute the asymptotic performance of a MMSE receiver. The expressions obtained through this method correspond to a fixed point of the recursion (2.16). It was further proved in [24] that, whenever the fixed point is unique, this prediction is asymptotically correct. For such values of the parameters, we deduce that the AMP algorithm is asymptotically equivalent to the MMSE receiver.

Let us point out that, in a practical setting, it might be inconvenient to estimate the noise variance and/or to change the function η_t across iterations. Several authors (see for instance [32]) used the function

$$\eta_t(x) = \tanh\{\beta x\}. \tag{2.17}$$

State evolution can be applied in this case as well (for any finite β) and reads

$$\tau_{t+1}^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}\Big\{\big[\tanh\big(\beta + \beta\tau_t Z\big) - 1\big]^2\Big\}. \tag{2.18}$$

On the other hand the case $\beta \to \infty$ is not covered by our Theorem 1, since it corresponds to the discontinuous function $\eta_t(x) = \mathrm{sign}(x)$.

III. 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 [4]. Related ideas can also be found in [14].

In the next section, we provide a high-level description of the conditioning technique, by using a simpler type of recursion as reference. We will then introduce some new notations and state and prove a more general result than Theorem 1.

A. The Conditioning Technique: An Informal Description

For understanding the conditioning technique, it is convenient to consider a somewhat simpler setting, namely the one of symmetric matrices. This will be discussed more formally in Section IV. Let $G = A^* + A$ where $A \in \mathbb{R}^{N\times N}$ has i.i.d. entries $A_{ij} \sim \mathsf{N}(0, (2N)^{-1})$. We consider the iteration

$$h^{t+1} = Gm^t - \lambda_t m^{t-1}, \tag{3.1}$$

$$m^t = f(h^t), \tag{3.2}$$

for $f : \mathbb{R} \to \mathbb{R}$ a nonlinear function and $m^{-1} = 0$. For the sake of simplicity, $h^0 = 0$. The correct expression for the scalar λ_t is provided in Section IV, and state evolution holds only if this value is used. On the other hand, this expression is not important for our informal discussion here.

Consider the first iteration. By definition $m^{-1} = 0$, whence $h^1 = Gf(0)$ is a vector with i.i.d. Gaussian components with variance $\|f(0)\|^2/N$. This follows in particular by the rotational invariance of the distribution of G, which implies that, for a deterministic vector v, Gv is distributed as $\|v\|Ge_1$ for e_1 the first vector of the standard basis of $\mathbb{R}^N$ (see also Lemma 2 below).

Now consider the t^{th} iteration (i.e., $h^{t+1} = Gm^t - \lambda_t m^{t-1}$). The problem in repeating the above argument is that G and $f(m^t)$ are dependent. For instance $f(m^t)$ might *a priori* align with the minimum eigenvector of G. More generally the problem is that G is not independent from the σ-algebra $\mathfrak{S}_t$ generated by $\{h^0, h^1, \dots, h^t\}$.

The key idea in the conditioning technique is to avoid computing the conditional distribution of m^t given G. We instead compute the *conditional distribution of G given $\mathfrak{S}_t$*.

The next important remark is that $\mathfrak{S}_t$ contains $\{m^0, m^1, \dots, m^t\}$ as well. Conditioning on $\mathfrak{S}_t$ is therefore equivalent to conditioning on the event

$$\mathcal{E}_t \equiv \{h^1 + \lambda^0 m^{-1} = Gm^0, \dots, h^t + \lambda^{t-1}m^{t-2} = Gm^{t-1}\}.$$

which is in turn equivalent to making a set of linear observations of G.

At this point, the assumption that G is Gaussian plays a crucial role. The conditional distribution of a Gaussian random variable G given linear observations is the same as its conditional expectation plus the projection of an independent Gaussian. In formulae:

$$G|_{\mathfrak{S}_t} \stackrel{\rm d}{=} G|_{\mathcal{E}_t} \stackrel{\rm d}{=} \mathbb{E}\{G|\mathfrak{S}_t\} + \mathrm{P}_\perp^t G^{\rm new}\mathrm{P}_\perp^t,$$

with $\mathrm{P}_\perp^t$ an appropriate projector. If we write $E_t \equiv \mathbb{E}\{G|\mathfrak{S}_t\}$, we have

$$Gm^t|_{\mathfrak{S}_t} \stackrel{\rm d}{=} G^{\rm new}(\mathrm{P}_\perp^t m^t) - (\mathsf{I} - \mathrm{P}_\perp^t)G^{\rm new}(\mathrm{P}_\perp^t m^t) + E_t m^t.$$

We refer to the actual proof for a calculation of the various terms involved.

Each of the above terms can be written explicitly as a function of the observed values $\{m^0, m^1, \dots, m^t\}$ and of the new Gaussian random variables $G^{\rm new}$. The first term $G^{\rm new}(\mathrm{P}_\perp^t m^t)$ is clearly Gaussian. The other terms are not. In order to control them, we will proceed by induction over t and use an appropriate strong law of large numbers for triangular arrays. The key phenomenon is that the only non-Gaussian term that does not vanish in the large system limit cancels with the term $-\lambda_t m^{t-1}$ in recursion (3.1), thus implying the claimed Gaussianity of h^{t+1}.

B. A General Result

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 case of the recursion defined here.

The algorithm is defined by two sequences of functions $\{f_t\}_{t\geq 0}$, $\{g_t\}_{t\geq 0}$, 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 Lipschitz continuous. Recall that Lipschitz functions are continuous, and are almost everywhere continuously differentiable with a bounded derivative. 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 $w\in\mathbb{R}^n$ and $x_0\in\mathbb{R}^N$, 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 obtaining $\{b^t\}_{t\geq 0}$, $\{m^t\}_{t\geq 0}$, $\{h^t\}_{t\geq 1}$, and $\{q^t\}_{t\geq 1}$ through

$$\begin{aligned} h^{t+1} &= A^* m^t - \xi_t\, q^t, \quad m^t = g_t(b^t,w),\\ b^t &= A\, q^t - \lambda_t m^{t-1}, \quad q^t = f_t(h^t,x_0), \end{aligned} \tag{3.3}$$

where $\xi_t=\langle g_t'(b^t,w)\rangle$, $\lambda_t=\frac{1}{\delta}\langle f_t'(h^t,x^0)\rangle$ (both derivatives are with respect to the first argument), and we recall that -by definition- $m^{-1}=0$.

Assume that the limit

$$\sigma_0^2\equiv\lim_{N\to\infty}\frac{1}{N\delta}\|q^0\|^2 \tag{3.4}$$

exists, is positive and finite, for a sequence of initial conditions of increasing dimensions. State evolution defines quantities $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 0}$ via

$$\begin{aligned} \tau_t^2 &= \mathbb{E}\big\{g_t(\sigma_t Z,W)^2\big\},\\ \sigma_t^2 &= \frac{1}{\delta}\,\mathbb{E}\big\{f_t(\tau_{t-1}Z,X_0)^2\big\}, \end{aligned} \tag{3.5}$$

where $W\sim p_W$ and $X_0\sim p_{X_0}$ are independent of $Z\sim\mathsf{N}(0,1)$. Further, recall the notion of pseudo-Lipschitz function for $k>1$ from Section I-A. We have the following general result.

Theorem 2: Let $\{q_0(N)\}_{N\geq 0}$ and $\{A(N)\}_{N\geq 0}$ be, respectively, a sequence of 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 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 k and all $t\geq 0$, almost surely

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N\psi(h_i^{t+1},x_{0,i})=\mathbb{E}\Big\{\psi\big(\tau_t Z,X_0\big)\Big\} \tag{3.6}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\psi(b_i^t,w_i)=\mathbb{E}\Big\{\psi\big(\sigma_t Z,W\big)\Big\} \tag{3.7}$$

where $X_0\sim p_{X_0}$ and $W\sim p_W$ are independent of $Z\sim\mathsf{N}(0,1)$, and σ_t,τ_t are determined by recursion (3.5).

C. Corollary of Theorem 2: AMP and Theorem 1

As already mentioned, the AMP algorithm (1.1) is a special case of recursion (3.3). The reduction is obtained by defining

$$h^{t+1}=x_0-(A^*z^t+x^t) \tag{3.8}$$
$$q^t=x^t-x_0 \tag{3.9}$$
$$b^t=w-z^t \tag{3.10}$$
$$m^t=-z^t. \tag{3.11}$$

The functions f_t and g_t are given by

$$f_t(s,x_0)=\eta_{t-1}(x_0-s)-x_0,\quad g_t(s,w)=s-w \tag{3.12}$$

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

Note 1:

(a) Although the recursions (1.1) and (3.3) are equivalent mathematically, only the former can be used as an algorithm. Indeed the recursion (3.3) tracks the difference of the current estimates x^t from x_0, and is initialized using x^0 itself. The recursion (3.3) 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).

D. Proof of Theorem 1

First note that (3.5) reduces to

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

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

$$\sigma_0^2=\frac{1}{\delta}\lim_{N\to\infty}\frac{1}{N}\|q^0\|^2=\frac{1}{\delta}\mathbb{E}_{p_{X_0}}(X_0^2)$$

and $\tau_0^2=\sigma^2+\sigma_0^2$. 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 almost surely

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

with $Z\sim\mathsf{N}(0,1)$ independent of $X_0\sim p_{X_0}$, which yields the claim as Z has the same distribution as $-Z$. Note that since η is Lipschitz continuous, when ψ belongs to $\mathrm{PL}(k)$ then $(h_i^t,x_{0,i})\mapsto\psi(\eta_{t-1}(x_{0,i}-h_i^t),x_{0,i})$ also belongs to $\mathrm{PL}(k)$.

E. Definitions and Notations

When the update equation for h^{t+1} in (3.3) 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} conditioned on all these known variables and also conditioned on x_0 and w. In particular, define $\mathfrak{S}_{t_1,t_2}$ to be the σ-algebra generated by $b^0,\dots,b^{t_1-1},m^0,\dots,m^{t_1-1},h^1,\dots,h^{t_2}$,

$q^0, \dots, q^{t_2}$, and x_0 and w. The basic idea of the proof is to compute the conditional distributions $b^t|_{\mathfrak{S}_{t,t}}$ 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 and 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

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

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

or in short $X_t = A^* M_t$ and $Y_t = A Q_t$. Here and below we use vertical lines to indicate columns of a matrix, i.e., $[a_1 | a_2 | \dots | a_k]$ is the matrix with columns $a_1, \dots, a_k$.

We also introduce the notation $m^t_{\|}$ for the projection of m^t onto the column space of M_t and define $m^t_{\perp} = m^t - m^t_{\|}$. Similarly, define $q^t_{\|}$ and $q^t_{\perp}$ to be the parallel and orthogonal projections of q^t onto column space of Q_t. In particular, let $\vec{\alpha} = \vec{\alpha}_t = (\alpha_0, \dots, \alpha_{t-1})$ and $\vec{\beta} = \vec{\beta}_t = (\beta_0, \dots, \beta_{t-1})$ be the vectors (in $\mathbb{R}^t$) of coefficients for these projections, i.e.

$$m^t_{\|} = \sum_{i=1}^{t-1} \alpha_i m^i, \quad q^t_{\|} = \sum_{i=0}^{t-1} \beta_i q^i. \tag{3.13}$$

We will show in Section III-I (cf. Corollary 2) that for any fixed t as N goes to infinity the quantities β_i's and α_j's have a finite limit.

Recall that D^* denotes the transpose of the matrix D and for a vector $u \in \mathbb{R}^m$: $\langle u \rangle = \sum_{i=1}^m u_i/m$. Also, for vectors $u, v \in \mathbb{R}^m$ we define the scalar product

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

Given two random variables X, Y, and a σ-algebra $\mathfrak{S}$, the notations $X|_{\mathfrak{S}} \stackrel{\mathrm{d}}{=} Y$ 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 words 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 $X \stackrel{\mathrm{d}}{=} Y$ (i.e., X and Y are equal in distribution). For random variables X, Y the notation $X \stackrel{\mathrm{a.s.}}{=} Y$ means that X and Y are equal almost surely.

The large system limit will be denoted either as $\lim_{N\to\infty}$ or as $\lim_{n\to\infty}$. It is understood that either of the two dimensions can index the sequence of problems under consideration, and that $n/N \to \delta$. In the large system limit, we use the notation $\vec{o}_t(1)$ to represent a vector in $\mathbb{R}^t$ (with t fixed) such that all of its coordinates converge to 0 almost surely as $N \to \infty$.

Finally, we will use $\mathsf{I}_{d\times d}$ to denote the $d \times d$ identity matrix (and drop the subscript when dimensions should be clear from the context). Similarly, $\mathsf{0}_{n\times m}$ is used to denote the $n \times m$ zero matrix. The indicator function of property $\mathcal{A}$ is denoted by $\mathbb{I}(\mathcal{A})$ or $\mathbb{I}_{\mathcal{A}}$. The normal distribution with mean μ and variance v^2 is $\mathsf{N}(\mu, v^2)$.

F. 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 defined uniquely by the recursion (3.5) 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)

$$h^{t+1}|_{\mathfrak{S}_{t+1,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1} \alpha_i h^{i+1} + \tilde{A}^* m^t_{\perp} + \tilde{Q}_{t+1} \vec{o}_{t+1}(1) \tag{3.14}$$

$$b^t|_{\mathfrak{S}_{t,t}} \stackrel{\mathrm{d}}{=} \sum_{i=0}^{t-1} \beta_i b^i + \tilde{A} q^t_{\perp} + \tilde{M}_t \vec{o}_t(1) \tag{3.15}$$

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 $\tilde{Q}_t^* \tilde{Q}_t = N\, \mathsf{I}_{t\times t}$ ($\tilde{M}_t^* \tilde{M}_t = n\, \mathsf{I}_{t\times t}$).

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

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \phi_h(h_i^1, \dots, h_i^{t+1}, x_{0,i}) \stackrel{\mathrm{a.s.}}{=} \mathbb{E}\{\phi_h(\tau_0 Z_0, \dots, \tau_t Z_t, X_0)\}, \tag{3.16}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \phi_b(b_i^0, \dots, b_i^t, w_i) \stackrel{\mathrm{a.s.}}{=} \mathbb{E}\{\phi_b(\sigma_0 \hat{Z}_0, \dots, \sigma_t \hat{Z}_t, W)\}, \tag{3.17}$$

where $(Z_0, \dots, Z_t)$ and $(\hat{Z}_0, \dots, \hat{Z}_t)$ 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):

$$\lim_{N\to\infty} \langle h^{r+1}, h^{s+1} \rangle \stackrel{\mathrm{a.s.}}{=} \lim_{n\to\infty} \langle m^r, m^s \rangle \tag{3.18}$$

$$\lim_{n\to\infty} \langle b^r, b^s \rangle \stackrel{\mathrm{a.s.}}{=} \frac{1}{\delta} \lim_{N\to\infty} \langle q^r, q^s \rangle. \tag{3.19}$$

(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):

$$\lim_{N\to\infty} \langle h^{r+1}, \varphi(h^{s+1}, x_0) \rangle \stackrel{\mathrm{a.s.}}{=} \lim_{N\to\infty} \langle h^{r+1}, h^{s+1} \rangle \langle \varphi'(h^{s+1}, x_0) \rangle, \tag{3.20}$$

$$\lim_{n\to\infty} \langle b^r, \varphi(b^s, w) \rangle \stackrel{\mathrm{a.s.}}{=} \lim_{n\to\infty} \langle b^r, b^s \rangle \langle \varphi'(b^s, w) \rangle. \tag{3.21}$$

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

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

$$\lim\sup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}(h_i^{t+1})^{2\ell} < \infty, \tag{3.22}$$

$$\lim\sup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(b_i^{t})^{2\ell} < \infty. \tag{3.23}$$

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

$$\lim_{N\to\infty} \frac{1}{N}\langle h^{r+1}, q^0\rangle \overset{\text{a.s.}}{=} 0. \tag{3.24}$$

(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

$$\lim_{N\to\infty} \langle q_\perp^r, q_\perp^r\rangle > \rho_r \tag{3.25}$$

$$\lim_{n\to\infty} \langle m_\perp^s, m_\perp^s\rangle > \varsigma_s. \tag{3.26}$$

Note 2: Equations (3.20) and (3.21) have the form of Stein's lemma [29] (cf. Lemma 3 in Section III-H).

Proof of Theorem 2: Assuming Lemma 1 is correct Theorem 2 easily follows. To be more precise, Theorem 2 is a obtained by applying Lemma 1(b) to functions $\phi_h(y_0,\dots,y_t,x_{0,i}) = \psi(y_t,x_{0,i})$ and $\phi_b(y_0,\dots,y_t,w_i) = \psi(y_t,w_i)$.

The rest of Section III focuses on proof of Lemma 1.

G. 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) for triangular arrays of independent but not identically distributed random variables. The form stated below follows immediately from [19, Th. 2.1].

Theorem 3 (SLLN, [19]): 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 to zero for each n and $n^{-1}\sum_{i=1}^{n}\mathbb{E}|X_{n,i}|^{2+\kappa} \le cn^{\kappa/2}$ for some $0<\kappa<1$, $c<\infty$. Then $\frac{1}{n}\sum_{i=1}^{n} X_{i,n} \to 0$ almost surely 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) $v^*\tilde{A}u \overset{\text{d}}{=} Z/\sqrt{n}$ 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 $P_W A u \overset{\text{d}}{=} Dx$ 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.

Lemma 3 (Stein's Lemma [29]): For jointly Gaussian random variables Z_1, Z_2 with zero mean, and any function $\varphi : \mathbb{R}\to\mathbb{R}$ where $\mathbb{E}\{\varphi'(Z_1)\}$ and $\mathbb{E}\{Z_1\varphi(Z_2)\}$ exist, 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 B-I.

Lemma 4: Let $k \ge 2$ and consider a sequence of vectors $\{v(N)\}_{N\ge 0}$, 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 k:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi(v_i) \overset{\text{a.s.}}{=} \mathbb{E}\big[\psi(V)\big]. \tag{3.27}$$

Next lemma is on weak convergence of Lipschitz functions and its proof is in Appendix B-II.

Lemma 5: Let $F : \mathbb{R}^2 \to \mathbb{R}$ be Lipschitz continuous 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) is 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 and Y are independent and that the distribution of X is absolutely continuous with respect to the Lebesgue measure. Then

$$\lim_{n\to\infty}\mathbb{E}\{F'(X_n,Y_n)\} = \mathbb{E}\{F'(X,Y)\}. \tag{3.28}$$

It is useful to remember a standard formula for the conditional variance of Gaussian random variables.

Lemma 6: Let $(Z_1,\dots,Z_t)$ be a normal random vector with zero mean, and assume that the covariance matrix of $(Z_1,\dots,Z_{t-1})$ (denoted by C) is invertible. Then

$$\mathrm{Var}(Z_t|Z_1,\dots,Z_{t-1}) = \mathbb{E}\{Z_t^2\} - u^*C^{-1}u,$$

where $u \in \mathbb{R}^{t-1}$ is given by $u_i \equiv \mathbb{E}\{Z_tZ_i\}$.

An immediate consequence is the following fact, proven in Appendix B-III.

Lemma 7: Let $Z_1,\dots,Z_t$ be a sequence of jointly Gaussian random variables and let $c_1,\dots,c_t$ be strictly positive constants such that for all $i=1,\dots,t$: $\mathrm{Var}(Z_i|Z_1,\dots,Z_{i-1}) \ge c_i$. Further assume $\mathbb{E}\{Z_i^2\} \le K$ for all i and some constant K. Let Y be a random variable in the same probability space.

Finally let $\ell : \mathbb{R}^2 \to \mathbb{R}$ be a Lipschitz function, with $z \mapsto \ell(z,Y)$ nonconstant with positive probability (with respect to Y).

Then there exist a positive constant c_t' (depending on $c_1,\dots,c_t$, on K, on the random variable Y, and on the function ℓ) such that

$$\mathbb{E}\{[\ell(Z_t,Y)]^2\} - u^*C^{-1}u > c_t'$$

where $u \in \mathbb{R}^{t-1}$ is given by $u_i \equiv \mathbb{E}\{\ell(Z_t,Y)\ell(Z_i,Y)\}$, and $C \in \mathbb{R}^{t-1\times t-1}$ satisfies $C_{ij} \equiv \mathbb{E}\{\ell(Z_i,Y)\,\ell(Z_j,Y)\}$ for $1 \le i,j \le t-1$.

Linear Algebra Facts: It is also convenient to recall some linear algebra facts. The first one is proved in Appendix B-IV.

Lemma 8: Let $v_1,\dots,v_t$ be a sequence vectors in $\mathbb{R}^n$ such that for all $i=1,\dots,t$

$$\frac{1}{n}\|v_i - P_{i-1}(v_i)\|^2 \ge c$$

for a positive constant c and let P_{i-1} be the orthogonal projector to the span of $v_1,\dots,v_{i-1}$. Then there is a constant c' (depending only on c and t), such that the matrix $C \in \mathbb{R}^{t\times t}$ with $C_{ij} = \langle v_i, v_j\rangle$ satisfies

$$\lambda_{\min}(C) \ge c'.$$

The second one is just a direct consequence of the fact that the mapping $S \mapsto \lambda_{\min}(S)$ is continuous at any matrix S that is invertible.

Lemma 9: Let $\{S_n\}_{n\ge 1}$ be a sequence of $t\times t$ matrices such that $\liminf_{n\to\infty}\lambda_{\min}(S_n) > c$ for a positive constant c. Also assume that $\lim_{n\to\infty} S_n = S_\infty$ where the limit is element-wise. Then, $\lambda_{\min}(S_\infty) \ge c$.

H. Conditional Distributions

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

Lemma 10: For $(t_1,t_2) = (t,t)$ or $(t_1,t_2) = (t+1,t)$, the conditional distribution of the random matrix A given the σ-algebra $\mathfrak{S}_{t_1,t_2}$, satisfies

$$A|_{\mathfrak{S}_{t_1,t_2}} \overset{\mathrm{d}}{=} E_{t_1,t_2} + \mathcal{P}_{t_1,t_2}(\tilde{A}). \tag{3.29}$$

Here $\tilde{A} \overset{\mathrm{d}}{=} A$ 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

$$\begin{aligned} E_{t_1,t_2} = {} & Y_{t_1}(Q_{t_1}^* Q_{t_1})^{-1}Q_{t_1}^* + M_{t_2}(M_{t_2}^* M_{t_2})^{-1}X_{t_2}^* \\ & - M_{t_2}(M_{t_2}^* M_{t_2})^{-1}M_{t_2}^* Y_{t_1}(Q_{t_1}^* Q_{t_1})^{-1}Q_{t_1}^*. \end{aligned} \tag{3.30}$$

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

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

Here $P_{M_{t_2}}^{\perp} = \mathsf{I} - P_{M_{t_2}}$, $P_{Q_{t_1}}^{\perp} = \mathsf{I} - P_{Q_{t_1}}$, 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 11: Let $z \in \mathbb{R}^n$ be a random vector with i.i.d. $\mathsf{N}(0,v^2)$ entries and let $\mathrm{D} \in \mathbb{R}^{m\times n}$ be a linear operator with full row rank. Then for any constant vector $b \in \mathbb{R}^m$ the distribution of z conditioned on $\mathrm{D}z = b$ satisfies:

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

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 i.i.d. $\mathsf{N}(0,v^2)$. Moreover, $\mathrm{D}^*(\mathrm{DD}^*)^{-1}b = \arg\min_z\{\|z\|^2 | \mathrm{D}z = b\}$.

Proof: The result is trivial if $\mathrm{D} = [\mathsf{I}_{m\times m}|\mathsf{0}_{m\times(n-m)}]$. 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 $\mathrm{D}^*(\mathrm{DD}^*)^{-1}b = \arg\min_z\{\|z\|^2|\mathrm{D}z=b\}$. □

Lemma 10 follows from applying Lemma 11 to the operator D that maps A to (AQ, M^*A). A detailed proof of Lemma 10 appears in Section III-H-I. Note that we can assume, without loss of generality f, g to be nonconstant 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 nonsingular almost surely, and hence the above expressions are well defined.

Lemma 12: The following holds:

$$\begin{aligned} E_{t+1,t}^* m^t = {} & X_t(M_t^* M_t)^{-1}M_t^* m_{\parallel}^t \\ & + Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1}Y_{t+1}^* m_{\perp}^t, \end{aligned} \tag{3.31}$$

$$\begin{aligned} E_{t,t} q^t = {} & Y_t(Q_t^* Q_t)^{-1}Q_t^* q_{\parallel}^t \\ & + M_t(M_t^* M_t)^{-1}X_t^* q_{\perp}^t. \end{aligned} \tag{3.32}$$

Proof: Write $m^t = m_{\parallel}^t + m_{\perp}^t$. Using (3.30) 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 $m_{\parallel}^t = \sum_{i=0}^{t-1}\alpha_i m^i = M_t\vec{\alpha}$. Then using $A^* M_t = X_t$, (3.29), and $[\mathcal{P}_{t+1,t}(\tilde{A})]^* m_{\parallel}^t = 0$ we have

$$\begin{aligned} E_{t+1,t}^* m_{\parallel}^t = {} & Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1}Y_{t+1}^* m_{\parallel}^t \\ & + X_t(M_t^* M_t)^{-1}M_t^* m_{\parallel}^t \\ & - Q_{t+1}(Q_{t+1}^* Q_{t+1})^{-1}Y_{t+1}^* m_{\parallel}^t \\ = {} & X_t(M_t^* M_t)^{-1}M_t^* m_{\parallel}^t. \end{aligned}$$

Similarly, writing $q^t = q_{\parallel}^t + q_{\perp}^t$, $q_{\parallel}^t = Q_t\vec{\beta}$, and using $X_t^* Q_t = M_t^* A Q_t = M_t^* Y_t$, $Q_t^* q_{\perp}^t = 0$ we obtain (3.32). □

Proof of Lemma 10: 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, just in Section III-H-I, we will drop all subindices t_1, t_2. The expression (3.30) for the conditional expectation $E = \mathbb{E}\{A|_{\mathfrak{S}_{t_1,t_2}}\}$ follows from Lemma 11 and the following calculation for:

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

where $\|A\|_F$ denotes the Frobenius norm of matrix A. 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

$$2A = \Theta Q^* + M\Gamma^*. \tag{3.33}$$

Equation (3.33) 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 $\Theta_R Q^* + M\Gamma_R^* = \Theta_0 Q^* + M\Gamma_0^* = 2A$. In particular for $R_1 = \Gamma_0^* Q(Q^*Q)^{-1}$ we have $Q^*\Gamma_{R_1} = 0$. Multiplying (3.33) by Q from right (using Θ_{R_1}, Γ_{R_1}) we have $2Y = \Theta_{R_1} Q^* Q$ or $\Theta_{R_1} = 2Y(Q^*Q)^{-1}$. Now multiplying (3.33) by M^* from left we obtain $2X^* = 2M^*Y(Q^*Q)^{-1}Q^* + M^*M\Gamma_{R_1}^*$ which leads to $\Gamma_{R_1}^* = 2(M^*M)^{-1}\left[X^* - M^*Y(Q^*Q)^{-1}Q^*\right]$. From these we see that $E = \mathbb{E}(A|\mathfrak{S}_{t_1,t_2})$ satisfies:

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

Now we are left with the task of proving that $\mathcal{P}_{t_1,t_2}(\tilde{A}) = P_M^\perp \tilde{A} P_Q^\perp$. We need to show that the linear operator $\mathcal{F} : A \mapsto P_M^\perp A P_Q^\perp$ 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 : $(\mathcal{F}(A), B) = (A, \mathcal{F}(B))$.

Now we check (a)–(d):

(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

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

I. *Proof of Lemma 1*

The proof is by induction on t. Let $\mathcal{H}_{t+1}$ be the property that (3.14), (3.16), (3.18), (3.20), (3.22), (3.24), and (3.25) hold. Similarly, let $\mathcal{B}_t$ be the property that (3.15), (3.17), (3.19), (3.21), (3.23), and (3.26) 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 several properties that we will denote by (a), (b), (c), (d), (e), and (g) according to their appearance in Lemma 1. For $\mathcal{H}$ we also need to prove a property (f).

It is immediate to check that our claims become trivial if $x \mapsto f_t(x, X_0)$ is constant (i.e., independent of x) almost surely (with respect to $X_0 \sim p_{X_0}$), or if $x \mapsto g_t(x, W)$ is constant almost surely (with respect to $W \sim p_W$). We will, therefore, assume that neither of these degenerate cases hold.

Step 1: $\mathcal{B}_0$: Note that $b^0 = Aq^0$.

(a) $\mathfrak{S}_{0,0}$ is generated by x_0, q^0 and w. Also $q^0 = q_\perp^0$ since Q_0 is an empty matrix. Hence

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

(b) Let $\phi_b : \mathbb{R}^2 \to \mathbb{R}$ be a pseudo-Lipschitz function of order 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(a) $[Aq^0]_i \stackrel{\rm d}{=} Z\|q^0\|/\sqrt{n}$ for $Z \sim \mathsf{N}(0,1)$. Hence, using

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

for all $p \ge 2$ 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

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}|\phi_b\left(b_i^0, w_i\right) - \mathbb{E}_A\{\phi_b(b_i^0, w_i)\}|^{2+\kappa} \le c' + \frac{L'c''}{n}\sum_{i=1}^n |w_i|^{(k-1)(2+\kappa)} \le cn^{\kappa/2}$$

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

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

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

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

Note that ψ belongs to $\mathrm{PL}(k)$ since ϕ_b belongs to $\mathrm{PL}(k)$.

(c) Using Lemma 2, conditioned on q^0

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

(d) Using $\mathcal{B}_0(b)$, and $\phi(x, w_i) = x\varphi(x, w_i)$ we obtain $\lim_{n\to\infty}\langle b^0, \varphi(b^0, w)\rangle \stackrel{\rm a.s.}{=} \mathbb{E}\{\sigma_0\hat{Z}\varphi(\sigma_0\hat{Z}, W)\}$, which is equal to $\sigma_0^2\mathbb{E}\{\varphi'(\sigma_0\hat{Z}, W)\}$ using Lemma 3. Note that $x\phi$ belongs to $\mathrm{PL}(k)$.
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 $\lim_{n\to\infty}\langle\varphi'(b^0, w)\rangle \stackrel{\rm a.s.}{=} \mathbb{E}\{\varphi'(\sigma_0\hat{Z}, W)\}$.

(e) Similar to (b), conditioning on q^0, the term $\sum_{i=1}^n([Aq^0]_i)^{2\ell}/n$ is sum of independent random variables (namely, Gaussians to the power 2ℓ) 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

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

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

(f) This case is trivial since there is no m^s with $s \le t-1 = -1$.

Step 2: $\mathcal{H}_1$: Note that $h^1 = A^* m^0 - \xi_0 q^0$.

(a) $\mathfrak{S}_{1,0}$ is generated by x_0, q^0, w, b^0 and m^0. Also $m^0 = m^0_\perp$ since M_0 is an empty matrix. Applying Lemma 10 we have

$$A|_{\mathfrak{S}_{1,0}} \stackrel{\mathrm{d}}{=} b^0 \|q^0\|^{-2}(q^0)^* + \tilde{A} P^\perp_{q^0}.$$

Hence

$$h^1|_{\mathfrak{S}_{1,0}} \stackrel{\mathrm{d}}{=} P^\perp_{q^0} \tilde{A}^* m^0 + \delta \frac{\langle b_0, m_0\rangle}{\langle q^0, q^0\rangle} q^0 - \xi_0 q^0.$$

But using $\mathcal{B}_0(d)$ for $\varphi = g_0$

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

Therefore

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

Also $\mathcal{B}_0(b)$, applied to the function $\phi_b(x,w) = g_0(x,w)^2$ gives

$$\lim_{n\to\infty} \langle m^0, m^0\rangle \stackrel{\text{a.s.}}{=} \mathbb{E}[g_0(\sigma_0 Z, W)^2] = \tau_0^2 < \infty. \tag{3.34}$$

Thus

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

where the last estimate follows from Lemma 2(c) and (3.34). Finally

$$h^1|_{\mathfrak{S}_{1,0}} \stackrel{\mathrm{d}}{=} \tilde{A}^* m^0 + \vec{o}_1(1) q^0. \tag{3.35}$$

(c) Using (3.35), (3.34), and Lemma 2, we get

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

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

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

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 $\sum_{i=1}^N ([\tilde{A}^* m^0]_i)^{2\ell}/N$.

(f) Using (3.35) and Lemma 2(a) we have almost surely

$$\begin{aligned}\lim_{N\to\infty} \langle h^1, q^0\rangle &\stackrel{\mathrm{d}}{=} \lim_{N\to\infty} \frac{Z\|m^0\|\|q^0\|}{N\sqrt{n}} \\ &\stackrel{\text{a.s.}}{=} \lim_{N\to\infty} \frac{Z}{\sqrt{N}} \sqrt{\langle m^0, m^0\rangle \langle q^0, q^0\rangle} \stackrel{\text{a.s.}}{=} 0.\end{aligned}$$

(b) This proof uses again (3.35) and is very similar to the proof of $\mathcal{B}_0(b)$. First we need to control the error term $\vec{o}_1(1)\tilde{q}^0 = \vec{o}_1(1) q^0$. In other words we need to show

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

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

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

Using Cauchy-Schwartz inequality

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N |\phi_h(a_i) - \phi_h(c_i)| \\ &\le L \max\left(\frac{1}{N}\sum_{i=1}^N \|a_i\|^{2k-2}, \frac{1}{N}\sum_{i=1}^N \|c_i\|^{2k-2} \right)^{1/2} \\ &\quad \times \langle q^0, q^0\rangle^{1/2} \, \vec{o}_1(1).\end{aligned}$$

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

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

which is bounded using part (e) and the original assumption on x_0. Similarly, using $\frac{1}{N}\sum_{i=1}^N \|q_i^0\|^{2k-2} < \infty$, we obtain $\frac{1}{N}\sum_{i=1}^N \|c_i\|^{2k-2} < \infty$.

Thus, from here we consider $\tilde{h}^1|_{\mathfrak{S}_{1,0}} \equiv \tilde{A}^* m^0$ whose components are distributed as $\|m^0\| Z/\sqrt{n}$ for Z a standard normal random variable, and will follow the steps taken in $\mathcal{B}_0(b)$. Conditionally on $\mathfrak{S}_{1,0}$, we can apply Theorem 3 to get

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

Then, using Lemma 4 for $v = x_0$ and $\psi(x_{0,i}) = \mathbb{E}_{\tilde{A}}\{\phi_h(\tilde{h}_i^1, x_{0,i})\}$, we obtain

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

The last equality used $\mathcal{B}_0(c)$.

(d) Using $\mathcal{H}_1(b)$ for $\phi_h(x, x_{0,i}) = x\varphi(x, x_{0,i})$ we obtain $\lim_{N\to\infty}\langle h^1, \varphi(h^1, x_0)\rangle \overset{\text{a.s.}}{=} \mathbb{E}\{\tau_0 Z\varphi(\tau_0 Z, X_0)\}$, 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 that $\lim_{N\to\infty}\langle h^1, h^1\rangle \overset{\text{a.s.}}{=} \tau_0^2$.
By part (b) the empirical distribution of (h^1, x_0) (i.e., the probability distribution on $\mathbb{R}^2$ that puts mass $1/N$ on each point $(h_i^1, x_{0,i})$, $i \in [N]$) converges weakly to $(\tau_0 Z, X_0)$. By applying Lemma 5 to the Lipschitz function φ, we get $\lim_{N\to\infty}\langle\varphi'(h^1, x_0)\rangle \overset{\text{a.s.}}{=} \mathbb{E}\{\varphi'(\tau_0 Z, X_0)\}$.
Since $t = 0$, and $q^0 = q_\perp^0$ then the result follows from (3.4) and that $\sigma_0^2 > 0$.

Step 3: $\mathcal{B}_t$: This part is analogous to step 1 albeit more complex.

First we prove (g).

(g) Note that using induction hypothesis $\mathcal{B}_{t-1}(b)$ for $\phi_b(b_i^r, b_i^s, w_i) = g_r(b_i^r, w_i)g_s(b_i^s, w_i)$, $0 \le r,\ s \le t-1$ we have almost surely

$$\lim_{n\to\infty}\langle m^r, m^s\rangle = \mathbb{E}\left\{g_r(\sigma_r\hat{Z}_r, W)\, g_s(\sigma_s\hat{Z}_s, W)\right\}. \quad (3.36)$$

On the other hand

$$\langle m_\perp^{t-1}, m_\perp^{t-1}\rangle = \langle m^{t-1}, m^{t-1}\rangle - \frac{(m^{t-1})^* M_{t-1}}{n}\left[\frac{M_{t-1}^* M_{t-1}}{n}\right]^{-1}\frac{M_{t-1}^* m^{t-1}}{n}. \quad (3.37)$$

But using induction hypothesis, we have $\lim_{n\to\infty}\langle m_\perp^r, m_\perp^r\rangle > \varsigma_r > 0$ for all $r < t-1$. So using Lemma 8, for large enough n the smallest eigenvalue of matrix $M_{t-1}^* M_{t-1}/n$ is larger than a positive constant c' that is independent of n. Hence, by Lemma 9 its inverse converges to an invertible limit. Thus, (3.36) and (3.37) lead to

$$\lim_{n\to\infty}\langle m_\perp^{t-1}, m_\perp^{t-1}\rangle \overset{\text{a.s.}}{=} \mathbb{E}\left\{[g_{t-1}(\sigma_{t-1}\hat{Z}_{t-1}, W)]^2\right\} - u^* C^{-1} u \quad (3.38)$$

with $u \in \mathbb{R}^{(t-1)}$ and $C \in \mathbb{R}^{(t-1)\times(t-1)}$ such that for $1 \le r,\ s \le t-1$

$$u_r = \mathbb{E}\left\{g_{r-1}(\sigma_{r-1}\hat{Z}_{r-1}, W)\, g_{t-1}(\sigma_{t-1}\hat{Z}_{t-1}, W)\right\},$$
$$C_{rs} = \mathbb{E}\left\{g_{r-1}(\sigma_{r-1}\hat{Z}_{r-1}, W)\, g_{s-1}(\sigma_{s-1}\hat{Z}_{s-1}, W)\right\}.$$

Now the result follows from Lemma 7 provided that we show for Gaussian random variables $\sigma_0\hat{Z}_0, \ldots, \sigma_{t-1}\hat{Z}_{t-1}$, all conditional variances $\mathrm{Var}[\sigma_r\hat{Z}_r|\ \sigma_0\hat{Z}_0, \ldots, \sigma_{r-1}\hat{Z}_{r-1}]$ are strictly positive for $r = 0, \ldots, t-1$. To prove the latter first using the induction hypothesis $\mathcal{B}_{t-1}(b)$, we have for all $0 \le r \le t-1$

$$\begin{aligned}\lim_{n\to\infty}&\langle b_\perp^r, b_\perp^r\rangle\\ &= \lim_{n\to\infty}\left(\langle b^r, b^r\rangle - \frac{(b^r)^* B_r}{n}\left[\frac{B_r^* B_r}{n}\right]^{-1}\frac{B_r^* b^r}{n}\right)\\ &\overset{\text{a.s.}}{=} \mathrm{Var}[\sigma_r\hat{Z}_r|\sigma_0\hat{Z}_0, \ldots, \sigma_{r-1}\hat{Z}_{r-1}].\end{aligned}$$

Similar as above we used the fact that for large enough n the matrix $B_r^* B_r/n$ has a smallest eigenvalue greater than a positive constant to obtain the limit of its inverse. On the other hand using induction hypothesis $\mathcal{B}_r(c)$ we have almost surely

$$\begin{aligned}\lim_{n\to\infty}&\langle b_\perp^r, b_\perp^r\rangle\\ &= \lim_{n\to\infty}\left(\langle b^r, b^r\rangle - \frac{(b^r)^* B_r}{n}\left[\frac{B_r^* B_r}{n}\right]^{-1}\frac{B_r^* b^r}{n}\right)\\ &= \frac{1}{\delta}\lim_{N\to\infty}\left(\langle q^r, q^r\rangle - \frac{(q^r)^* Q_r}{N}\left[\frac{Q_r^* Q_r}{N}\right]^{-1}\frac{Q_r^* q^r}{N}\right)\\ &= \frac{1}{\delta}\lim_{N\to\infty}\langle q_\perp^r, q_\perp^r\rangle. \end{aligned}\quad (3.39)$$

And now by induction hypothesis $\mathcal{H}_r(g)$ we have $\lim_{N\to\infty}\langle q_\perp^r, q_\perp^r\rangle > \rho_r$. Hence the result follows.

Corollary 2: The vectors

$$\vec{\alpha} = (\alpha_0, \ldots, \alpha_{t-1}) = \left[\frac{M_t^* M_t}{n}\right]^{-1}\frac{M_t^*\, m^t}{n}$$
$$\vec{\beta} = (\beta_0, \ldots, \beta_{t-1}) = \left[\frac{Q_t^* Q_t}{N}\right]^{-1}\frac{Q_t^*\, q^t}{N}$$

have finite limits as n and N converge to ∞.

Proof: We can apply Lemma 8 to obtain that for large enough n the smallest eigenvalue of $M_t^* M_t/n$ is larger than a positive constant c'. Hence by Lemma 9 its inverse has a finite limit. Similarly, we can apply induction hypothesis $\mathcal{H}_t(g)$ and Lemmas 8 and 9 to the matrix $Q_t^* Q_t/N$. □

(a) Recall definition of Y_t and X_t from Section III-E.

$$X_t = H_t + Q_t\Xi_t, \quad Y_t = B_t + [0|M_{t-1}]\Lambda_t \quad (3.40)$$

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

Lemma 13: The following holds

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

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

Proof: In light of Lemmas 10 and 12, we have

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

Now using (3.40), we only need to show

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

Recall that $m_{\|}^t = M_t\vec{\alpha}$ and $q_{\|}^t = Q_t\vec{\beta}$. On the other hand $Y_{t+1}^* m_\perp^t = B_{t+1}^* m_\perp^t$ because $M_t^* m_\perp^t = 0$. Similarly, $X_t^* q_\perp^t = H_t^* q_\perp^t$. Hence, we need to show

$$\begin{aligned} &Q_t\Xi_t\vec{\alpha} + Q_{t+1}(Q_{t+1}^*Q_{t+1})^{-1}B_{t+1}^* m_\perp^t - \xi_t q^t \\ &\quad = Q_t\vec{o}_t(1) \qquad (3.41)\\ &[0|M_{t-1}]\Lambda_t\vec{\beta} + M_t(M_t^*M_t)^{-1}H_t^* q_\perp^t - \lambda_t m^{t-1} \\ &\quad = M_t\vec{o}_t(1). \qquad (3.42)\end{aligned}$$

Here is our strategy to prove (3.42) [proof of (3.41) is similar]. The left-hand side (LHS) is a linear combination of vectors $m^0,\dots,m^{t-1}$. For any $\ell = 1,\dots,t$ we will prove that the coefficient of $m^{\ell-1}\in\mathbb{R}^n$ converges to 0. This coefficient in the LHS is equal to

$$\begin{aligned}&\left[(M_t^*M_t)^{-1}H_t^*q_\perp^t\right]_\ell - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}} \\ &\quad = \sum_{r=1}^t \left[\left(\frac{M_t^*M_t}{n}\right)^{-1}\right]_{\ell,r} \frac{\langle h^r, q^t - \sum_{s=0}^{t-1}\beta_s q^s\rangle}{\delta} \\ &\quad - \lambda_\ell(-\beta_\ell)^{\mathbb{I}_{\ell\neq t}}.\end{aligned}$$

To simplify the notation denote the matrix $M_t^*M_t/n$ by G. Therefore

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

But using the induction hypothesis $\mathcal{H}_t(d)$ for $\varphi = f_1,\dots,f_t$, and $\mathcal{H}_t(f)$, 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

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

Notice that the above series of equalities hold because G has, almost surely, a nonsingular limit as $N\to\infty$ as shown in point (g) above. Equation (3.41) is proved analogously, using $\xi_t = \langle g'(b^t, w)\rangle$. ☐

The proof of (3.15) follows immediately since the last lemma yields

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

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

$$M_t(M_t^*M_t)^{-1}M_t^*\tilde{A}q_\perp^t \overset{\text{d}}{=} \tilde{M}_t\vec{o}_t(1)$$

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 the induction hypothesis. For $s = t,\ r < t$, we can apply Lemma 13 to b^t (proved above), thus obtaining

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

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 $\mathcal{B}_{t-1}(c)$ 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 $\mathcal{B}_{t-1}(c)$ for each term of the form $\langle b^i, b^r\rangle$ (noting that $i,\ r\le t-1$) and Corollary 2 we have

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

The 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

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

The contribution of other terms is $o(1)$ because

- $\langle P_{M_t}^\perp\tilde{A}q_\perp^t, M_t\vec{o}_t(1)\rangle = \langle\tilde{A}q_\perp^t, P_{M_t}^\perp M_t\vec{o}_t(1)\rangle = 0$.
- $\langle\sum_{i=0}^{t-1}\beta_i b^i, M_t\vec{o}_t(1)\rangle = o(1)$, using Corollary 2 and induction hypothesis $\mathcal{B}_{t-1}(d)$ for $\varphi = g_j$.
- $\langle\sum_{i=0}^{t-1}\beta_i b^i, P_{M_t}^\perp\tilde{A}q_\perp^t\rangle = o(1)$ follows from Lemma 2 and Corollary 2.

The arguments at the last two points are completely analogous to the one carried out in the case $s = t,\ r < t$ above.

Now, using Lemma 2

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

Hence, from the induction hypothesis $\mathcal{B}_{t-1}(c)$

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

(e) Conditioning on $\mathfrak{S}_{t,t}$ and using Lemma 13 (proved at point (a) above), almost surely

$$\begin{aligned}\sum_{i=1}^n\frac{1}{n}(b^t_i)^{2\ell} \le{}& \frac{C}{n}\sum_{i=1}^n\left(\sum_{r=0}^{t-1}\beta_r b^r_i\right)^{2\ell} \\ &+\frac{C}{n}\sum_{i=1}^n([P^{\perp}_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell} + o(1)\frac{C}{n}\sum_{r=0}^{t-1}\sum_{i=1}^n([m^r]_i)^{2\ell}\end{aligned}$$

for some constant $C = C(\ell, t) < \infty$. We will bound each of the above summands.

— The term $n^{-1}\sum_{i=1}^n(\sum_{r=0}^{t-1}\beta_r b^r_i)^{2\ell}$ is finite since we can write

$$\frac{1}{n}\sum_{i=1}^n\left(\sum_{r=0}^{t-1}\beta_r b^r_i\right)^{2\ell} = O\left(\sum_{r=0}^{t-1}\beta_r^{2\ell}\frac{1}{n}\sum_{i=1}^n(b^r_i)^{2\ell}\right)$$

then use Corollary 2 and induction hypothesis $\mathcal{B}_{t-1}(e)$ for each of $n^{-1}\sum_{i=1}^n(b^r_i)^{2\ell}$.

— For the term $n^{-1}\sum_{i=1}^n([m^r]_i)^{2\ell}$ we use

$$(m^r_i)^2 = g_r(b^r_i, w_{i,0})^2 = O\Big((b^r_i)^2 + w_{i,0}^2 + g(0,0)^2\Big)$$

that follows from the Lipschitz assumption on g_r. Thus

$$\frac{1}{n}\sum_{i=1}^n(m^r_i)^{2\ell} = O\left(\frac{1}{n}\sum_{i=1}^n(b^r_i)^{2\ell} + \frac{1}{n}\sum_{i=1}^n w_{i,0}^{2\ell} + g(0,0)^{2\ell}\right) \tag{3.43}$$

which has a finite limit almost surely, using the induction hypothesis $\mathcal{B}_{t-1}(e)$ and the assumption on w.

— The term $n^{-1}\sum_{i=1}^n([P^{\perp}_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell}$ can be written as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n([P^{\perp}_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell} ={}& O\left(\frac{1}{n}\sum_{i=1}^n([\tilde{A}q^t_{\perp}]_i)^{2\ell}\right) \\ &+O\left(\frac{1}{n}\sum_{i=1}^n([P_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell}\right).\end{aligned}$$

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

Finally, for $n^{-1}\sum_{i=1}^n([P_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell}$ using Lemma 2 and Corollary 2 we can write

$$\begin{aligned}P_{M_t}&\tilde{A}q^t_{\perp} \\ &\overset{\text{d}}{=} M_t\left[\frac{M_t^*M_t}{n}\right]^{-1} \\ &\quad\times\left[\frac{Z_0\|m^0\|\|q^t_{\perp}\|}{n\sqrt{n}}\Big|\cdots\Big|\frac{Z_{t-1}\|m^{t-1}\|\|q^t_{\perp}\|}{n\sqrt{n}}\right]^* \\ &= \frac{1}{\sqrt{n}}\sum_{r=0}^{t-1}c_r m^r Z_r\end{aligned}$$

where $Z_0,\dots,Z_{t-1}$ are iid with distribution $\mathsf{N}(0,1)$. and $c_0,\dots,c_r$ are almost surely bounded for all N large enough. Therefore, almost surely,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n([P_{M_t}\tilde{A}q^t_{\perp}]_i)^{2\ell} &\overset{\text{d}}{=} \frac{1}{n}\sum_{i=1}^n\left(\frac{1}{\sqrt{n}}\sum_{r=0}^{t-1}c_r m^r_i Z_r\right)^{2\ell} \\ &\le C\sum_{r=0}^{t-1}\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{\sqrt{n}}m^r_i Z_r\right)^{2\ell} \\ &\le C'\sum_{r=0}^{t-1}\frac{1}{n}\sum_{i=1}^n(m^r_i)^{2\ell}.\end{aligned}$$

Now each term is finite using the same argument as in (3.43).

Using part (a) we can write

$$\begin{aligned}\phi_b(b^0_i,\dots,b^t_i,w_i)|_{\mathfrak{S}_{t,t}} &\overset{\text{d}}{=} \\ \phi_b&\left(b^0_i,\dots,b^{t-1}_i,\left[\sum_{r=0}^{t-1}\beta_r b^r + \tilde{A}q^t_{\perp} + \tilde{M}_t\vec{o}_t(1)\right]_i, w_i\right).\end{aligned}$$

Similar to the proof of $\mathcal{H}_1(b)$ we can drop the error term $\tilde{M}_t\vec{o}_t(1)$. Indeed, defining

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

by the pseudo-Lipschitz assumption

$$\begin{aligned}|\phi_b(a_i) - \phi_b(c_i)| \\ \le L\Big\{1 + \max\left(\|a_i\|^{k-1}, \|c_i\|^{k-1}\right)\Big\}\big|[\tilde{M}_t\vec{o}_t(1)]_i\big|.\end{aligned}$$

Therefore, using Cauchy-Schwartz inequality twice, we have

$$\frac{1}{n}\left|\sum_{i=1}^{n}\phi_b(a_i)-\sum_{i=1}^{n}\phi_b(c_i)\right| \leq L\left[\max\left(\frac{\sum_{i=1}^{n}\|a_i\|^{2k-2}}{n},\frac{\sum_{i=1}^{n}\|c_i\|^{2k-2}}{n}\right)\right]^{\frac{1}{2}} \times\left[\sum_{r=0}^{t-1}\langle\tilde{m}^r,\tilde{m}^r\rangle\right]^{\frac{1}{2}}\vec{o}_1(1). \quad (3.44)$$

Also note that

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

which is finite almost surely using the induction hypothesis $\mathcal{B}_t(e)$ proved above and the assumption on w. The term $n^{-1}\sum_{i=1}^{n}\|c_i\|^{2\ell}$ is bounded almost surely since

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|c_i\|^{2\ell} &\leq \frac{C}{n}\sum_{i=1}^{n}\|a_i\|^{2\ell}+C\sum_{r=0}^{t-1}\frac{1}{n}\sum_{i=1}^{n}(\tilde{m}^r)^{2\ell}\vec{o}_1(1)\\ &\leq \frac{C}{n}\sum_{i=1}^{n}\|a_i\|^{2\ell}+C'\sum_{r=0}^{t-1}\frac{1}{n}\sum_{i=1}^{n}(m^r)^{2\ell}\vec{o}_1(1)\end{aligned}$$

where the last inequality follows from the fact that $[M_t^*M_t/n]$ has almost surely a nonsingular limit as $N\to\infty$, as proved in point (g) above. Finally, for $r\leq t-1$, each term $(1/n)\sum_{i=1}^{n}(\tilde{m}^r)^{2\ell}$ is bounded using the induction hypothesis $\mathcal{B}_{t-1}(e)$, and the argument in (3.43).

Hence for any fixed t, (3.44) vanishes almost surely when n goes to ∞.

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

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

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

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left[\phi_b\left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r+\tilde{A}q_\perp^t\right]_i,w_i\right)-\mathbb{E}_{\tilde{A}}\left\{\phi_b\left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r+\tilde{A}q_\perp^t\right]_i,w_i\right)\right\}\right]\overset{\text{a.s.}}{=}0. \quad (3.45)$$

Note that $[\tilde{A}q_\perp^t]_i$ is a Gaussian random variable with variance $\|q_\perp^t\|^2/n$. Further $\|q_\perp^t\|^2/n$ converges to a finite limit γ_t^2 almost surely as $N\to\infty$. Indeed $\|q_\perp^t\|^2/N=\|q^t\|^2/N-\|q_\|^t\|^2/N$. By induction hypthesis $\mathcal{H}_t(b)$ applied to the pseudo-Lipshitz function $\phi_h(h_i^t,x_{0,i})=f_t(h_i^t,x_{0,i})^2$, $\|q^t\|^2/N=\langle f_t(h^t,x_0),f_t(h^t,x_0)\rangle$ converges to a finite limit. Further $\|q_\|^t\|^2/N=\sum_{r,s=0}^{t-1}\beta_r\beta_r\langle q^r,q^s\rangle$ also converges since the products $\langle q^r,q^s\rangle$ do and the coefficients β_r, $r\leq t-1$ converge by Corollary 2.

Hence we can use induction hypothesis $\mathcal{B}_{t-1}(b)$ and Corollary 2 for

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

where Z is an independent $\mathsf{N}(0,1)$ random variable to show (3.46) at the bottom of the page. Note that $\sum_{r=0}^{t-1}\beta_r\sigma_r Z_r+\gamma_t Z$ is Gaussian. All that we need, is to show that the variance of this Gaussian is σ_t^2. But using a combination of (3.45) and (3.46) for the pseudo-Lipschitz function $\phi_b(y_0,\ldots,y_t,w_i)=y_t^2$

$$\lim_{n\to\infty}\langle b^t,b^t\rangle\overset{\text{a.s.}}{=}\mathbb{E}\left\{\left(\sum_{r=0}^{t-1}\beta_r\sigma_r Z_r+\gamma_t Z\right)^2\right\}. \quad (3.47)$$

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 $\phi_h(y_0,\ldots,y_t,x_{0,i})=f(y_t,x_{0,i})^2$ we get $\lim_{n\to\infty}\delta^{-1}\langle f(h^t,x_0),f(h^t,x_0)\rangle\overset{\text{a.s.}}{=}\delta^{-1}\mathbb{E}\left\{f(\tau_{t-1}Z,X_0)^2\right\}$. So by definition (3.5), both sides of (3.47) are equal to σ_t^2.

(d) In a manner very similar to the proof of $\mathcal{B}_0(d)$, using part (b) for the pseudo-Lipschitz function $\phi_b:\mathbb{R}^{t+2}\to\mathbb{R}$

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\mathbb{E}_{\tilde{A}}\left\{\phi_b\left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\beta_r b^r+\tilde{A}q_\perp^t\right]_i,w_i\right)\right\}}{n}\overset{\text{a.s.}}{=}\mathbb{E}\,\mathbb{E}_Z\left\{\phi_b\left(\sigma_0 Z_0,\ldots\sigma_{t-1}Z_{t-1},\sum_{r=0}^{t-1}\beta_r\sigma_r Z_r+\gamma_t Z,W\right)\right\}. \quad (3.46)$$

that is given by $\phi_b(y_0,\dots,y_t,w_i) = y_t\varphi(y_s,w_i)$ we can obtain

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

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 $\lim_{n\to\infty}\langle b^t,b^s\rangle$. Similar to $\mathcal{B}_0(d)$ we can use Lemma 5 to transform $\mathbb{E}\{\varphi'(\sigma_s\hat{Z}_s,W)\}$ to $\lim_{n\to\infty}\langle\varphi'(b^t,w)\rangle$ almost surely. This finishes the proof of (d).

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.

(g) This part is very similar to the one of $\mathcal{B}_t(g)$.

(a) To prove (3.14) we use Lemma 13(a) as for $\mathcal{B}_t(a)$ to obtain

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

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

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

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

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

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

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

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

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

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

Where we used the definition of α_i and $m_\perp^t\perp m^r$.

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

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

Note that we used similar argument as in proof of $\mathcal{B}(c)$ 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

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

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

(f) Using $\mathcal{H}_t(a)$ and Lemma 2(a) we have almost surely

$$\lim_{N\to\infty}\langle h^{t+1},q^0\rangle \overset{\text{d}}{=} \lim_{N\to\infty}\frac{Z\|m_\perp^t\|\|q^0\|}{\sqrt{n}N} + \sum_{i=0}^{t-1}\lim_{N\to\infty}\alpha_i\langle h^{i+1},q^0\rangle.$$

But this limit is 0 almost surely, using the induction hypothesis $\mathcal{H}_r(e)$ for $r<t$ and $\mathcal{B}_t(c)$.

(b) Using part (a) we can write

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

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 (3.48), shown at the bottom of the next page. 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

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

where Z is an independent $\mathsf{N}(0,1)$ random variable, to show (3.49) at the bottom of the page. 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 (3.48) and (3.49) for the pseudo-Lipschitz function $\phi_h(y_0, \dots, y_t, x_{0,i}) = y_t^2$

$$\lim_{N\to\infty} \langle h^{t+1}, h^{t+1} \rangle \overset{\text{a.s.}}{=} \mathbb{E}\left\{ \left(\sum_{r=0}^{t-1} \alpha_r \tau_r Z_r + \frac{\|m_\perp^t\| Z}{\sqrt{n}} \right)^2 \right\}. \tag{3.50}$$

On the other hand in part (c) we proved $\lim_{N\to\infty} \langle h^{t+1}, h^{t+1} \rangle \overset{\text{a.s.}}{=} \lim_{N\to\infty} \langle g_t(b^t, w), g_t(b^t, w) \rangle$. 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 $\lim_{n\to\infty} \langle g_t(b^t, w), g_t(b^t, w) \rangle \overset{\text{a.s.}}{=} \mathbb{E}\{g_t(\sigma_t Z, W)^2\}$. So by the definition (1.4), both sides of (3.50) 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 $\lim_{N\to\infty} \langle h^{t+1}, h^{s+1} \rangle$. Similar to $\mathcal{H}_1(d)$ using Lemma 5, $\mathbb{E}\{\varphi'(\tau_s Z_s, X_0)\}$ can be transformed to $\lim_{N\to\infty} \langle \varphi'(h^{t+1}, x_0) \rangle$ almost surely. This finishes the proof of (d).

J. *Proof of Corollary 1*

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

$$\psi(x_1, \dots, x_\ell, y_1, \dots, y_\ell) = \psi_1(x_1, y_1) \cdots \psi_\ell(x_\ell, y_\ell), \tag{3.51}$$

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

$$\begin{aligned} &\mathsf{E}\, \psi(x_{J(1)}^t, \dots, x_{J(\ell)}^t, x_{0,J(1)}, \dots, x_{0,J(\ell)}) \\ &\quad = \prod_{s=1}^{\ell} \left(\frac{1}{N} \sum_{i=1}^N \psi_s(x_i^t, x_{0,i}) \right) + O\left(\frac{1}{N}\right). \end{aligned} \tag{3.52}$$

IV. Symmetric Case

Let $k \geq 2$, $G = A^* + A$ with $A \in \mathbb{R}^{N\times N}$, and assume that the entries of A are i.i.d. $\mathsf{N}(0, (2N)^{-1})$. Also let $f : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function. Start with m^0 and m^1 in $\mathbb{R}^N$ where $m^0 = 0_{N\times 1}$ and m^1 is a fixed deterministic vector in $\mathbb{R}^N$ with $\limsup_{N\to\infty} N^{-1} \sum_{i=1}^N (m_{1,i})^{2k-2} < \infty$, and proceed by the following iteration

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

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 i.i.d. entries $A_{ij} \sim \mathsf{N}(0, 1/(2N)^{-1})$. Then, for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order k and all $t \in \mathbb{N}$, almost surely

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

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \Bigg(&\phi_h\left(h_i^1, \dots, h_i^t, \left[\sum_{r=0}^{t-1} \alpha_r h^{r+1} + \tilde{A}^* m_\perp^t \right]_i, x_{0,i} \right) \\ &- \mathbb{E}_{\tilde{A}} \left\{ \phi_h\left(h_i^1, \dots, h_i^t, \left[\sum_{r=0}^{t-1} \alpha_r b^{r+1} + \tilde{A}^* m_\perp^t \right]_i, x_{0,i} \right) \right\} \Bigg) \overset{\text{a.s.}}{=} 0. \end{aligned} \tag{3.48}$$

$$\begin{aligned} \lim_{N\to\infty} &\frac{\sum_{i=1}^N \mathbb{E}_{\tilde{A}} \left\{ \phi_h\left(h_i^1, \dots, h_i^t, \left[\sum_{r=0}^{t-1} \alpha_r b^{r+1} + \tilde{A}^* m_\perp^t \right]_i, x_{i,0} \right) \right\}}{N} \\ &\overset{\text{a.s.}}{=} \mathbb{E}\, \mathbb{E}_Z \left\{ \phi_h\left(\tau_0 Z_0, \dots \tau_{t-1} Z_{t-1}, \sum_{r=0}^{t-1} \alpha_r \tau_r Z_r + \frac{\|m_\perp^t\| Z}{\sqrt{n}}, X_0 \right) \right\}. \end{aligned} \tag{3.49}$$

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 (4.1):

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

The analogue of Lemma 1 is the following.

Lemma 14: Let $\{A(N)\}_N$ be a sequence of random matrices as in Theorem 4. Then the following hold for all $t \in \mathbb{N}$

(a)

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

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 $\tilde{M}_t^* \tilde{M}_t = n\, \mathsf{I}_{t\times t}$. Recall that, $\vec{o}_t(1) \in \mathbb{R}^t$ 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 k,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \phi(h_i^2, \dots, h_i^{t+1}) \stackrel{\text{a.s.}}{=} \mathbb{E}\left[\phi(\tau_1 Z_1, \dots, \tau_t Z_t)\right] \quad (4.5)$$

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 have degenerate distribution (i.e., they are constant random variables):

$$\lim_{N\to\infty} \langle h^{r+1}, h^{s+1}\rangle \stackrel{\text{a.s.}}{=} \lim_{N\to\infty} \langle m^r, m^s\rangle \qquad (4.6)$$

(d) For all $1 \le r,\ s \le t$, and for any Lipschitz continuous function φ, the following equations hold and all limits exist, are bounded and have degenerate distribution (i.e., they are constant random variables):

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

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

(f) For all $0 \le r \le t$ the following limit exists and there are positive constants ρ_r (independent of N) such that almost surely

$$\lim_{N\to\infty} \langle m_\perp^r, m_\perp^r\rangle > \rho_r. \qquad (4.8)$$

Appendix A
AMP Algorithm: An Heuristic Derivation

In this appendix we present an heuristic derivation of the AMP iteration (1.1) starting from the standard message passing formulation (1.2). Let us stress that such derivation is not relevant for the proof of our Theorem 1. Our objective is to help the reader develop an intuitive understanding of the AMP iteration. For further discussion of the connection with belief propagation, we refer to [11] and [12].

Let us rewrite the message passing iteration for greater convenience of the reader

$$z_{a\to i}^t = y_a - \sum_{j\in[N]\setminus i} A_{aj} x_{j\to a}^t \qquad (A.1)$$

$$x_{i\to a}^{t+1} = \eta_t \left(\sum_{b\in[n]\setminus a} A_{bi} z_{b\to i}^t \right). \qquad (A.2)$$

Notice that on the RHS of both equations the messages appears in sums of $\Theta(N)$ terms. Consider for instance the messages $\{z_{a\to i}^t\}_{i\in[N]}$ for a fixed node $a \in [n]$. These depend on $i \in [N]$ only because the excluded term changes. It is therefore natural to guess that $z_{a\to i}^t = z_a^t + O(N^{-1/2})$ and $x_{i\to a}^t = x_i^t + O(n^{-1/2})$, where z_a^t only depends on the index a (and not on i), and x_i^t only depends on i (and not on a).

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

$$z_{a\to i}^t = z_a^t + \delta z_{a\to i}^t, \quad x_{i\to a}^t = x_i^t + \delta x_{i\to a}^t.$$

Substituting in (A.1), we get

$$\begin{aligned}
&z_a^t + \delta z_{a\to i}^t \\
&\quad = y_a - \sum_{j\in[N]} A_{aj}(x_j^t + \delta x_{j\to a}^t) \\
&\qquad + A_{ai}(x_i^t + \delta x_{i\to a}^t), \\
&x_i^{t+1} + \delta x_{i\to a}^{t+1} \\
&\quad = \eta_t \left(\sum_{b\in[n]} A_{bi}(z_b^t + \delta z_{b\to i}^t) - A_{ai}(z_a^t + \delta z_{a\to i}^t) \right).
\end{aligned}$$

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

$$\begin{aligned}
&z_a^t + \delta z_{a\to i}^t \\
&\quad = y_a - \sum_{j\in[N]} A_{aj}(x_j^t + \delta x_{j\to a}^t) + A_{ai} x_i^t, \\
&x_i^{t+1} + \delta x_{i\to a}^{t+1} \\
&\quad = \eta_t \left(\sum_{b\in[n]} A_{bi}(z_b^t + \delta z_{b\to i}^t) - A_{ai} z_a^t \right).
\end{aligned}$$

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

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

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

$$z^t_a = y_a - \sum_{j\in[N]} A_{aj}(x^t_j + \delta x^t_{j\to a}), \tag{A.3}$$

$$\delta z^t_{a\to i} = A_{ai}x^t_i. \tag{A.4}$$

Analogously for the second equation we get

$$x^{t+1}_i = \eta_t\left(\sum_{b\in[n]} A_{bi}(z^t_b + \delta z^t_{b\to i})\right), \tag{A.5}$$

$$\delta x^{t+1}_{i\to a} = -\eta'_t\left(\sum_{b\in[n]} A_{bi}(z^t_b + \delta z^t_{b\to i})\right) A_{ai}z^t_a. \tag{A.6}$$

Substituting (A.4) in (A.5) to eliminate $\delta z^t_{b\to i}$ we get

$$x^{t+1}_i = \eta_t\left(\sum_{b\in[n]} A_{bi}z^t_b + \sum_{b\in[n]} A^2_{bi}x^t_i\right) \tag{A.7}$$

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

$$x^{t+1} = \eta_t(x^t + A^* z^t). \tag{A.8}$$

Analogously substituting (A.6) in (A.3), we get

$$z^t_a = y_a - \sum_{j\in[N]} A_{aj}x^t_j + \sum_{j\in[N]} A^2_{aj}\eta'_t(x^t_j + (A^*z^t)_j)z^t_a. \tag{A.9}$$

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

$$\begin{aligned} &\sum_{j\in[N]} A^2_{aj}\eta'_t(x^t_j + (A^*z^t)_j) \\ &\quad \approx \frac{1}{n}\sum_{j\in[N]} \eta'_t(x^t_j + (A^*z^t)_j) \\ &\quad \to \frac{1}{\delta}\langle \eta'_t(x^t_j + (A^*z^t)_j)\rangle \end{aligned} \tag{A.10}$$

whence substituting in (A.9), we obtain the second equation in (1.1). This finishes our derivation.

Appendix B
Proof of Probability and Linear Algebra Lemmas

In this Appendix we provide proofs of two probability lemmas stated in Section III-G.

A) 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

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

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 $\psi_B(x)$, we have

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

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

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

From this we obtain

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

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

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{E}_{\hat{p}_{v(N)}}\left\{L(1+|V|^k)\mathbb{I}_{\{|V|^k\le\frac{B}{L}-1\}}\right\} \\ &\qquad = \mathbb{E}_{p_V}\left\{L(1+|V|^k)\mathbb{I}_{\{|V|^k\le\frac{B}{L}-1\}}\right\}. \end{aligned}$$

Therefore

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

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 $k^{\rm th}$ 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.

B) 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 $\{(X_n,Y_n)\}_{n\ge 1}$ and (X,Y) on this space, such that letting

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

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. Since F is Lipschitz continuous, $\mathcal{C}_F$ has full Lebesgue measure. Since the probability distribution of (X,Y) is absolutely continuous with respect to Lebesgue, $\mathcal{C}_F$ has measure 1 under this measure.

Hence if we let

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

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

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

Since F is Lipschitz $|Z_n(\omega)|\le C$, and hence the bounded convergence theorem implies $\mathbb{E}\{Z_n(\omega)\}\to\mathbb{E}\{Z(\omega)\}$ which proves our claim.

C) Proof of Lemma 7: Let us denote by Q the covariance of the Gaussian vector $Z_1,\dots,Z_t$. The set of matrices Q satisfying the constraints with constants $c_1,\dots,c_t$, K is compact. Hence if the thesis does not hold, there must exist a specific covariance matrix satisfying these constrains, and such that

$$\mathbb{E}\{[\ell(Z_t,Y)]^2\}-u^*C^{-1}u=0. \tag{B.2}$$

Fix Q to be such a matrix, and let $S\in\mathbb{R}^{t\times t}$ be the matrix with entries $S_{i,j}\equiv\mathbb{E}\{\ell(Z_i,Y)\ell(Z_j,Y)\}$. Then (B.2) implies that S is not invertible (by Schur complement formula). Therefore there exist nonvanishing constants $a_1,\dots,a_t$ such that

$$a_1\,\ell(Z_1,Y)+a_2\,\ell(Z_2,Y)+\dots+a_t\,\ell(Z_t,Y)\overset{\rm a.s.}{=}0. \tag{B.3}$$

The function $(z_1,\dots,z_t)\mapsto a_1\,\ell(z_1,Y)+\dots+a_t\,\ell(z_t,Y)$ is Lipschitz and nonconstant. Hence there is a set $\mathcal{A}\subseteq\mathbb{R}^t$ of positive Lebesgue measure such that it is nonvanishing on $\mathcal{A}$. Therefore, $\mathcal{A}$ must have zero measure under the law of $(Z_1,\dots,Z_t)$, i.e., $\lambda_{\min}(Q)=0$. This implies that there exists nonvanishing constants $a'_1,\dots,a'_t$ such that

$$a'_1\,Z_1+a'_2\,Z_2+\dots+a'_t\,Z_t\overset{\rm a.s.}{=}0.$$

If $t_*=\max\{i\in\{1,\dots,t\}:a'_i\neq 0\}$, this implies

$$Z_{t_*}\overset{\rm a.s.}{=}\sum_{i=1}^{t_*-1}\left(\frac{-a'_i}{a'_{t_*}}\right)Z_i,$$

which contradicts the assumption $\mathrm{Var}(Z_{t_*}|Z_1,\dots,Z_{t_*-1})>0$.

D) Proof of Lemma 8: We will prove the thesis by induction over t. The case $t=1$ is trivial, and assume that the claim is true up for any $(t-1)$ vectors $v_1,\dots,v_{t-1}$, with constant c'_{t-1}. Without loss of generality, we will assume $\|v_i\|^2/n\le K$ for some constant K independent of n (increasing the norm of the v_i's increases $\lambda_{\min}(C)$).

Let $V\in\mathbb{R}^{n\times t}$ be the matrix with columns $v_1,\dots,v_t$. Then $C=V^*V/n$. By Gram-Schmidt orthonormalization, we can construct A upper triangular, and $U\in\mathbb{R}^{n\times t}$ orthonormal (i.e., with $U^*U=\mathsf{I}_{t\times t}$) such that

$$U=VA.$$

It follows that

$$\begin{aligned}\lambda_{\min}(C)&=\frac{1}{n}\,\lambda_{\min}(V^*V)=\frac{1}{n}\,\lambda_{\min}((A^{-1})^*A^{-1})\\&=\frac{1}{n}\,\lambda_{\max}(AA^*)^{-1}=\frac{1}{n}\,\sigma_{\max}(A)^{-2}.\end{aligned} \tag{B.4}$$

Defining u_i to be the columns of U, Gram-Schmidt orthonormalization prescribes

$$u_i=\frac{v_i-P_{i-1}(v_i)}{\|v_i-P_{i-1}(v_i)\|}.$$

Which implies $A_{ii}=\|v_i-P_{i-1}(v_i)\|^{-1}\le(cn)^{-1/2}$ and

$$A_{ji}=-\frac{1}{\|v_i-P_{i-1}(v_i)\|}(\tilde V_{i-1}^*\tilde V_{i-1})^{-1}\tilde V_{i-1}^*v_i.$$

We then have

$$\begin{aligned}|A_{ji}|&\le(cn)^{-1/2}\lambda_{\min}(\tilde V_{i-1}^*\tilde V_{i-1})^{-1}(i-1)Kn\\&\le t(cn)^{-1/2}(c'_{t-1}n)^{-1}Kn\le c'n^{-1/2}.\end{aligned}$$

It follows that $\sigma_{\max}(A)\le c'''n^{-1/2}$ (with c''' depending on n) whence the thesis follows by (B.4).

ACKNOWLEDGMENT

The authors are deeply indebted to E. Bolthausen, who first presented the conditioning technique during the EURANDOM workshop on "Order, disorder and double disorder" in September 2009. They are also grateful to D. Donoho and A. Maleki for a number of insightful exchanges.

REFERENCES

[1] D. Aldous, "The $\zeta(2)$ limit in the random assignment problem," *Random Struct. Algorith.*, vol. 18, pp. 381–418, 2001.

[2] D. Aldous and J. M. Steele, "The objective method: Probabilistic combinatorial optimization and local weak convergence," in *Probability On Discrete Structures*, H. Kesten, Ed. New York: Springer-Verlag, 2003, pp. 1–72.

[3] M. Bayati and A. Montanari, The LASSO risk for gaussian matrices, 2010, preprint. [Online]. Available: http://arxiv.org/abs/1008.2581

[4] E. Bolthausen, On the high-temperature phase of the Sherrington-Kirkpatrick model, Seminar presented at EURANDOM. Eindhoven, The Netherlands, September 2009 [Online]. Available: http://www.eurandom.nl/events/workshops/2009/OD/Order/Bolthausen.pdf

[5] I. Benjamini and O. Schramm, "Percolation beyond Z^d, many questions and a few answers," *Electron. Commun. Probab.*, vol. 1, pp. 71–82, 1996.

[6] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Commun. Pure and Appl. Math.*, vol. 75, pp. 1412–1457, 2004.
[7] D. L. Donoho and I. M. Johnstone, “Ideal spatial adaptation via wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425–455, 1994.
[8] D. L. Donoho and I. M. Johnstone, “Minimax risk over l_p balls,” *Prob. Theory Rel. Fields*, vol. 99, pp. 277–303, 1994.
[9] D. L. Donoho and A. Maleki, “Optimally tuned iterative thresholding algorithm for compressed sensing,” *IEEE J. Sel. Areas Signal Process.* 2009 [Online]. Available: http://arxiv.org/abs/0909.0777, to be published
[10] D. L. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing,” *Proc. Nat. Acad. Sci.*, vol. 106, pp. 18914–18919, 2009.
[11] D. L. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing: I. Motivation and construction,” in *Proc. IEEE Inf. Theory Workshop*, Cairo, Egypt, 2010.
[12] D. L. Donoho, A. Maleki, and A. Montanari, “On the construction of approximate message passing algorithms”, to be published.
[13] D. L. Donoho, A. Maleki, and A. Montanari, The Noise-Sensitivity Phase Transition in Compressed Sensing [Online]. Available: http://arxiv.org/abs/1004. 1218
[14] D. Donoho, “For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution,” *Commun. Pure and Appl. Math.*, vol. 59, pp. 907–934, 2006.
[15] D. L. Donoho and J. Tanner, “Neighborliness of randomly-projected simplices in high dimensions,” *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9452–9457, 2005.
[16] D. L. Donoho and J. Tanner, “Counting faces of randomly projected polytopes when the projection radically lowers dimension,” *J. Amer. Math. Soc.*, vol. 22, pp. 1–53, 2009.
[17] D. Guo, D. Baron, and S. Shamai, “A single-letter characterization of optimal noisy compressed sensing,” in *Proc. 47th Ann. Allerton Conf.*, Monticello, IL, Sep. 2009.
[18] D. Guo and S. Verdu, “Randomly spread CDMA: Asymptotics via statistical physics,” *IEEE Trans. Inf. Theory*, vol. 51, pp. 1982–2010, 2005.
[19] T. C. Hu and R. L. Taylor, “Strong law for arrays and for the bootstrap mean and variance,” *Int. J. Math. and Math. Sci.*, vol. 20, pp. 375–383, 1997.
[20] Y. Kabashima, “A CDMA multiuser detection algorithm on the basis of belief propagation,” *J. Phys. A*, vol. 36, pp. 11111–11121, 2003.
[21] Y. Kabashima, T. Wadayama, and T. Tanaka, “A typical reconstruction limit for compressed sensing based on l_p-norm minimization,” *J. Stat. Mech.*, 2009.
[22] M. Mézard and A. Montanari, *Information, Physics and Computation*. Oxford, U.K.: Oxford Univ. Press, 2009.
[23] V. A. Marcenko and L. A. Pastur, “Distribution of eignevalues for some sets of random matrices,” *Math. USSR-Sbornik*, vol. 1, 1967.
[24] A. Montanari and D. Tse, “Analysis of belief propagation for non-linear problems: The example of CDMA (or: how to prove Tanaka’s formula),” in *Proc. IEEE Inf. Theory Workshop*, Punta de l’Este, Uruguay, 2006.
[25] J. P. Neirotti and D. Saad, “Improved message passing for inference in densely connected systems,” *Europhys. Lett.*, vol. 71, pp. 866–872, 2005.
[26] S. Rangan, A. K. Fletcher, and V. K. Goyal, “Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009, vol. 22, pp. 1545–1553.
[27] T. J. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[28] J. Salez and D. Shah, “Belief propagation: Optimal algorithm for random assignment problem,” *Math. Operat. Res.*, vol. 32, pp. 468–480, 2009.
[29] C. Stein, “A bound for the error in the normal approximation to the distribution of a sum of dependent random variables,” in *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, 1972.
[30] M. Talagrand, *Mean Field Models for Spin Glasses*. Berlin, Germany: Springer, 2010, Vol. I: Basic Examples.
[31] T. Tanaka, “A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors,” *IEEE Trans. Inf. Theory*, vol. 48, pp. 2888–2910, 2002.
[32] T. Tanaka, “Statistical physics of information processing in mobile communications,” *Progr. Theory Phys. Suppl.*, vol. 157, pp. 176–183, 2005.
[33] D. N. C. Tse and S. V. Hanly, “Linear multiuser receivers: Effective interference, effective bandwidth and user capacity,” *IEEE Trans. Inf. Theory*, vol. 45, pp. 641–657, 1999.
[34] S. Verdú and S. Shamai, “Spectral efficiency of CDMA with random spreading,” *IEEE Trans. Inf. Theory*, vol. 45, pp. 622–640, 1999.

Mohsen Bayati received the Ph.D. degree from Stanford University, Stanford, CA, in 2007. His dissertation was on optimization and modeling of large-scale networks.

During summer 2005 and 2006, he interned with IBM Research and Microsoft Research (MSR). He has been a Postdoctoral Researcher with MSR Redmond (2007–2008) and MSR New England (2008–2009) working mainly on applications of machine learning and optimization in healthcare and online advertising. He is currently a Postdoctoral Researcher at Stanford University.

Andrea Montanari 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 de l’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 the Departments of Electrical Engineering and of Statistics, Stanford University, Stanford, CA. Since 2010, he has been an Associate Professor in the same departments.

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.