Information and Inference: A Journal of the IMA (2019) **00**, 1–47
doi:10.1093/imaiai/iay021

State evolution for approximate message passing with non-separable functions

Raphaël Berthier
Département de Mathématiques et Applications, Ecole Normale Supérieure, Paris, 75005, France and Département de Mathématiques d'Orsay, Université Paris-Sud, Orsay, 91405, France

Andrea Montanari†
Department of Electrical Engineering and Department of Statistics, Stanford University, California, 94305, USA
†Corresponding author. Email: montanari@stanford.edu

and

Phan-Minh Nguyen
Department of Electrical Engineering, Stanford University, California, 94305, USA

[Received on 23 August 2017; revised on 26 July 2018; accepted on 5 September 2018]

Given a high-dimensional data matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, approximate message passing (AMP) algorithms construct sequences of vectors $\boldsymbol{u}^t \in \mathbb{R}^n$, $\boldsymbol{v}^t \in \mathbb{R}^m$, indexed by $t \in \{0,1,2\ldots\}$ by iteratively applying $\boldsymbol{A}$ or $\boldsymbol{A}^{\mathsf{T}}$ and suitable nonlinear functions, which depend on the specific application. Special instances of this approach have been developed—among other applications—for compressed sensing reconstruction, robust regression, Bayesian estimation, low-rank matrix recovery, phase retrieval and community detection in graphs. For certain classes of random matrices $\boldsymbol{A}$, AMP admits an asymptotically exact description in the high-dimensional limit $m,n \to \infty$, which goes under the name of *state evolution.* Earlier work established state evolution for separable nonlinearities (under certain regularity conditions). Nevertheless, empirical work demonstrated several important applications that require non-separable functions. In this paper we generalize state evolution to Lipschitz continuous non-separable nonlinearities, for Gaussian matrices $\boldsymbol{A}$. Our proof makes use of Bolthausen's conditioning technique along with several approximation arguments. In particular, we introduce a modified algorithm (called LoAMP for Long AMP), which is of independent interest.

Keywords: message passing; compressed sensing; statistical estimation; random matrices.

1. Introduction

Over the past few years approximate message passing (AMP) algorithms have been applied to a broad range of statistical estimation problems, including compressed sensing [18], robust regression [17], Bayesian estimation [25], low-rank matrix recovery [24], phase retrieval [42] and community detection in graphs [12]. In a fairly generic formulation,[1] AMP takes as input a random data matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and generates sequences of vectors $\boldsymbol{u}^t \in \mathbb{R}^n$, $\boldsymbol{v}^t \in \mathbb{R}^m$, indexed by $t \in \mathbb{N}$ according to the iteration

$$\boldsymbol{u}^{t+1} = \boldsymbol{A}^{\mathsf{T}} g_t(\boldsymbol{v}^t) - \mathsf{d}_t e_t(\boldsymbol{u}^t), \tag{1}$$

$$\boldsymbol{v}^t = \boldsymbol{A} e_t(\boldsymbol{u}^t) - \mathsf{b}_t g_{t-1}(\boldsymbol{v}^{t-1}). \tag{2}$$

[1] More general settings have also been developed, see for instance [21].

Here $g_t : \mathbb{R}^m \to \mathbb{R}^m$ and $e_t : \mathbb{R}^n \to \mathbb{R}^n$ are two sequences of functions indexed by the iteration number t that encode the specific application. The coefficients $\mathsf{d}_t, \mathsf{b}_t \in \mathbb{R}$ are completely fixed by the choice of these functions. For instance, assuming $\mathbb{E}\{A_{ij}^2\} = 1/m$, we can use

$$\mathsf{d}_t^{\text{emp}} = \frac{1}{m}\text{div}\, g_t(\boldsymbol{v}^t)\,, \quad \mathsf{b}_t^{\text{emp}} = \frac{1}{m}\text{div}\, e_t(\boldsymbol{u}^t). \tag{3}$$

(A slightly different definition, that is more convenient for proofs, will be adopted in Section 3.) Here $\text{div} f(\boldsymbol{x}) = \sum_i \frac{\partial f}{\partial x_i}(\boldsymbol{x})$ is the divergence operator.

As a concrete example, in applications to compressed sensing, AMP takes the form

$$\hat{\boldsymbol{\theta}}^{t+1} = \eta_t\big(\hat{\boldsymbol{\theta}}^t + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t\big), \tag{4}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\theta}}^t + \hat{\mathsf{b}}_t \boldsymbol{r}^{t-1}. \tag{5}$$

Here $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta}_0 + \boldsymbol{w}$ are noisy measurements of an unknown signal $\boldsymbol{\theta}_0$, $\boldsymbol{A}$ is a known sensing matrix, $\boldsymbol{w}$ is a noise vector and $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is a general denoiser. This iteration can be shown to be a special case of the above, via the mapping $\boldsymbol{u}^{t+1} = \boldsymbol{\theta}_0 - (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t + \hat{\boldsymbol{\theta}}^t)$, $\boldsymbol{v}^t = \boldsymbol{w} - \boldsymbol{r}^t$. While earlier work was limited to separable denoisers η_t, the results in this paper allow to treat a broad class of non-separable denoisers. We will devote Section 7 to such general compressed sensing applications.

Apart from being broadly applicable, AMP algorithms admit an asymptotically exact characterization in the high-dimensional limit $m, n \to \infty$ with m/n converging to a limit, which is known as *state evolution*. Informally, for any t fixed, in the high-dimensional limit, $\boldsymbol{u}^t$ is approximately Gaussian with mean zero and covariance $\tau_t^2 \boldsymbol{I}_n$, while $\boldsymbol{v}^t$ is approximately $\mathsf{N}(0, \sigma_t^2 \boldsymbol{I}_m)$. The variance parameters τ_t^2, σ_t^2 can be computed via a one-dimensional recursion.

State evolution was proved in [5] for the recursion (1) and (2) under the following two key assumptions:

- $\boldsymbol{A}$ a Gaussian random matrix with i.i.d. entries $(A_{ij})_{i \le m, j \le n} \sim \mathsf{N}(0, 1/m)$.
- The functions $g_t(\,\cdot\,), e_t(\,\cdot\,)$ are separable[2] and Lipschitz continuous.

This paper relaxes the second assumption and establishes state evolution for functions $g_t(\,\cdot\,), e_t(\,\cdot\,)$ that are Lipschitz continuous, but not necessarily separable. Our proof uses (as the original paper [5]) a conditioning technique initially developed by Bolthausen [7] to study the TAP equations in spin glass theory. A key difficulty with non-separable denoisers is that the iterates $g_1(\boldsymbol{v}^1), g_2(\boldsymbol{v}^2), \ldots, g_t(\boldsymbol{v}^t) \in \mathbb{R}^m$ might be collinear and lie in a subspace of dimension smaller than t, for large m. This degeneracy (or a similar problem with the $e_1(\boldsymbol{u}^1), e_2(\boldsymbol{u}^2), \ldots e_t(\boldsymbol{u}^t)$) would cause a naive adaptation of the proof of [5] to break down. In order to circumvent this problem without introducing *ad hoc* assumptions, we proceed in two steps:

1. We introduce a random perturbation of the functions $e_t(\,\cdot\,), g_t(\,\cdot\,)$. We prove that, with probability one with respect to this random perturbation, the new iteration satisfies the required non-degeneracy assumption.

[2] We say that $f : \mathbb{R}^d \to \mathbb{R}^d$ is separable if $f(x_1, \ldots, x_d)_i = f_i(x_i)$ for some functions $f_i : \mathbb{R} \to \mathbb{R}$.

2. We prove that both AMP and state evolution are uniformly continuous in the size of the perturbation, and hence we can let the perturbation vanish recovering state evolution for the original unperturbed problem.

Further, we obtain a streamlined proof with respect to the strategy of [5], by introducing a different algorithm, that we call LoAMP (Long AMP). State evolution is proved first for LoAMP, and then the latter is shown to be closely approximated by the original AMP. We believe that LoAMP is potentially of independent interest and will be further investigated in [27].

In the rest of this introduction we will briefly describe two applications of AMP with non-separable nonlinearities, and show how state evolution can be used to characterize its behavior. Both of these are examples of generalized compressed sensing, cf. Section 7. We will then review some related work in Section 2, and state our results in Section 3 (for the asymmetric iteration (1) and Section 4 (for the analogue case in which $\boldsymbol{A}$ is a random symmetric matrix)). Proofs are presented in Sections 5 and 6. In fact, we will first prove state evolution in the case in which $\boldsymbol{A}$ is a symmetric random matrix, and then reduce the asymmetric case to the symmetric one. Finally, Section 7 applies the general theory to compressed sensing reconstruction with a variety of denoisers. In particular, we derive a bound on the convergence rate for denoisers that are projectors onto convex sets. Several technical elements are deferred to the Appendices.

For a summary of notations used throughout the paper, the reader is urged to consult Section 5.1.

1.1 *Vignette #1: matrix compressed sensing*

We want to reconstruct an unknown matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n_1 \times n_2}$ from linear measurements $\boldsymbol{y} \in \mathbb{R}^m$, where

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_0). \tag{6}$$

Here $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a Gaussian linear operator. Concretely $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}_0 \rangle$ where $\boldsymbol{A}_i \in \mathbb{R}^{n_1 \times n_2}$ are i.i.d. matrices with independent entries $(\boldsymbol{A}_i)_{r,c} \sim \mathsf{N}(0, 1/m)$. This setting was first studied in [37] and can be used as a simple model for system identification and matrix completion.

When $\boldsymbol{X}_0$ has a low rank, the following AMP algorithm can be useful to reconstruct $\boldsymbol{X}_0$ from observations $\boldsymbol{y}$:

$$\boldsymbol{X}^{t+1} = \mathsf{S}\big(\boldsymbol{X}^t + \mathcal{A}^{\mathsf{T}} \boldsymbol{r}^t; \lambda_t\big), \tag{7}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \mathcal{A}\big(\boldsymbol{X}^t\big) + \mathsf{b}_t \boldsymbol{r}^{t-1}, \tag{8}$$

with initialization $\boldsymbol{X}^0 = \boldsymbol{0}$. After t iterations, the algorithm produces an estimate $\boldsymbol{X}^t$ and a residual $\boldsymbol{r}^t$. Here $\mathcal{A}^{\mathsf{T}}$ is the adjoint[3] of the operator $\mathcal{A}$,

$$\mathsf{b}_t = \frac{1}{m} \operatorname{div} \mathsf{S}\big(\boldsymbol{X}^{t-1} + \mathcal{A}^{\mathsf{T}} \boldsymbol{r}^{t-1}; \lambda_{t-1}\big), \tag{9}$$

[3] We can represent the action $\mathcal{A}(\boldsymbol{X})$ by vectorizing $\boldsymbol{X}$ as $\mathrm{vec}(\boldsymbol{X}) \in \mathbb{R}^n$, $n = n_1 n_2$. If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the matrix whose ith row is $\boldsymbol{A}_i$, then $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{A}\mathrm{vec}(\boldsymbol{X})$. Then the adjoint $\mathcal{A}^{\mathsf{T}}$ corresponds to the transpose $\boldsymbol{A}^{\mathsf{T}}$.

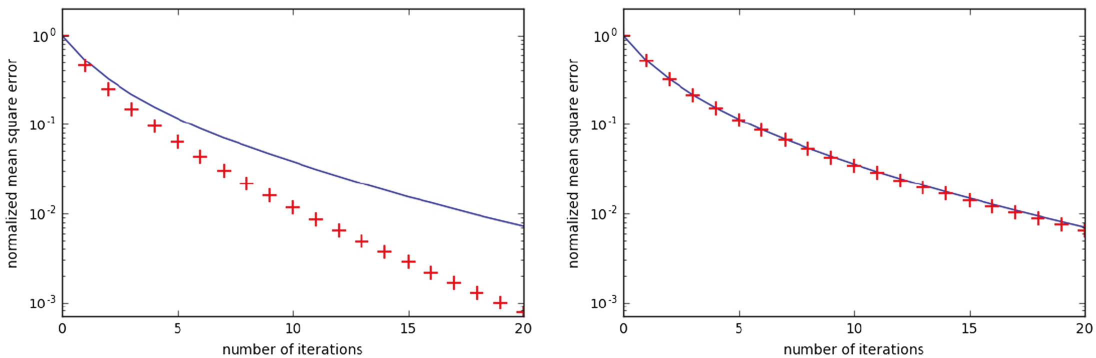

FIG. 1. Matrix compressed sensing reconstruction using AMP: normalized mean square error as a function of the number of iterations. Left: 30 × 30 matrices of rank 3. Right: 170 × 170 matrices of rank 17. Pluses (+): simulations. Solid line: state evolution prediction.

and $\mathsf{S}(\cdot\,;\lambda)$ is the singular value thresholding (SVT) operator, defined as follows. For a matrix $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$, with singular value decomposition

$$\boldsymbol{Y} = \sum_{i=1}^{n_1 \wedge n_2} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}, \tag{10}$$

the SVT operator yields

$$\mathsf{S}(\boldsymbol{Y};\lambda) = \sum_{i=1}^{n_1 \wedge n_2} (\sigma_i - \lambda)_+ \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}. \tag{11}$$

The divergence in Equation (9) can be computed explicitly using a formula from [11,14], see Appendix A.1. The sequence of parameters $(\lambda_t)_{t\geq 0}$ can be chosen to optimize the algorithm performance.

Fixed points of this AMP algorithm are minimum nuclear norm solution of the constraint $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{X})$. This algorithm was implemented in Donoho (personal communication) and partly motivated the predictions of [15]. A recent detailed study (and generalizations) can be found in [38], showing that its phase transition matches the one of nuclear norm minimization, predicted in [15] and proved in [1,34].

With a change of variables, the algorithm (7) and (8) can be recast in the general form (1) and (2) with one of the functions being non-separable and given by the SVT operator (the change of variables is described in Section 7).

In Fig. 1 we report the results of numerical simulations using this algorithm. We generated $\boldsymbol{X}_0 \in \mathbb{R}^{n_1 \times n_2}$ of rank r by letting $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}$ for $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$ uniformly random orthogonal matrices, and computed m measurements $\boldsymbol{y}$ as per Equation (6). We took $n_1 = n_2$, $r = 0.1 \cdot n_1$, $m = 0.65 \cdot n_1 n_2$. We chose the threshold parameter λ_t to be proportional to the noise level as estimated via the residual [28]:

$$\lambda_t = 2\sqrt{n_1}\,\frac{\|\boldsymbol{r}^t\|_2}{\sqrt{m}}. \tag{12}$$

We plot the normalized mean square error as a function of the iteration number (with $n = n_1 n_2$ the number of unknowns):

$$\mathrm{NMSE}(t;n) = \frac{\|\boldsymbol{X}^t - \boldsymbol{X}_0\|_F^2}{\|\boldsymbol{X}_0\|_F^2}\,. \tag{13}$$

State evolution allows to predict the value $\lim_{n\to\infty} \mathrm{NMSE}(t;n)$. The prediction is already very accurate for $n_1 = n_2 = 170$.

1.2 *Vignette #2: compressed sensing with images*

We represent an image as a two-dimensional array $\boldsymbol{x} = (x_{i,j})_{i\le n_1, j\le n_2}$, which we identify with its vectorization $\mathrm{vec}(\boldsymbol{x}) \in \mathbb{R}^n$, $n = n_1 n_2$. In compressed sensing we acquire a small number of incoherent measurements $\boldsymbol{y} \in \mathbb{R}^m$ according to

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \tag{14}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a known sensing matrix for which we assume the simple Gaussian model $(A_{ij})_{i\le m, j\le n} \sim_{iid} \mathsf{N}(0, 1/m)$ and $\boldsymbol{w} \sim \mathsf{N}(0, \sigma_w^2 \boldsymbol{I}_m)$ is noise.

A broad class of AMP reconstruction algorithms take the form

$$\boldsymbol{x}^{t+1} = \eta_t\big(\boldsymbol{x}^t + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t\big), \tag{15}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^t + \mathsf{b}_t \boldsymbol{r}^{t-1}, \tag{16}$$

where $\boldsymbol{x}^0 = 0$, $\eta_t : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ is a sequence of image denoisers, and

$$\mathsf{b}_t = \frac{1}{m}\mathrm{div}\, \eta_{t-1}\big(\boldsymbol{x}^{t-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^{t-1}\big). \tag{17}$$

The compressed sensing reconstruction algorithm in [**18**] was a special case of this iteration with $\eta_t(\,\cdot\,)$ corresponding to coordinate-wise soft thresholding (in a suitable basis), hence leading to a separable AMP. Several authors studied the same algorithm with non-separable denoisers, including Hidden Markov Models [**41**,**43**], total variation and block thresholding denoisers [**16**], universal denoising [**31**], and restricted Boltzmann machines [**46**]. As documented in these papers, a good choice of the denoiser yields a significant performance boost over classical compressed sensing reconstruction methods, such as ℓ_1 minimization. State-of-the-art performances were achieved in [**26**] using block-matching and 3-D filtering (BM3D) denoising [**13**].

Again, the iteration (16) and (15) can be put in the form (1) and (2) with a change of variables described in Section 7. A non-separable denoiser η_t translates into non-separable nonlinearities g_t, e_t.

Here we use Non-Local Means (NLM) denoising [**2**]. Given a noisy image $\boldsymbol{z}$, NLM estimates pixel (i, j) as a weighted average of the pixels of $\boldsymbol{z}$:

$$\eta(\boldsymbol{z})_{i,j} = \frac{\sum_{(k,l)} W_{(k,l),(i,j)}(\boldsymbol{z})\, z_{k,l}}{\sum_{(k,l)} W_{(k,l),(i,j)}(\boldsymbol{z})}\,. \tag{18}$$

The weights $W_{(k,l),(i,j)}(\boldsymbol{z})$ depend on the similarity between the patches in $\boldsymbol{z}$ centered around (k, l) and (i, j), respectively, as well as on the distance between the two pixels. In a simple instantiation, we choose a patch size $L \in \mathbb{N}_{>0}$, a range $R > 0$ and a precision parameter $h > 0$. For a position (k,l) in the image, denote by $P_{(k,l)}(z)$ the subimage of z (or patch) centered in (k,l), of size $L \times L$. Then

$$W_{(k,l),(i,j)}(z) = \mathbf{1}_{\|(i,j)-(k,l)\| \le R} \exp\left(- \frac{\|P_{(k,l)}(z) - P_{(i,j)}(z)\|_2^2}{L^2 h^2} \right). \tag{19}$$

In other words, NLM averages patches that are similar to each other. The recent paper [**26**] studies this algorithm and demonstrates good performances.[4] Here we carry out similar simulations to demonstrate the accuracy of the state evolution prediction. At each iteration we can choose three parameters: L_t, R_t and h_t. We fix $L_t = 7$, $R_t = 11$ and adapt h_t to the noise level. The theory developed in the next sections suggests that $\|\boldsymbol{r}^t\|_2/\sqrt{m}$ is a good measure of the effective noise level after t iterations. We therefore set

$$h_t = 0.9 \cdot \frac{\|\boldsymbol{r}^t\|_2}{\sqrt{m}}, \tag{20}$$

where the coefficient 0.9 was selected empirically.

One difficulty is to compute the divergence of NLM denoisers $\operatorname{div} \eta_t$. Rather than computing explicitly the divergence from Equations (18) and (19), we use a trick suggested in [**36**]. The trick is based on the formula

$$\operatorname{div} \eta_t(x) = \lim_{\varepsilon \to 0} \mathbb{E}\left[\left\langle \boldsymbol{Z}, \frac{1}{\varepsilon}\left(\eta_t(\boldsymbol{x} + \varepsilon \boldsymbol{Z}) - \eta_t(\boldsymbol{x})\right) \right\rangle \right], \quad \boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_n). \tag{21}$$

Rather than taking the limit, we fix ε very small and evaluate the expectation by Monte Carlo. In high dimensions, concentration of measure helps and it is sufficient to use only one or a few samples to approximate the integral.

In Fig. 2, we demonstrate the algorithm performance for an image of size 170×170 (i.e. $n_1 = n_2 = 170$) with $m = 0.5 \cdot n_1 n_2$ measurements and noise level $\sigma_w = 0.034 \cdot \|\boldsymbol{x}_0\|_2/\sqrt{170}$. For each iteration $t \in \{0,1,2,3,4\}$, we show the estimates $\boldsymbol{x}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{r}^t$ (left column) together with the denoised versions $\boldsymbol{x}^{t+1} = \eta_t(\boldsymbol{x}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{r}^t)$ (right column). In Fig. 3, we report the evolution of the normalized square error $\mathrm{NMSE}(t;n) = \|\boldsymbol{x}^t - \boldsymbol{x}_0\|_2^2/\|\boldsymbol{x}_0\|_2^2$, as a function of the number of iterations. State evolution appears to track very closely the simulation results.

2. Further related work

AMP algorithms are motivated by ideas in spin glass theory, where they correspond to an iterative version of the celebrated Thouless-Anderson-Palmer (TAP) equations [**7**,**45**]. They can also be derived from graphical model ideas, by viewing them as approximations of belief propagation [**23**,**28**]. In both of these cases, the AMP nonlinearities turn out to be related to conditional expectation with respect to certain prior distributions. The theorems proved here apply more broadly, as demonstrated by the example in Section 1.2.

---

[4] While performances are obtained in the same paper using BM3D denoising [**26**], we consider NLM here because of its simplicity.

FIG. 2. Compressed sensing reconstruction of Lena using NLM-AMP, at undersampling ratio $m/n = 0.5$: iterates $\boldsymbol{x}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{r}^t$ (left column) and $\boldsymbol{x}^{t+1} = \eta_t(\boldsymbol{x}^t + \boldsymbol{A}^\mathsf{T}\boldsymbol{r}^t)$ (right column) for $t \in \{0, 1, 2, 3, 4\}$. (Details in the main text.)

The state evolution analysis of [5] was generalized in a number of directions over the past few years. State evolution was proven to hold for matrices $\boldsymbol{A}$ with i.i.d. sub-Gaussian entries in [4], under the assumption that the nonlinearity is a separable polynomial. The proof of [4] is based on the moment method, and hence is entirely different from the one presented here. Several generalizations of the basic iteration (1), (2) were studied in [21]. The framework of [21] allows to treat some classes of matrices with independent Gaussian, but not identically distributed entries, as well as algorithms in which $\boldsymbol{u}^t \in \mathbb{R}^{n\times k}$, $\boldsymbol{v}^t \in \mathbb{R}^{m\times k}$ are matrices with k fixed as $m, n \to \infty$.

A generalization of AMP to right-invariant random matrices was introduced and analyzed in [29,39], using the conditioning technique also applied here. This allows to treat classes of matrices with dependent entries and potentially large condition numbers. In the same direction, [9,32] develops iterative algorithms analogous to (1) and (2) for unitarily invariant symmetric matrices and for compressed sensing. The analysis in these works is based on non-rigorous density functional methods from statistical physics.

All results discussed above are asymptotic and characterize the limit $m, n \to \infty$ with m/n converging to a limit. Nevertheless, the conditioning technique does rely on central limit theorem and concentration of measure arguments and, as demonstrated in [40], it can be sharpened to obtain non-asymptotic results.

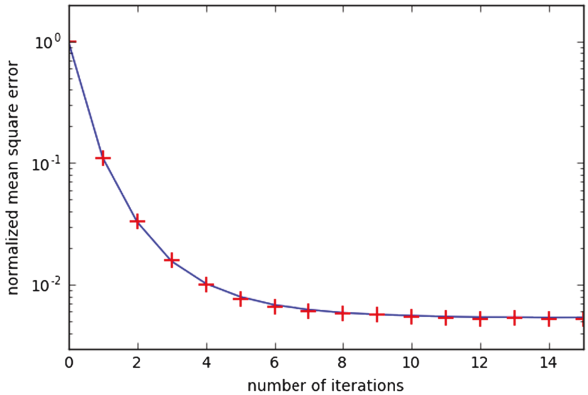

FIG. 3. Compressed sensing reconstruction of Lena using NLM-AMP. Red pluses (+): evolution of the normalized square error. Blue line: state evolution prediction.

Finally, a recent paper by Ma *et al.* [30] states a theorem establishing state evolution for compressed sensing reconstruction via AMP with a non-separable sliding-window denoiser. The result of [30] is not directly comparable with ours, since it concerns a special class of non-separable nonlinearities, but provides non-asymptotic guarantees.

3. Main results

In this section we state our main result for the asymmetric AMP iteration of Equations (1), (2). A similar result for symmetric AMP will be stated in Section 4 (and proven in 5).

3.1 *Definitions*

For two sequences (in n) of random variables X_n and Y_n, we write $X_n \overset{\mathrm{P}}{\simeq} Y_n$ when their difference converges in probability to 0, i.e. $X_n - Y_n \xrightarrow{\mathrm{P}} 0$.

For $\mathrm{K} = (\mathrm{K}_{s,r})_{1\le s,r\le t}$ a $t \times t$ covariance matrix, we will write $(\boldsymbol{Z}^1, \dots, \boldsymbol{Z}^t) \sim \mathsf{N}(0, \mathrm{K} \otimes \boldsymbol{I}_n)$ to mean that $\boldsymbol{Z}^1, \dots, \boldsymbol{Z}^t$ is a collection of centered, jointly Gaussian random vectors in $\mathbb{R}^n$, with covariances $\mathbb{E}[\boldsymbol{Z}^s(\boldsymbol{Z}^r)^\mathsf{T}] = \mathrm{K}_{s,r}\boldsymbol{I}_n$ for $1 \le s, r \le t$.

For $k \in \mathbb{N}_{>0}$ and any $n, m \in \mathbb{N}_{>0}$, a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is called *pseudo-Lipschitz of order k* if there exists a constant L such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\frac{\|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\|_2}{\sqrt{m}} \le L\left(1 + \left(\frac{\|\boldsymbol{x}\|_2}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\boldsymbol{y}\|_2}{\sqrt{n}}\right)^{k-1}\right)\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\sqrt{n}}. \tag{22}$$

L is then called the pseudo-Lipschitz constant of ϕ. Note that this definition is the same as introduced in [5], apart from a different scaling of the norm $\|\cdot\|_2$. The normalization factors are introduced to simplify the analysis that follows. For $k = 1$, this definition coincides with the standard definition of a Lipschitz function, for mapping between the normed spaces $(\mathbb{R}^n, \|\cdot\|_2/\sqrt{n})$ and $(\mathbb{R}^m, \|\cdot\|_2/\sqrt{m})$. In this case L is the Lipschitz constant of ϕ.

A sequence (in n) of pseudo-Lipschitz functions $\{\phi_n\}_{n\in\mathbb{N}_{>0}}$ is called *uniformly* pseudo-Lipschitz of order k if, denoting by L_n is the pseudo-Lipschitz constant of order k of ϕ_n, we have $L_n < \infty$ for each n and $\limsup_{n\to\infty} L_n < \infty$. Note that the input and output dimensions of each ϕ_n can depend on n. We call any $L > \limsup_{n\to\infty} L_n$ a pseudo-Lipschitz constant of the sequence.

3.2 *State evolution*

Fix $\delta > 0$ and consider a sequence $m = m(n) \in \mathbb{N}$ such that $m/n \to \delta$ as $n \to \infty$. For all n, we are given two sequences of (deterministic) functions, $\{e_t : \mathbb{R}^n \to \mathbb{R}^n\}_{t\in\mathbb{N}}$ and $\{g_t : \mathbb{R}^m \to \mathbb{R}^m\}_{t\in\mathbb{N}}$, as well as a sequence of (deterministic) vectors, $\boldsymbol{u}^0 = \boldsymbol{u}^0(n) \in \mathbb{R}^n$, and a sequence of random rectangular matrices $\boldsymbol{A} = \boldsymbol{A}(n) \in \mathbb{R}^{m\times n}$.

We next list our assumptions (we refer to Section 5.1 for a summary of notations used in the paper):

- **(B1)** $\boldsymbol{A}$ has entries $(A_{ij})_{i\le m, j\le n} \sim_{iid} \mathsf{N}(0, 1/m)$.
- **(B2)** For each $t \in \mathbb{N}$, the functions $e_t : \mathbb{R}^n \to \mathbb{R}^n$, $g_t : \mathbb{R}^m \to \mathbb{R}^m$ are uniformly Lipschitz (where uniformly is understood with respect to n).
- **(B3)** $\|\boldsymbol{u}^0\|_2/\sqrt{n}$ converges to a finite constant as $n \to \infty$.
- **(B4)** The following limit exists and is finite:

$$\Sigma_{0,0} \equiv \lim_{n\to\infty} \frac{1}{m}\langle e_0(\boldsymbol{u}^0), e_0(\boldsymbol{u}^0)\rangle. \tag{23}$$

- **(B5)** For any $t \in \mathbb{N}_{>0}$ and any $s \ge 0$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{m}\mathbb{E}\big[\langle e_0(\boldsymbol{u}^0), e_t(\boldsymbol{Z})\rangle\big], \tag{24}$$

  where $\boldsymbol{Z} \sim \mathsf{N}(0, s\boldsymbol{I}_n)$.
- **(B6)** For any $s, t \in \mathbb{N}_{>0}$ and any $\boldsymbol{S} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{S} \succeq \boldsymbol{0}$, the following limits exist and are finite:

$$\lim_{n\to\infty} \frac{1}{m}\mathbb{E}\left[\langle e_s(\boldsymbol{Z}_1), e_t(\boldsymbol{Z}_2)\rangle\right], \tag{25}$$

$$\lim_{n\to\infty} \frac{1}{m}\mathbb{E}\left[\langle g_s(\boldsymbol{Z}_3), g_t(\boldsymbol{Z}_4)\rangle\right], \tag{26}$$

  where $(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \sim \mathsf{N}(0, \boldsymbol{S}\otimes\boldsymbol{I}_n)$ and $(\boldsymbol{Z}_3, \boldsymbol{Z}_4) \sim \mathsf{N}(0, \boldsymbol{S}\otimes\boldsymbol{I}_m)$.

The technical Assumptions (B4), (B5) and (B6) allow to define two doubly infinite arrays, $(\Sigma_{s,r})_{s,r\ge 0}$ and $(\mathrm{T}_{s,r})_{s,r\ge 1}$, through the following recursion, known as *state evolution*.

We set $\Sigma_{0,0}$ using Assumption (B4). Then for each $t \ge 0$, given $(T_{s,r})_{1\le s,r\le t}$ and $(\Sigma_{s,r})_{0\le s,r\le t}$, we compute $(T_{s,r})_{1\le s,r\le t+1}$ and $(\Sigma_{s,r})_{0\le s,r\le t+1}$ using the following iteration:

- Set $(\boldsymbol{Z}_\sigma^0, \dots, \boldsymbol{Z}_\sigma^t) \sim \mathsf{N}(0, (\Sigma_{s,r})_{0\le s,r\le t} \otimes \boldsymbol{I}_m)$ and define

$$T_{t+1,s+1} = T_{s+1,t+1} = \lim_{n\to\infty} \frac{1}{m}\mathbb{E}\left[\langle g_s\left(\boldsymbol{Z}_\sigma^s\right), g_t\left(\boldsymbol{Z}_\sigma^t\right)\rangle\right], \qquad 0 \le s \le t. \tag{27}$$

  This defines the array $(T_{s,r})_{1\le s,r\le t+1}$.

- Set $\boldsymbol{Z}_\tau^0 = \boldsymbol{u}^0$ and $(\boldsymbol{Z}_\tau^1, \ldots, \boldsymbol{Z}_\tau^{t+1}) \sim \mathsf{N}(0, (T_{s,r})_{1 \le s,r \le t+1} \otimes \boldsymbol{I}_n)$ and define

$$\Sigma_{t+1,s} = \Sigma_{s,t+1} = \lim_{n\to\infty} \frac{1}{m} \mathbb{E}\left[\left\langle e_s\left(\boldsymbol{Z}_\tau^s\right), e_{t+1}\left(\boldsymbol{Z}_\tau^{t+1}\right)\right\rangle\right], \qquad 0 \le s \le t+1. \tag{28}$$

  This defines the array $(\Sigma_{s,r})_{0 \le s,r \le t+1}$.

We will refer to the arrays $(\Sigma_{s,r})_{s,r\ge 0}$ and $(\mathrm{T}_{s,r})_{s,r\ge 1}$ as to the *state evolution iterates* (and sometimes simply *state evolution*) and denote them by $\{\mathrm{T}_{s,t}, \Sigma_{s,t} | \, e_t, g_t, \boldsymbol{u}^0\}$ to make explicit the nonlinearities and initialization.

The state evolution characterizes the AMP iteration of Equations (1) and (2), which we copy here for the reader's convenience:

$$\boldsymbol{u}^{t+1} = \boldsymbol{A}^{\mathsf{T}} g_t(\boldsymbol{v}^t) - \mathsf{d}_t e_t(\boldsymbol{u}^t), \tag{29}$$

$$\boldsymbol{v}^t = \boldsymbol{A} e_t(\boldsymbol{u}^t) - \mathsf{b}_t g_{t-1}(\boldsymbol{v}^{t-1}), \tag{30}$$

where the initial condition is given by $\boldsymbol{u}^0$, and we let $g_{-1}(\,\cdot\,) = 0$ by convention. Further, we use the following expression for the memory terms (which we shall refer to as 'Onsager terms', following the physics tradition):

$$\mathsf{d}_t = \frac{1}{m}\mathbb{E}\left[\operatorname{div} g_t\left(\boldsymbol{Z}_\sigma^t\right)\right], \quad \mathsf{b}_t = \frac{1}{m}\mathbb{E}\left[\operatorname{div} e_t\left(\boldsymbol{Z}_\tau^t\right)\right], \tag{31}$$

where $\boldsymbol{Z}_\sigma^t \sim \mathsf{N}(0, \Sigma_{t,t}\boldsymbol{I}_m)$ and $\boldsymbol{Z}_\tau^t \sim \mathsf{N}(0, \mathrm{T}_{t,t}\boldsymbol{I}_n)$. We denote the asymmetric AMP iterates $(\boldsymbol{u}^t, \boldsymbol{v}^t)_{t\ge 0}$ by $\{\boldsymbol{u}^t, \boldsymbol{v}^t | \, e_t, g_t, \boldsymbol{u}^0\}$.

We are now in position to state our main result.

Theorem 1 Under Assumptions (B1)–(B6), consider the asymmetric AMP iteration $\{\boldsymbol{u}^t, \boldsymbol{v}^t | \, e_t, g_t, \boldsymbol{u}^0\}$ along with its state evolution $\{\mathrm{T}_{s,t}, \Sigma_{s,t} | \, e_t, g_t, \boldsymbol{u}^0\}$. Define for all n,

$$\left(\boldsymbol{Z}_\sigma^0, \ldots, \boldsymbol{Z}_\sigma^{t-1}\right) \sim \mathsf{N}\left(0, (\Sigma_{s,r})_{0\le s,r\le t-1} \otimes \boldsymbol{I}_m\right), \tag{32}$$

$$\left(\boldsymbol{Z}_\tau^1, \ldots, \boldsymbol{Z}_\tau^t\right) \sim \mathsf{N}\left(0, (\mathrm{T}_{s,r})_{1\le s,r\le t} \otimes \boldsymbol{I}_n\right), \tag{33}$$

such that the two collections, $(\boldsymbol{Z}_\sigma^0, \ldots, \boldsymbol{Z}_\sigma^{t-1})$ and $(\boldsymbol{Z}_\tau^1, \ldots, \boldsymbol{Z}_\tau^t)$, are independent of each other. Assume further that $\Sigma_{0,0}, \ldots, \Sigma_{t-1,t-1}, \mathrm{T}_{1,1}, \ldots, \mathrm{T}_{t,t} > 0$.

Then for any deterministic sequence $\phi_n : (\mathbb{R}^n \times \mathbb{R}^m)^t \times \mathbb{R}^n \to \mathbb{R}$ of uniformly pseudo-Lipschitz functions of order k,

$$\phi_n\left(\boldsymbol{u}^0, \boldsymbol{v}^0, \boldsymbol{u}^1, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{t-1}, \boldsymbol{u}^t\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{u}^0, \boldsymbol{Z}_\sigma^0, \boldsymbol{Z}_\tau^1, \boldsymbol{Z}_\sigma^1, \ldots, \boldsymbol{Z}_\sigma^{t-1}, \boldsymbol{Z}_\tau^t\right)\right]. \tag{34}$$

The proof of this theorem is presented in Section 6 and is obtained by reduction to the symmetric case, which is treated in the next section.

Let us emphasize a difference with respect to earlier work. In [5,21], the state evolution iterates are not two-dimensional arrays, but sequences describing the asymptotic variances of the iterates. Here the arrays of $(\Sigma_{s,t})_{s,t\ge 0}$, $(T_{s,t})_{s,t\ge 0}$ describe the asymptotic covariances across AMP iterates, as asserted by

Theorem 1. Applications in [6,17] illustrate the usefulness of controlling covariances as well. Notice the diagonals $\Sigma_{t,t}$, $t \geq 0$ and $T_{t,t}$, $t \geq 1$ correspond to the more standard state evolution iterates of [5,21]. The description given above implies that these diagonal elements can be computed through a simple one-dimensional recursion (which generalizes in the obvious way the recursion of [5]), without computing the off-diagonal elements. Namely,

$$T_{t+1,t+1} = \lim_{n\to\infty} \frac{1}{m}\mathbb{E}\left[\left\langle g_t\left(\boldsymbol{Z}_\sigma^t\right), g_t\left(\boldsymbol{Z}_\sigma^t\right)\right\rangle\right], \qquad \boldsymbol{Z}_\sigma^t \sim \mathsf{N}\left(0, \Sigma_{t,t}\cdot \boldsymbol{I}_m\right), \tag{35}$$

$$\Sigma_{t+1,t+1} = \lim_{n\to\infty} \frac{1}{m}\mathbb{E}\left[\left\langle e_{t+1}\left(\boldsymbol{Z}_\tau^{t+1}\right), e_{t+1}\left(\boldsymbol{Z}_\tau^{t+1}\right)\right\rangle\right], \qquad \boldsymbol{Z}_\tau^{t+1} \sim \mathsf{N}\left(0, T_{t+1,t+1}\cdot \boldsymbol{I}_n\right). \tag{36}$$

As mentioned above, we use Equation (31) to define the coefficients b_t, d_t because this simplifies the proofs. In practice, this definition is replaced by an empirical estimate, e.g. as in Equation (3). State evolution follows for these versions of AMP provided such estimates of b_t, d_t are consistent.

Corollary 2 Consider the modified AMP iteration whereby Equations (29) and (30) are replaced by

$$\hat{\boldsymbol{u}}^{t+1} = \boldsymbol{A}^\mathsf{T} g_t(\hat{\boldsymbol{v}}^t) - \hat{\mathsf{d}}_t e_t(\hat{\boldsymbol{u}}^t), \tag{37}$$

$$\hat{\boldsymbol{v}}^t = \boldsymbol{A} e_t(\hat{\boldsymbol{u}}^t) - \hat{\mathsf{b}}_t g_{t-1}(\hat{\boldsymbol{v}}^{t-1}), \tag{38}$$

with the initialization $\hat{\boldsymbol{u}}^0 = \boldsymbol{u}^0$, where $\hat{\mathsf{b}}_t = \hat{\mathsf{b}}_t(\hat{\boldsymbol{u}}^0, \hat{\boldsymbol{v}}^0, \ldots, \hat{\boldsymbol{v}}^{t-1}, \hat{\boldsymbol{u}}^t)$ and $\hat{\mathsf{d}}_t = \hat{\mathsf{d}}_t(\hat{\boldsymbol{u}}^0, \hat{\boldsymbol{v}}^0, \ldots, \hat{\boldsymbol{v}}^{t-1}, \hat{\boldsymbol{u}}^t, \hat{\boldsymbol{v}}^t)$ are two estimators of b_t, d_t. Assume the same conditions as Theorem 1. If for each t, $\hat{\mathsf{b}}_t(\,\cdot\,)$, $\hat{\mathsf{d}}_t(\,\cdot\,)$ are such that

$$\hat{\mathsf{b}}_t(\hat{\boldsymbol{u}}^0, \hat{\boldsymbol{v}}^0, \ldots, \hat{\boldsymbol{v}}^{t-1}, \hat{\boldsymbol{u}}^t) \overset{\mathrm{P}}{\simeq} \mathsf{b}_t, \quad \hat{\mathsf{d}}_t(\hat{\boldsymbol{u}}^0, \hat{\boldsymbol{v}}^0, \ldots, \hat{\boldsymbol{v}}^{t-1}, \hat{\boldsymbol{u}}^t, \hat{\boldsymbol{v}}^t) \overset{\mathrm{P}}{\simeq} \mathsf{d}_t, \tag{39}$$

then the iterates $(\hat{\boldsymbol{u}}^t, \hat{\boldsymbol{v}}^t)_{t\geq 0}$ satisfy state evolution, namely Equation (34) holds with $(\boldsymbol{u}^t, \boldsymbol{v}^t)_{t\geq 0}$ replaced by $(\hat{\boldsymbol{u}}^t, \hat{\boldsymbol{v}}^t)_{t\geq 0}$.

The proof of this statement is deferred to Section 6.

Two choices of $\hat{\mathsf{b}}_t$, $\hat{\mathsf{d}}_t$ that satisfy the assumptions are as follows:

- The empirical values

$$\hat{\mathsf{b}}_t = \frac{1}{m}\operatorname{div} e_t(\hat{\boldsymbol{u}}^t), \quad \hat{\mathsf{d}}_t = \frac{1}{m}\operatorname{div} g_t(\hat{\boldsymbol{v}}^t). \tag{40}$$

  By Theorem 1, if $\operatorname{div} e_t(\,\cdot\,)/m$, $\operatorname{div} g_t(\,\cdot\,)/m$ are uniformly pseudo-Lipschitz, then the assumptions of Corollary 2 hold, and hence we can apply state evolution.

- As an alternative,

$$\hat{\mathsf{b}}_t = \frac{n\langle \hat{\boldsymbol{u}}^t, e_t(\hat{\boldsymbol{u}}^t)\rangle}{m\|\hat{\boldsymbol{u}}^t\|_2^2}, \quad \hat{\mathsf{d}}_t = \frac{\langle \hat{\boldsymbol{v}}^t, g_t(\hat{\boldsymbol{v}}^t)\rangle}{\|\hat{\boldsymbol{v}}^t\|_2^2}. \tag{41}$$

  Consistency follows (for $e_t(\,\cdot\,)$, $g_t(\,\cdot\,)$ uniformly Lipschitz) from Theorem 1 and Gaussian integration by parts (in particular, Stein's lemma; see Lemma C.2).

4. Symmetric AMP

For all n, we are given a (deterministic) vector $\boldsymbol{x}^0 \in \mathbb{R}^n$ and a sequence of (deterministic) functions $\{f_t : \mathbb{R}^n \to \mathbb{R}^n\}_{t\in\mathbb{N}}$. These will be referred to as the *setting* $\{\boldsymbol{x}^0, f_t\}$. Given a sequence of (random) symmetric matrices $\boldsymbol{A} = \boldsymbol{A}(n) \in \mathbb{R}^{n\times n}$, we consider the following symmetric AMP iteration

$$\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{m}^t - \mathsf{b}_t \boldsymbol{m}^{t-1}, \tag{42}$$

$$\boldsymbol{m}^t = f_t(\boldsymbol{x}^t), \tag{43}$$

for $t \in \mathbb{N}$, with initialization $\boldsymbol{x}^0$ (and $\boldsymbol{m}^{-1} = 0$). Here

$$\mathsf{b}_t = \frac{1}{n}\mathbb{E}\left[\operatorname{div} f_t(\boldsymbol{Z}^t)\right], \tag{44}$$

where $\boldsymbol{Z}^t \sim \mathsf{N}(0, \mathrm{K}_{t,t}\boldsymbol{I}_n)$ and $\mathrm{K}_{t,t}$ will be defined via the state evolution recursion below (see, in particular, Equation (48)). We denote this AMP recursion as $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$, to make explicit the dependence on the setting.

We insist on the fact that $\boldsymbol{A}, f_t$ and $\boldsymbol{x}^0$ depend on n. However, we will drop this dependence most of the time to ease the reading.

We make the following assumptions.

(**A1**) $\boldsymbol{A}$ is sampled from the Gaussian orthogonal ensemble GOE(n), i.e. $\boldsymbol{A} = \boldsymbol{G} + \boldsymbol{G}^{\mathsf{T}}$ for $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ with i.i.d. entries $G_{ij} \sim \mathsf{N}(0, 1/(2n))$.

(**A2**) For each $t \in \mathbb{N}, f_t : \mathbb{R}^n \to \mathbb{R}^n$ is uniformly Lipschitz (as a sequence in n).

(**A3**) $\|\boldsymbol{x}^0\|_2/\sqrt{n}$ converges to a finite constant as $n \to \infty$.

(**A4**) The following limit exists and is finite:

$$\mathrm{K}_{1,1} \equiv \lim_{n\to\infty} \frac{1}{n}\langle f_0(\boldsymbol{x}^0), f_0(\boldsymbol{x}^0)\rangle. \tag{45}$$

(**A5**) For any $t \in \mathbb{N}_{>0}$ and any $s \geq 0$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\langle f_0(\boldsymbol{x}^0), f_t(\boldsymbol{Z})\rangle], \tag{46}$$

where $\boldsymbol{Z} \in \mathbb{R}^n$, $\boldsymbol{Z} \sim \mathsf{N}(0, s\boldsymbol{I}_n)$.

(**A6**) For any $s, t \in \mathbb{N}_{>0}$ and any $\boldsymbol{S} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{S} \succeq \boldsymbol{0}$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\langle f_s(\boldsymbol{Z}), f_t(\boldsymbol{Z}')\rangle\right], \tag{47}$$

where $(\boldsymbol{Z}, \boldsymbol{Z}') \in (\mathbb{R}^n)^2$, $(\boldsymbol{Z}, \boldsymbol{Z}') \sim \mathsf{N}(0, \boldsymbol{S} \otimes \boldsymbol{I}_n)$.

Given assumptions (A4), (A5) and (A6) we can define a doubly infinite array $(\mathrm{K}_{s,r})_{s,r\geq 1}$ via a *state evolution* recursion as follows.

The initial condition $\mathrm{K}_{1,1}$ is given by assumption (A4). Once $\mathbf{K}^{(t)} = (\mathrm{K}_{s,r})_{s,r\le t}$ is defined, let $(\boldsymbol{Z}^1,\dots,\boldsymbol{Z}^t) \sim \mathsf{N}(0,\mathrm{K}^{(t)} \otimes \boldsymbol{I}_n)$ and define, for $0 \le s \le t$,

$$\mathrm{K}_{t+1,s+1} = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\langle f_s(\boldsymbol{Z}^s), f_t(\boldsymbol{Z}^t)\rangle\right], \tag{48}$$

where it is understood that $\boldsymbol{Z}^0 = \boldsymbol{x}^0$ and $\mathrm{K}_{s+1,t+1} = \mathrm{K}_{t+1,s+1}$ is fixed by symmetry. We will refer to $(\mathrm{K}_{s,t})_{s,t\ge 1}$ as to the state evolution iterates, and we will emphasize their dependence on the setting denoting them by $\{\mathrm{K}_{s,t}|f_t,\boldsymbol{x}^0\}$. The Onsager term in Equation (42) is defined as per Equation (44), with $\boldsymbol{Z}^t \sim \mathsf{N}(0,\mathrm{K}_{t,t}\boldsymbol{I}_n)$ and $\mathrm{K}_{t,t}$ given by state evolution.

We can now state the following state evolution characterization of symmetric AMP, which is analogous to Theorem 1.

THEOREM 3 Under assumptions (A1)–(A6), consider the AMP iteration $\{\boldsymbol{x}^t,\boldsymbol{m}^t|f_t,\boldsymbol{x}^0\}$. Define for all n,

$$\left(\boldsymbol{Z}^1,\dots,\boldsymbol{Z}^{t+1}\right) \sim \mathsf{N}\left(0, \left(\mathrm{K}_{s,r}\right)_{s,r\le t+1} \otimes \boldsymbol{I}_n\right). \tag{49}$$

Assume further that $\mathrm{K}_{1,1},\dots,\mathrm{K}_{t,t} > 0$. For any sequence of uniformly pseudo-Lipschitz functions $\{\phi_n : (\mathbb{R}^n)^{t+2} \to \mathbb{R}\}$,

$$\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\dots,\boldsymbol{x}^{t+1}\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0,\boldsymbol{Z}^1,\dots,\boldsymbol{Z}^{t+1}\right)\right]. \tag{50}$$

The proof of this theorem is presented in Section 5. We also note that an analogue of Corollary 2 applies to this case as well and b_t can be replaced by a consistent estimator $\hat{\mathsf{b}}_t$.

5. Proof of Theorem 3 (symmetric AMP)

In this section we prove Theorem 3 using a sequence of lemmas, whose proofs are postponed to Section 5.5. We will also try to motivate the main steps. Throughout this section and the next, assumptions (A1)–(A6) hold.

5.1 *Notations*

We generally denote scalars by lower case letters, e.g. a, b, c, vectors by lower case boldface, e.g. $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and matrices by upper case boldface, e.g. $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$. We also use the upper case to emphasize that we are referring to a random variable, and—with a slight abuse of the convention—upper case boldface for random vectors.

For two random variables X and Y and a σ-algebra $\mathfrak{S}$, we use $X|_{\mathfrak{S}} \overset{\mathrm{d}}{=} Y$ to mean that for any integrable function ϕ and any $\mathfrak{S}$-measurable bounded random variable Z, $\mathbb{E}[\phi(X)Z] = \mathbb{E}[\phi(Y)Z]$. In other words, X is distributed as Y *conditional on* $\mathfrak{S}$. If $\mathfrak{S}$ is the trivial σ-algebra, we simply write $X \overset{\mathrm{d}}{=} Y$, i.e. X is distributed as Y.

For two vectors $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$, we denote their inner product by $\langle \boldsymbol{x},\boldsymbol{y}\rangle = \sum_{i=1}^n x_i y_i$ and the associated norm by $\|\boldsymbol{x}\|_2$. For two matrices, $\boldsymbol{X}$, $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$, $\langle \boldsymbol{X},\boldsymbol{Y}\rangle = \mathrm{Tr}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y})$ is their scalar product when viewed as vectors.

We use $\boldsymbol{I}_n$ to denote the $n \times n$ identity matrix. We use $\sigma_{\min}(\boldsymbol{Q})$ and $\sigma_{\max}(\boldsymbol{Q}) = \|\boldsymbol{Q}\|_{\mathrm{op}}$ to denote the minimum and maximum singular values of the matrix $\boldsymbol{Q}$. For two matrices, $\boldsymbol{Q}$ and $\boldsymbol{P}$, of the same number of rows, $[\boldsymbol{Q}|\boldsymbol{P}]$ denotes the matrix by concatenating $\boldsymbol{Q}$ and $\boldsymbol{P}$ horizontally. For any matrix $\boldsymbol{M}$, we denote the orthogonal projection onto its range $\boldsymbol{P}_M$ and we let $\boldsymbol{P}_M^{\perp} = \boldsymbol{I} - \boldsymbol{P}_M$. When $\boldsymbol{M}$ is an empty matrix, $\boldsymbol{P}_M = 0$ and $\boldsymbol{P}_M^{\perp} = \boldsymbol{I}$. When $\boldsymbol{M}$ has full column rank, $\boldsymbol{P}_M = \boldsymbol{M}(\boldsymbol{M}^{\mathsf{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathsf{T}}$.

If $f : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz function, it is almost everywhere differentiable (w.r.t. the Lebesgue measure), and thus we can define almost everywhere the quantity

$$\operatorname{div} f(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i}(\boldsymbol{x}), \tag{51}$$

where $f_i(\boldsymbol{x})$ is the ith coordinate of $f(\boldsymbol{x})$.

We say that a sequence of events, which depends on n, holds with high probability (w.h.p.) if it holds with probability converging to 1 as $n \to \infty$.

We define the Wasserstein distance (of order 2) between two probability measures μ and ν as

$$W_2(\mu, \nu) = \inf_{(X,Y)} \mathbb{E}\big[\,(X - Y)^2\,\big]^{1/2}, \tag{52}$$

where the infimum is taken over all couplings of μ and ν, i.e. all random variables (X, Y) such that $X \sim \mu$ and $Y \sim \nu$ marginally.

5.2 *Long AMP and proof outline*

The main idea of the proof is to analyze a different recursion than the AMP recursion (42) and (43). This new recursion satisfies the conclusion of Theorem 3 and will be a good approximation of the AMP recursion in the asymptotic $n \to \infty$. It is defined as follows:

$$\boldsymbol{h}^{t+1} = \boldsymbol{P}^{\perp}_{Q_{t-1}} \boldsymbol{A} \boldsymbol{P}^{\perp}_{Q_{t-1}} \boldsymbol{q}^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t, \tag{53}$$

$$\boldsymbol{q}^t = f_t(\boldsymbol{h}^t), \tag{54}$$

where at each step t, we have defined

$$\boldsymbol{Q}_{t-1} = \left[\boldsymbol{q}^0|\boldsymbol{q}^1|\cdots|\boldsymbol{q}^{t-1}\right], \tag{55}$$

$$\boldsymbol{\alpha}^t = \left(\boldsymbol{Q}_{t-1}^{\mathsf{T}}\boldsymbol{Q}_{t-1}\right)^{-1}\boldsymbol{Q}_{t-1}^{\mathsf{T}}\boldsymbol{q}^t, \tag{56}$$

$$\boldsymbol{H}_{t-1} = \left[\boldsymbol{h}^1|\boldsymbol{h}^2|\cdots|\boldsymbol{h}^t\right]. \tag{57}$$

The initialization is $\boldsymbol{q}^0 = f_0(\boldsymbol{x}^0)$ and $\boldsymbol{h}^1 = \boldsymbol{A}\boldsymbol{q}^0$. This recursion will be referred as the *Long AMP* recursion, or LoAMP $\{\boldsymbol{h}^t, \boldsymbol{q}^t | f_t, \boldsymbol{x}^0\}$.

Note that for the LoAMP recursion to be well defined, the matrices $\boldsymbol{Q}_{t-1}^{\mathsf{T}}\boldsymbol{Q}_{t-1}$ must be invertible, that is to say the family $\boldsymbol{q}^0, \boldsymbol{q}^1, \ldots, \boldsymbol{q}^{t-1}$ must be linearly independent. This has no reason to be true, since $\boldsymbol{q}^s = f_s(\boldsymbol{h}^s)$ and f_s is a generic sequence of Lipschitz functions (satisfying assumptions (A4)–(A6)). For instance, if all f_s, $s = 0, \ldots, t-1$, have images included in a same subspace of dimension lower than t, this cannot be true. This difficulty leads to some technicalities in the proof. However, we will start by studying the case where $\boldsymbol{Q}_{t-1}^{\mathsf{T}}\boldsymbol{Q}_{t-1}$ is invertible, with $\sigma_{\min}(\boldsymbol{Q}_{t-1})/\sqrt{n} \geq c_t > 0$, for n large enough, where c_t is a constant independent of n. More formally, we make the following assumption.

Assumption (non-degeneracy): We say that the LoAMP iterates satisfy the non-degeneracy assumption if

- almost surely, for all $t \in \mathbb{N}$ and all $n \geq t$, $\boldsymbol{Q}_{t-1}$ has full column rank;
- for all $t \in \mathbb{N}_{>0}$, there exists some constant $c_t > 0$ -independent of n- such that almost surely, there exists n_0 (random) such that, for $n \geq n_0$, $\sigma_{\min}(\boldsymbol{Q}_{t-1})/\sqrt{n} \geq c_t > 0$.

At this point we introduced all the basic concepts needed for the proof, and it is useful to pause in order to describe the proof strategy:

Step 1. We prove that state evolution (namely, a version of Theorem 3) holds for LoAMP under non-degeneracy. This proof is outlined in Section 5.3, with technical lemmas proved in Section 5.5.
The proof proceeds by induction over the number of steps t. The induction hypothesis at step t effectively captures the joint distribution of vectors $\boldsymbol{x}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{h}^t$. In order to obtain the conditional distribution of $\boldsymbol{h}^{t+1}$ given the previous iterates, we use an exact characterization of the conditional distribution of $\boldsymbol{A}$ given the past $\boldsymbol{x}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{h}^t$. This is possible thanks to the fact that $\boldsymbol{A}$ is Gaussian.

Step 2. We prove that LoAMP iterates are well approximated by AMP iterates, and hence state evolution also holds for AMP under the non-degeneracy condition, cf. Lemma 6 and Theorem 7.

Step 3. We then consider general nonlinearities $f_t(\,\cdot\,)$ (no non-degeneracy assumption) and define the perturbation $f_t^{\epsilon y}(\,\cdot\,) = f_t(\,\cdot\,) + \epsilon \boldsymbol{y}^t$, where the vectors $(\boldsymbol{y}^t)_{t\geq 0}$ are i.i.d. standard normal. We prove that, for any fixed $\varepsilon > 0$, the non-degeneracy assumption holds for this perturbed iteration with high probability, and hence an analogous of Theorem 3 holds for this case (cf. Lemma 9).

Step 4. We prove that AMP at $\varepsilon > 0$ is approximated by AMP at $\varepsilon = 0$ (i.e. the unperturbed case) up to an error that vanishes as $\varepsilon \to 0$, uniformly in n, cf. Lemma 12. Analogously, state evolution at $\varepsilon > 0$ is well approximated by state evolution at $\varepsilon = 0$, up to an error that vanishes as $\varepsilon \to 0$, cf. Lemma 11. Using these approximation results, we establish state evolution for the general case.

5.3 *The non-degenerate case*

The LoAMP recursion is of interest because it behaves well with Gaussian conditioning, so that the sequence of iterates becomes easier to study. The following lemma makes this idea explicit.

LEMMA 4 Consider the LoAMP $\{\boldsymbol{h}^t, \boldsymbol{q}^t | f_t, \boldsymbol{x}^0\}$ and assume it satisfies the non-degeneracy assumption. Fix $t \in \mathbb{N}_{>0}$. Let $\mathfrak{S}_t$ be the σ-algebra generated by $\boldsymbol{h}^1, \ldots, \boldsymbol{h}^t$ and denote $\boldsymbol{q}_\perp^t = \boldsymbol{P}_{Q_{t-1}}^\perp \boldsymbol{q}^t$ and $\boldsymbol{q}_\parallel^t = \boldsymbol{P}_{Q_{t-1}} \boldsymbol{q}^t$. Then

$$\boldsymbol{h}^{t+1}|_{\mathfrak{S}_t} \overset{\mathrm{d}}{=} \boldsymbol{P}_{Q_{t-1}}^\perp \tilde{\boldsymbol{A}} \boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1} \boldsymbol{\alpha}^t, \tag{58}$$

where $\tilde{\boldsymbol{A}}$ is an independent copy of $\boldsymbol{A}$.

Here, we decompose $\boldsymbol{h}^{t+1}$ as a sum of past iterates $\boldsymbol{h}^1, \ldots, \boldsymbol{h}^t$ and of a new Gaussian vector $\boldsymbol{P}_{Q_{t-1}}^\perp \tilde{\boldsymbol{A}} \boldsymbol{q}_\perp^t$, whose conditional law knowing the past $\mathfrak{S}_t$ is well understood. The key property is that we have replaced $\boldsymbol{A}$ by a new matrix $\tilde{\boldsymbol{A}}$ decoupled from the past iterates. This enables us to show that the sets of points $\boldsymbol{q}^0, \boldsymbol{q}^1, \ldots, \boldsymbol{q}^t$ and $\boldsymbol{h}^1, \boldsymbol{h}^2, \ldots, \boldsymbol{h}^{t+1}$ have asymptotically the same geometry and that the conclusion of Theorem 3 holds for LoAMP. The following lemma gives a precise statement.

LEMMA 5 Consider the LoAMP $\{\boldsymbol{h}^t, \boldsymbol{q}^t | f_t, \boldsymbol{x}^0\}$ and suppose it satisfies the non-degeneracy assumption. Then

(a) For all $0 \le s, r \le t$,

$$\frac{1}{n}\left\langle \boldsymbol{h}^{s+1}, \boldsymbol{h}^{r+1} \right\rangle \overset{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{q}^s, \boldsymbol{q}^r \right\rangle. \tag{59}$$

(b) For any $t \in \mathbb{N}$, for any sequence of uniformly order-k pseudo-Lipschitz functions $\{\phi_n : (\mathbb{R}^n)^{t+2} \to \mathbb{R}\}$,

$$\phi_n\left(\boldsymbol{x}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{h}^{t+1}\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0, \boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1}\right)\right], \tag{60}$$

where

$$(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1}) \sim \mathsf{N}\left(0, \left(\mathrm{K}_{s,r}\right)_{s,r \le t+1} \otimes \boldsymbol{I}_n\right). \tag{61}$$

Here the state evolution $\{\mathrm{K}_{s,t} | f_t, \boldsymbol{x}^0\}$ is described in Section 4.

To conclude that Theorem 3 holds in this case, we only need to show that LoAMP is a good approximation of AMP.

LEMMA 6 Consider the AMP $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$ and the LoAMP $\{\boldsymbol{h}^t, \boldsymbol{q}^t | f_t, \boldsymbol{x}^0\}$. Suppose the LoAMP satisfies the non-degeneracy assumption. For any $t \in \mathbb{N}$,

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{h}^{t+1} - \boldsymbol{x}^{t+1}\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0 \quad \text{and} \quad \frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^t - \boldsymbol{m}^t\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{62}$$

Wrapping things together, we have shown the following weaker form of Theorem 3.

THEOREM 7 Assume (A1)–(A6) and that the LoAMP iterates satisfy the non-degeneracy assumption. Consider the AMP $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$. For any sequence of uniformly order-k pseudo-Lipschitz functions $\{\phi_n : (\mathbb{R}^n)^{t+2} \to \mathbb{R}\}$,

$$\phi_n\left(\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^{t+1}\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0, \boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1}\right)\right], \tag{63}$$

where

$$\left(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1}\right) \sim \mathsf{N}\left(0, \left(\mathrm{K}_{s,r}\right)_{s,r \le t+1} \otimes \boldsymbol{I}_n\right). \tag{64}$$

5.4 *The general case*

To treat the case where the matrix $\boldsymbol{Q}_{t-1}$ is ill-conditioned, we add a small perturbation to the functions f_s so that the perturbed AMP behaves well. We then make sure that the perturbed AMP approximates well the original one.

A convenient way to implement this program is to perturb *randomly* the functions. We then show that, almost surely, the perturbation has the required properties (A4)–(A6). Specifically, consider

$$f_t^{\epsilon y}(\,\cdot\,) = f_t(\,\cdot\,) + \epsilon \boldsymbol{y}^t, \tag{65}$$

where $\epsilon \geq 0$ and $\boldsymbol{y}^0, \boldsymbol{y}^1, \boldsymbol{y}^2, \ldots$ are generated as i.i.d. $\mathsf{N}(0, \boldsymbol{I}_n)$, independent of the matrix A. The perturbation vectors $\boldsymbol{y}^0, \boldsymbol{y}^1, \boldsymbol{y}^2, \ldots$ are called collectively as $\boldsymbol{y}$ for brevity.

LEMMA 8 Almost surely (w.r.t. $\boldsymbol{y}$), the setting $\{\boldsymbol{x}^0, f_t^{\epsilon y}\}$ satisfies assumptions (A4)–(A6). As a consequence, we can define an associated state evolution $\{\mathrm{K}^\epsilon_{s,t} | f_t^{\epsilon y}, \boldsymbol{x}^0\}$:

$$\mathrm{K}^\epsilon_{1,1} = \lim_{n\to\infty} \frac{1}{n} \| f_0^{\epsilon y}(\boldsymbol{x}^0) \|_2^2, \tag{66}$$

and once $\mathrm{K}^\epsilon = (\mathrm{K}^\epsilon_{s,r})_{s,r\leq t}$ is defined, take $(\boldsymbol{Z}^{\epsilon,1}, \ldots, \boldsymbol{Z}^{\epsilon,t}) \sim \mathsf{N}(0, \mathrm{K}^\epsilon \otimes \boldsymbol{I}_n)$ independently of $\boldsymbol{y}$ and define

$$\mathrm{K}^\epsilon_{1,t+1} = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\big[\langle f_0^{\epsilon y}(\boldsymbol{x}^0), f_t^{\epsilon y}\left(\boldsymbol{Z}^{\epsilon,t}\right) \rangle\big], \tag{67}$$

$$\mathrm{K}^\epsilon_{s+1,t+1} = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\langle f_s^{\epsilon y}\left(\boldsymbol{Z}^{\epsilon,s}\right), f_t^{\epsilon y}\left(\boldsymbol{Z}^{\epsilon,t}\right)\rangle\right], \tag{68}$$

where the expectations are taken w.r.t. $\boldsymbol{Z}^{\epsilon,1}, \ldots, \boldsymbol{Z}^{\epsilon,t}$, but not $\boldsymbol{y}$. Moreover, the resulting state evolution is almost surely equal to a constant, thus justifying that we drop the dependence on $\boldsymbol{y}$ in $\mathrm{K}^\epsilon_{s,t}$.

LEMMA 9 Denote as $\boldsymbol{Q}^{\epsilon y}_{t-1}$ the matrix associated with the LoAMP iterates $\{\boldsymbol{h}^{\epsilon y,t}, \boldsymbol{q}^{\epsilon y,t} | f_t^{\epsilon y}, \boldsymbol{x}^0\}$, according to Equation (55). Assume $\epsilon > 0$. Then as soon as $n \geq t$, almost surely the matrix $\boldsymbol{Q}^{\epsilon y}_{t-1}$ is of full column rank. Furthermore, there exists a constant $c_{t,\epsilon} > 0$—independent of n—such that almost surely, there exists n_0 (random) such that for $n \geq n_0$, $\sigma_{\min}(\boldsymbol{Q}^{\epsilon y}_{t-1})/\sqrt{n} \geq c_{t,\epsilon}$.

The last two lemmas imply that almost surely, we can apply Theorem 7 to $\{f_t^{\epsilon y}\}_{t\geq 0}$. The next three lemmas quantify how this result approximates our original one.

LEMMA 10 Let $\{\phi_n : (\mathbb{R}^n)^t \to \mathbb{R}\}_{n>0}$ be a sequence of uniformly pseudo-Lipschitz functions of order k. Let $\mathrm{K}, \tilde{\mathrm{K}}$ be two $t \times t$ covariance matrices and $\boldsymbol{Z} \sim \mathsf{N}(0, \mathrm{K} \otimes \boldsymbol{I}_n)$, $\tilde{\boldsymbol{Z}} \sim \mathsf{N}(0, \tilde{\mathrm{K}} \otimes \boldsymbol{I}_n)$. Then

$$\lim_{\tilde{\mathrm{K}}\to \mathrm{K}} \sup_{n\geq 1} \left|\mathbb{E}[\phi_n(\boldsymbol{Z})] - \mathbb{E}[\phi_n(\tilde{\boldsymbol{Z}})]\right| = 0. \tag{69}$$

LEMMA 11 For any $s, t \geq 1$, $\mathrm{K}^\epsilon_{s,t} \xrightarrow[\epsilon\to 0]{} \mathrm{K}_{s,t}$.

LEMMA 12 Consider the AMP iterates in two different settings, $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$ and $\{\boldsymbol{x}^{\epsilon y,t}, \boldsymbol{m}^{\epsilon y,t} | f_t^{\epsilon y}, \boldsymbol{x}^0\}$. Assume further that for some $t \in \mathbb{N}$, $\mathrm{K}_{1,1}, \ldots, \mathrm{K}_{t,t} > 0$. Then there exist functions $h_t(\epsilon)$, $h_t'(\epsilon)$, independent of n, such that

$$\lim_{\epsilon\to 0} h_t(\epsilon) = \lim_{\epsilon\to 0} h_t'(\epsilon) = 0, \tag{70}$$

and for all $\epsilon \leq 1$, with high probability,

$$\frac{1}{\sqrt{n}} \left\|\boldsymbol{m}^{\epsilon y,t} - \boldsymbol{m}^t\right\|_2 \leq h_t'(\epsilon), \tag{71}$$

$$\frac{1}{\sqrt{n}} \left\|\boldsymbol{x}^{\epsilon y,t+1} - \boldsymbol{x}^{t+1}\right\|_2 \leq h_t(\epsilon). \tag{72}$$

Proof of Theorem 3 The proof combines three elements that follow from the previous lemmas:

- Thanks to Lemmas 8 and 9, *almost surely w.r.t. the perturbation* $\boldsymbol{y}^0, \boldsymbol{y}^1, \ldots$, the assumptions of Theorem 7 are satisfied for the perturbed setting $\{\boldsymbol{x}^0, f_t^{\epsilon y}\}$. We get that a.s., for any sequence of uniformly pseudo-Lipschitz functions $\{\phi_n : (\mathbb{R}^n)^{t+2} \to \mathbb{R}\}$,

$$\phi_n\big(\boldsymbol{x}^0, \boldsymbol{x}^{\epsilon y,1}, \ldots, \boldsymbol{x}^{\epsilon y,t+1}\big) \overset{\mathrm{P}}{\simeq} \mathbb{E}\big[\phi_n\big(\boldsymbol{x}^0, \boldsymbol{Z}^{\epsilon,1}, \ldots, \boldsymbol{Z}^{\epsilon,t+1}\big)\big], \tag{73}$$

where $\boldsymbol{Z}^{\epsilon,1}, \ldots, \boldsymbol{Z}^{\epsilon,t+1} \sim \mathsf{N}(0, (\mathrm{K}^{\epsilon}_{r,s})_{r,s\le t+1} \otimes \boldsymbol{I}_n)$. To obtain the desired result, we shall take the limit $\epsilon \to 0$, the technicalities of which are presented in the following two elements.

- Let $\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1} \sim \mathsf{N}(0, (\mathrm{K}_{r,s})_{r,s\le t+1} \otimes \boldsymbol{I}_n)$. Since, by Lemma 11, the perturbed state evolution converges to the original one when $\epsilon \to 0$, so we can apply Lemma 10 to get

$$\sup_{n\ge 1} \Big|\mathbb{E}\big[\phi_n\big(\boldsymbol{x}^0, \boldsymbol{Z}^{\epsilon,1}, \ldots, \boldsymbol{Z}^{\epsilon,t+1}\big)\big] - \mathbb{E}\big[\phi_n\big(\boldsymbol{x}^0, \boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{t+1}\big)\big]\Big| \xrightarrow[\epsilon\to 0]{} 0. \tag{74}$$

- Using that ϕ_n is uniformly pseudo-Lipschitz of order k and the triangle inequality,

$$\Big|\phi_n\big(\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^{t+1}\big) - \phi_n\big(\boldsymbol{x}^0, \boldsymbol{x}^{\epsilon y,1}, \ldots, \boldsymbol{x}^{\epsilon y,t+1}\big)\Big| \tag{75}$$

$$\le LC_1(k,t)\left(1 + \frac{\|\boldsymbol{x}^0\|_2^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^{t+1} \frac{\|\boldsymbol{x}^{\epsilon y,i}\|_2^{k-1}}{n^{(k-1)/2}} + \sum_{i=1}^{t+1} \frac{\|\boldsymbol{x}^{i}\|_2^{k-1}}{n^{(k-1)/2}}\right) \sum_{i=1}^{t+1} \frac{\|\boldsymbol{x}^{\epsilon y,i} - \boldsymbol{x}^i\|_2}{\sqrt{n}}, \tag{76}$$

where here $C_j(k,t)$ is a constant depending only on k and t. Lemma 12 ensures that w.h.p. $\|\boldsymbol{x}^{\epsilon y,i} - \boldsymbol{x}^i\|_2/\sqrt{n} \le h_i(\epsilon)$. We also know by assumption (A3) that $\|\boldsymbol{x}^0\|_2/\sqrt{n}$ converges to a finite limit. Furthermore, one can use Theorem 7 to bound w.h.p.

$$\frac{\|\boldsymbol{x}^{\epsilon y,i}\|_2^{k-1}}{n^{(k-1)/2}} = \frac{\mathbb{E}\left[\|\boldsymbol{Z}^{\epsilon,i}\|_2^{k-1}\right]}{n^{(k-1)/2}} + \left(\frac{\|\boldsymbol{x}^{\epsilon y,i}\|_2^{k-1}}{n^{(k-1)/2}} - \frac{\mathbb{E}\left[\|\boldsymbol{Z}^{\epsilon,i}\|_2^{k-1}\right]}{n^{(k-1)/2}}\right) \le C_2(k) \left\|\left(\mathrm{K}^{\epsilon}_{s,r}\right)_{s,r\le t+1}\right\|_{\mathrm{op}}^{(k-1)/2} + 1. \tag{77}$$

Finally, using the triangle inequality, w.h.p.,

$$\frac{\|\boldsymbol{x}^{i}\|_2^{k-1}}{n^{(k-1)/2}} \le C_3(k)\left(\frac{\|\boldsymbol{x}^{\epsilon y,i}\|_2^{k-1}}{n^{(k-1)/2}} + \frac{\|\boldsymbol{x}^{\epsilon y,i} - \boldsymbol{x}^i\|_2^{k-1}}{n^{(k-1)/2}}\right) \tag{78}$$

$$\le C_4(k)\left(\left\|\left(\mathrm{K}^{\epsilon}_{s,r}\right)_{s,r\le t+1}\right\|_{\mathrm{op}}^{(k-1)/2} + 1 + h_i(\epsilon)^{k-1}\right). \tag{79}$$

Putting things together, we get w.h.p.,

$$\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^{\epsilon y,1},\ldots,\boldsymbol{x}^{\epsilon y,t+1}\right)\right| \tag{80}$$

$$\leq LC_5(k,t)\left(1+\left\|\left(\mathrm{K}^{\epsilon}_{s,r}\right)_{s,r\leq t+1}\right\|_{\mathrm{op}}^{(k-1)/2}+\sum_{i=1}^{t+1}h_i(\epsilon)^{k-1}\right)\sum_{i=1}^{t+1}h_i(\epsilon). \tag{81}$$

As this upper bound converges to 0 as $\epsilon\to 0$, we have for any $\eta>0$,

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty}\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^{\epsilon y,1},\ldots,\boldsymbol{x}^{\epsilon y,t+1}\right)\right|\geq\eta\right)=0. \tag{82}$$

Let us now combine the three elements together. Let $\eta>0$. We have the following:

$$\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0,\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^{t+1}\right)\right]\right|\geq\eta\right) \tag{83}$$

$$\leq\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^{\epsilon y,1},\ldots,\boldsymbol{x}^{\epsilon y,t+1}\right)\right|\geq\frac{\eta}{3}\right) \tag{84}$$

$$+\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^{\epsilon y,1},\ldots,\boldsymbol{x}^{\epsilon y,t+1}\right)-\mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0,\boldsymbol{Z}^{\epsilon,1},\ldots,\boldsymbol{Z}^{\epsilon,t+1}\right)\right]\right|\geq\frac{\eta}{3}\right) \tag{85}$$

$$+\mathbb{I}_{\{|\mathbb{E}[\phi_n(\boldsymbol{x}^0,\boldsymbol{Z}^{\epsilon,1},\ldots,\boldsymbol{Z}^{\epsilon,t+1})]-\mathbb{E}[\phi_n(\boldsymbol{x}^0,\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^{t+1})]|\geq\eta/3\}}. \tag{86}$$

Taking $\limsup$ as $n\to\infty$, the second term vanishes because of (73):

$$\limsup_{n\to\infty}\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0,\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^{t+1}\right)\right]\right|\geq\eta\right) \tag{87}$$

$$\leq\limsup_{n\to\infty}\Pr\left(\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^{\epsilon y,1},\ldots,\boldsymbol{x}^{\epsilon y,t+1}\right)\right|\geq\frac{\eta}{3}\right) \tag{88}$$

$$+\mathbb{I}_{\{\sup_{n\geq 1}|\mathbb{E}[\phi_n(\boldsymbol{x}^0,\boldsymbol{Z}^{\epsilon,1},\ldots,\boldsymbol{Z}^{\epsilon,t+1})]-\mathbb{E}[\phi_n(\boldsymbol{x}^0,\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^{t+1})]|\geq\eta/3\}}. \tag{89}$$

Because of (82) and (74), this upper bound converges to 0 as $\epsilon\to 0$. We can then conclude that

$$\left|\phi_n\left(\boldsymbol{x}^0,\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{t+1}\right)-\mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0,\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^{t+1}\right)\right]\right|\xrightarrow[n\to\infty]{P}0. \tag{90}$$

□

5.5 *Proof of the Lemmas*

5.5.1 *Proof of Lemma* 4 The claim for $t = 0$ is immediate from the fact that $\mathfrak{S}_0$ is the trivial σ-algebra and $\boldsymbol{P}^{\perp}_{Q_{t-1}} = \boldsymbol{I}_n$. For $t \geq 1$, let us rewrite (53) as

$$\boldsymbol{h}^{t+1} = \boldsymbol{A}\boldsymbol{q}^t_\perp - \boldsymbol{P}_{Q_{t-1}}\boldsymbol{A}\boldsymbol{q}^t_\perp + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t \tag{91}$$

$$= \boldsymbol{A}\boldsymbol{q}^t_\perp - \tilde{\boldsymbol{Q}}_{t-1}\left(\tilde{\boldsymbol{Q}}^{\mathsf{T}}_{t-1}\tilde{\boldsymbol{Q}}_{t-1}\right)^{-1}\boldsymbol{Y}^{\mathsf{T}}_{t-1}\boldsymbol{q}^t_\perp + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t, \tag{92}$$

where $\boldsymbol{q}^t_\perp = \boldsymbol{P}^{\perp}_{Q_{t-1}}\boldsymbol{q}^t$, $\tilde{\boldsymbol{Q}}_{t-1} = [\boldsymbol{q}^0_\perp|\boldsymbol{q}^1_\perp|\dots|\boldsymbol{q}^{t-1}_\perp]$ and $\boldsymbol{Y}_{t-1} = [\boldsymbol{y}^0|\boldsymbol{y}^1|\dots|\boldsymbol{y}^{t-1}]$ with $\boldsymbol{y}^s = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{q}^s_\perp = \boldsymbol{A}\boldsymbol{q}^s_\perp$. Here we use the fact that $\boldsymbol{P}_{Q_{t-1}} = \boldsymbol{P}_{\tilde{Q}_{t-1}}$. Notice that

$$\boldsymbol{y}^0 = \boldsymbol{h}^1, \tag{93}$$

$$\boldsymbol{y}^s = \boldsymbol{h}^{s+1} + \tilde{\boldsymbol{Q}}_{s-1}\left(\tilde{\boldsymbol{Q}}^{\mathsf{T}}_{s-1}\tilde{\boldsymbol{Q}}_{s-1}\right)^{-1}\boldsymbol{Y}^{\mathsf{T}}_{s-1}\boldsymbol{q}^s_\perp - \boldsymbol{H}_{s-1}\boldsymbol{\alpha}^s \tag{94}$$

for any $s \geq 1$. Also, $\boldsymbol{H}_{s-1}$, $\boldsymbol{Q}_{s-1}$ and $\tilde{\boldsymbol{Q}}_{s-1}$ are $\mathfrak{S}_t$-measurable for $1 \leq s \leq t$. Then a simple induction yields that $\boldsymbol{Y}_{t-1}$ is $\mathfrak{S}_t$-measurable. Hence, to find $\boldsymbol{A}|_{\mathfrak{S}_t}$, conditioning on $\mathfrak{S}_t$ is equivalent to conditioning on the linear constraint $\boldsymbol{A}\tilde{\boldsymbol{Q}}_{t-1} = \boldsymbol{Y}_{t-1}$. As shown in [21, Lemma 3] and [5, Lemma 10], $\boldsymbol{A}|_{\mathfrak{S}_t} \stackrel{\mathrm{d}}{=} \mathbb{E}[\boldsymbol{A}|\mathfrak{S}_t] + \mathcal{P}_t(\tilde{\boldsymbol{A}})$, where $\tilde{\boldsymbol{A}} \stackrel{\mathrm{d}}{=} \boldsymbol{A}$ independent of $\mathfrak{S}_t$ and $\mathcal{P}_t$ is the orthogonal projector onto the subspace $\{\hat{\boldsymbol{A}} \in \mathbb{R}^{n\times n}|\hat{\boldsymbol{A}}\tilde{\boldsymbol{Q}}_{t-1} = 0, \hat{\boldsymbol{A}} = \hat{\boldsymbol{A}}^{\mathsf{T}}\}$:

$$\mathbb{E}\left[\boldsymbol{A}\middle|\mathfrak{S}_t\right] = \boldsymbol{A} - \boldsymbol{P}^{\perp}_{\tilde{Q}_{t-1}}\boldsymbol{A}\boldsymbol{P}^{\perp}_{\tilde{Q}_{t-1}} = \boldsymbol{A} - \boldsymbol{P}^{\perp}_{Q_{t-1}}\boldsymbol{A}\boldsymbol{P}^{\perp}_{Q_{t-1}}, \tag{95}$$

$$\mathcal{P}_t\left(\tilde{\boldsymbol{A}}\right) = \boldsymbol{P}^{\perp}_{\tilde{Q}_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{P}^{\perp}_{\tilde{Q}_{t-1}} = \boldsymbol{P}^{\perp}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{P}^{\perp}_{Q_{t-1}}, \tag{96}$$

where we use $\boldsymbol{P}^{\perp}_{\tilde{Q}_{t-1}} = \boldsymbol{P}^{\perp}_{Q_{t-1}}$. Then from (53),

$$\boldsymbol{h}^{t+1}|_{\mathfrak{S}_t} \stackrel{\mathrm{d}}{=} \boldsymbol{P}^{\perp}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{P}^{\perp}_{Q_{t-1}}\boldsymbol{q}^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t \tag{97}$$

since $\boldsymbol{P}^{\perp}_{Q_{t-1}}\mathbb{E}[\boldsymbol{A}|\mathfrak{S}_t]\boldsymbol{P}^{\perp}_{Q_{t-1}} = \boldsymbol{P}^{\perp}_{Q_{t-1}}(\boldsymbol{A} - \boldsymbol{P}^{\perp}_{Q_{t-1}}\boldsymbol{A}\boldsymbol{P}^{\perp}_{Q_{t-1}})\boldsymbol{P}^{\perp}_{Q_{t-1}} = 0.$ □

5.5.2 *Proof of Lemma* 5 We prove the results by induction over $t \in \mathbb{N}$. Let the statement for t be $\mathcal{H}_t$.

Proof of $\mathcal{H}_0$. Recall that $\boldsymbol{h}^1 = \boldsymbol{A}\boldsymbol{q}^0$. Then (a) follows immediately from Lemma C.4, and (b) is from Lemmas C.4, C.6, C.8.

Proof of $\mathcal{H}_t$. We assume $\mathcal{H}_0, \dots, \mathcal{H}_{t-1}$ hold and prove $\mathcal{H}_t$. First note that $\boldsymbol{\alpha}^t \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$ a constant vector in $\mathbb{R}^t$, using $\mathcal{H}_{t-1}(b)$, Lemma C.5 and the non-degeneracy assumption.

(a) We only need to prove the claim for $r = t$.

Consider the case $s < t$. Since $\boldsymbol{h}^{s+1}$ and $\langle \boldsymbol{q}^s, \boldsymbol{q}^t\rangle$ are $\mathfrak{S}_t$-measurable, by Lemma 4,

$$\left(\langle \boldsymbol{h}^{s+1}, \boldsymbol{h}^{t+1}\rangle - \langle \boldsymbol{q}^s, \boldsymbol{q}^t\rangle\right)\Big|_{\mathfrak{S}_t} \stackrel{\mathrm{d}}{=} \left\langle \boldsymbol{P}^{\perp}_{Q_{t-1}}\boldsymbol{h}^{s+1}, \tilde{\boldsymbol{A}}\boldsymbol{q}^t_\perp\right\rangle + \left\langle \boldsymbol{H}^{\mathsf{T}}_{t-1}\boldsymbol{h}^{s+1}, \boldsymbol{\alpha}^t\right\rangle - \langle \boldsymbol{q}^s, \boldsymbol{q}^t\rangle. \tag{98}$$

Note that by $\mathcal{H}_{t-1}(a)$, $\frac{1}{n}\|\boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{h}^{s+1} - \boldsymbol{Q}_{t-1}^\mathsf{T}\boldsymbol{q}^s\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$. Hence,

$$\frac{1}{n}\left|\left\langle \boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{h}^{s+1},\boldsymbol{\alpha}^t\right\rangle - \langle \boldsymbol{q}^s,\boldsymbol{q}^t\rangle\right| = \frac{1}{n}\left|\left\langle \boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{h}^{s+1},\boldsymbol{\alpha}^t\right\rangle - \left\langle \boldsymbol{P}_{Q_{t-1}}\boldsymbol{q}^s,\boldsymbol{q}^t\right\rangle\right| \tag{99}$$

$$= \frac{1}{n}\left|\left\langle \boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{h}^{s+1} - \boldsymbol{Q}_{t-1}^\mathsf{T}\boldsymbol{q}^s,\boldsymbol{\alpha}^t\right\rangle\right| \tag{100}$$

$$\le \frac{1}{n}\left\|\boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{h}^{s+1} - \boldsymbol{Q}_{t-1}^\mathsf{T}\boldsymbol{q}^s\right\|_2 \|\boldsymbol{\alpha}^t\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0, \tag{101}$$

where we use $\boldsymbol{\alpha}^t \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$ (which holds by the induction hypothesis). Furthermore, since $\tilde{\boldsymbol{A}}$ is independent of $\boldsymbol{q}_\perp^t$ and $\boldsymbol{P}_{Q_{t-1}}^\perp\boldsymbol{h}^{s+1}$, by Lemma C.4,

$$\frac{1}{n}\left|\left\langle \boldsymbol{P}_{Q_{t-1}}^\perp\boldsymbol{h}^{s+1},\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\rangle\right| \xrightarrow[n\to\infty]{\mathrm{P}} 0 \tag{102}$$

since $\frac{1}{\sqrt{n}}\|\boldsymbol{h}^{s+1}\|_2$ and $\frac{1}{\sqrt{n}}\|\boldsymbol{q}^t\|_2$ concentrate at finite constants by $\mathcal{H}_{t-1}(b)$ and Lemma C.5 and $\|\boldsymbol{P}_{Q_{t-1}}^\perp\boldsymbol{h}^{s+1}\|_2 \le \|\boldsymbol{h}^{s+1}\|_2$, $\|\boldsymbol{q}_\perp^t\|_2 \le \|\boldsymbol{q}^t\|_2$. It follows that $\frac{1}{n}\langle \boldsymbol{h}^{s+1},\boldsymbol{h}^{t+1}\rangle \overset{\mathrm{P}}{\simeq} \frac{1}{n}\langle \boldsymbol{q}^s,\boldsymbol{q}^t\rangle$.

Consider the case $s = t$. Since $\boldsymbol{q}^t$ is $\mathfrak{S}_t$-measurable, by Lemma 4,

$$\left(\left\|\boldsymbol{h}^{t+1}\right\|_2^2 - \|\boldsymbol{q}^t\|_2^2\right)\Big|_{\mathfrak{S}_t} \overset{\mathrm{d}}{=} \left\|\boldsymbol{P}_{Q_{t-1}}^\perp\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2^2 + 2\left\langle \boldsymbol{P}_{Q_{t-1}}^\perp\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t,\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right\rangle + \|\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\|_2^2 - \|\boldsymbol{q}^t\|_2^2 \tag{103}$$

$$= \left\|\boldsymbol{P}_{Q_{t-1}}^\perp\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2^2 + 2\left\langle \tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t,\boldsymbol{P}_{Q_{t-1}}^\perp\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right\rangle + \left\langle \boldsymbol{\alpha}^t,\boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right\rangle - \|\boldsymbol{q}^t\|_2^2. \tag{104}$$

Again, $\frac{1}{n}\left\langle \tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t,\boldsymbol{P}_{Q_{t-1}}^\perp\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right\rangle \xrightarrow[n\to\infty]{\mathrm{P}} 0$. By independence of $\tilde{\boldsymbol{A}}$ and Lemma C.4, we get

$$\frac{1}{n}\left\|\boldsymbol{P}_{Q_{t-1}}^\perp\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2^2 = \frac{1}{n}\left\|\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2^2 - \frac{1}{n}\left\|\boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2^2 \overset{\mathrm{P}}{\simeq} \frac{1}{n}\|\boldsymbol{q}_\perp^t\|_2^2. \tag{105}$$

Using $\mathcal{H}_{t-1}$ (a) and that $\boldsymbol{\alpha}^t \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$,

$$\frac{1}{n}\left\langle \boldsymbol{\alpha}^t,\left(\boldsymbol{H}_{t-1}^\mathsf{T}\boldsymbol{H}_{t-1} - \boldsymbol{Q}_{t-1}^\mathsf{T}\boldsymbol{Q}_{t-1}\right)\boldsymbol{\alpha}^t\right\rangle \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{106}$$

Notice that $\left\langle \boldsymbol{\alpha}^t,\boldsymbol{Q}_{t-1}^\mathsf{T}\boldsymbol{Q}_{t-1}\boldsymbol{\alpha}^t\right\rangle = \left\|\boldsymbol{q}_\parallel^t\right\|_2^2$. The claim is proven.

(b) First note that $\frac{1}{n}\|\boldsymbol{h}^{t+1}\|_2^2 \overset{\mathrm{P}}{\simeq} \frac{1}{n}\|\boldsymbol{q}^t\|_2^2 \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{K}_{t+1,t+1}$ by $\mathcal{H}_{t-1}$. By Lemma 4,

$$\phi_n(\boldsymbol{x}^0,\boldsymbol{h}^1,\ldots,\boldsymbol{h}^t,\boldsymbol{h}^{t+1})\Big|_{\mathfrak{S}_t} \overset{\mathrm{d}}{=} \phi_n\left(\boldsymbol{x}^0,\boldsymbol{h}^1,\ldots,\boldsymbol{h}^t,\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t - \boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right), \tag{107}$$

and we denote the right-hand side by $\phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t - \boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right)$ for brevity. Since ϕ_n is uniformly pseudo-Lipschitz and by the induction hypothesis,

$$\begin{aligned}
&\left|\phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t - \boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right) - \phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right)\right| \\
&\le L_n C(k,t)\left\{1 + \left(\frac{\|\boldsymbol{x}^0\|_2}{\sqrt{n}}\right)^{k-1} + \sum_{s=1}^{t}\left(\frac{\|\boldsymbol{h}^s\|_2}{\sqrt{n}}\right)^{k-1}\right. \\
&\qquad \left. + \left(\frac{\|\boldsymbol{h}^{t+1}\|_2}{\sqrt{n}}\right)^{k-1} + \left(\frac{\left\|\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\|_2}{\sqrt{n}}\right)^{k-1}\right\} \frac{\left\|\boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2}{\sqrt{n}},
\end{aligned}$$

where $C(k,t)$ is a constant depending only on k and t. We have

$$\frac{1}{\sqrt{n}}\|\boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\|_2 \le \left\|\frac{1}{\sqrt{n}}\boldsymbol{H}_{t-1}\right\|_2 \|\boldsymbol{\alpha}^t\|_2 \le \left\|\frac{1}{\sqrt{n}}\boldsymbol{H}_{t-1}\right\|_{\mathrm{F}} \|\boldsymbol{\alpha}^t\|_2 = \sqrt{\frac{1}{n}\sum_{s=1}^{t}\|\boldsymbol{h}^s\|_2^2}\cdot\|\boldsymbol{\alpha}^t\|_2, \tag{108}$$

which converges to a finite constant by $\mathcal{H}_{t-1}$ (b) and that $\boldsymbol{\alpha}^t \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$. We also have $\frac{1}{\sqrt{n}}\left\|\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2 \le \frac{1}{\sqrt{n}}\left\|\tilde{\boldsymbol{A}}\right\|_{\mathrm{op}}\|\boldsymbol{q}^t\|_2$, which converges to a finite constant due to $\mathcal{H}_{t-1}$ (b) and Theorem C.1. Furthermore, by independence of $\tilde{\boldsymbol{A}}$, recalling $\mathrm{rank}\left(\boldsymbol{P}_{Q_{t-1}}\right) \le t$, we have $\frac{1}{\sqrt{n}}\left\|\boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$ by Lemma C.4. Therefore,

$$\begin{aligned}
\phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t - \boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right) &\overset{\mathrm{P}}{\simeq} \phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right) &(109)\\
&\overset{\mathrm{P}}{\simeq} \phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^{t,*}\right). &(110)
\end{aligned}$$

Notice that $\frac{1}{n}\|\boldsymbol{q}_\perp^t\|_2^2 = \frac{1}{n}\|\boldsymbol{q}^t\|_2^2 - \frac{1}{n}\left\langle\boldsymbol{\alpha}^t, \boldsymbol{Q}_{t-1}^\top\boldsymbol{Q}_{t-1}\boldsymbol{\alpha}^t\right\rangle$, which converges to a constant a^2 due to $\mathcal{H}_{t-1}$ (b) and that $\boldsymbol{\alpha}^t \xrightarrow[n\to\infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$. Then by Lemma C.4, there exists $\boldsymbol{Z} \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$ independent of $\mathfrak{S}_t$ such that

$$\begin{aligned}
\phi_n'\left(\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t - \boldsymbol{P}_{Q_{t-1}}\tilde{\boldsymbol{A}}\boldsymbol{q}_\perp^t + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^t\right) &\overset{\mathrm{P}}{\simeq} \phi_n'\left(a\boldsymbol{Z} + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^{t,*}\right) &(111)\\
&\overset{\mathrm{P}}{\simeq} \mathbb{E}_z\left[\phi_n'\left(a\boldsymbol{Z} + \boldsymbol{H}_{t-1}\boldsymbol{\alpha}^{t,*}\right)\right] &(112)\\
&\overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{x}^0, \boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t, a\boldsymbol{Z} + \sum_{s=1}^{t}\alpha_s^{t,*}\boldsymbol{Z}^s\right)\right], &(113)
\end{aligned}$$

where we use Lemma C.8 in the second step and $\mathcal{H}_{t-1}$ (b) and Lemma C.7 in the third step. (Here with an abuse of notation, we let $\boldsymbol{Z}$ to be on the same joint space as and independent of $\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t$.) The thesis follows immediately from that

$$\left(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t, a\boldsymbol{Z} + \sum_{s=1}^{t} \alpha_s^{t,*} \boldsymbol{Z}^s\right) \stackrel{\mathrm{d}}{=} \left(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t, \boldsymbol{Z}^{t+1}\right), \tag{114}$$

which we now prove.

Let $\tilde{\boldsymbol{Z}} = a\boldsymbol{Z} + \sum_{s=1}^{t} \alpha_s^{t,*} \boldsymbol{Z}^s$ for brevity. Observe that $\tilde{\boldsymbol{Z}}$ is Gaussian with zero mean and i.i.d. entries, $\mathrm{Var}\left[\tilde{Z}_i\right]$ is a constant independent of n and $\mathbb{E}\left[\boldsymbol{Z}^s \tilde{\boldsymbol{Z}}^{\mathsf{T}}\right] = \gamma_s \boldsymbol{I}_n$ for some constant γ_s independent of n, for $1 \leq s \leq t$. It suffices to show that $\mathrm{Var}\ [\tilde{Z}_i] = \mathrm{K}_{t+1,t+1}$ and $\gamma_s = \mathrm{K}_{s,t+1}$. From the above, $\mathcal{H}_t$ (a) and $\mathcal{H}_{t-1}$ (b), we have

$$\mathrm{Var}\left[\tilde{Z}_i\right] = \frac{1}{n}\mathbb{E}\left[\|\tilde{\boldsymbol{Z}}\|_2^2\right] \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\|\boldsymbol{h}^{t+1}\right\|_2^2 \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\|\boldsymbol{q}^t\right\|_2^2 \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{K}_{t+1,t+1}. \tag{115}$$

Similarly, for $s \geq 2$,

$$\gamma_s = \frac{1}{n}\mathbb{E}\left[\left\langle \boldsymbol{Z}^s, \tilde{\boldsymbol{Z}}\right\rangle\right] \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{h}^s, \boldsymbol{h}^{t+1}\right\rangle \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{q}^{s-1}, \boldsymbol{q}^t\right\rangle \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{K}_{s,t+1}, \tag{116}$$

and for $s = 1$,

$$\gamma_1 = \frac{1}{n}\mathbb{E}\left[\left\langle \boldsymbol{Z}^1, \tilde{\boldsymbol{Z}}\right\rangle\right] \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{h}^1, \boldsymbol{h}^{t+1}\right\rangle \stackrel{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{q}^0, \boldsymbol{q}^t\right\rangle \xrightarrow[n\to\infty]{\mathrm{P}} \mathrm{K}_{1,t+1}. \tag{117}$$

This completes the proof. □

5.5.3 *Proof of Lemma 6* For the recursion (53)–(54), define the following quantity for each $t \in \mathbb{N}$,

$$\hat{\boldsymbol{h}}^{t+1} = \boldsymbol{A}\boldsymbol{q}^t - \mathsf{b}_t \boldsymbol{q}^{t-1}, \qquad \mathsf{b}_t = \mathbb{E}\left[\frac{1}{n}\mathrm{div}\, f_t\left(\boldsymbol{Z}^t\right)\right], \tag{118}$$

where we take $\hat{\boldsymbol{h}}^1 = \boldsymbol{A}\boldsymbol{q}^0$. □

LEMMA 13 For any $t \in \mathbb{N}_{>0}$, $\frac{1}{\sqrt{n}}\left\|\boldsymbol{h}^{t+1} - \hat{\boldsymbol{h}}^{t+1}\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$.

Proof. Denoting the claim as $\mathcal{H}_t$, we prove it by induction. The base case $\mathcal{H}_1$ is immediate since $\boldsymbol{h}^1 = \hat{\boldsymbol{h}}^1 = \boldsymbol{A}\boldsymbol{q}^0$. Assuming $\mathcal{H}_1, \ldots, \mathcal{H}_{t-1}$, we prove $\mathcal{H}_t$. Letting $\mathsf{B}_t = \mathrm{diag}\left(0, \mathsf{b}_1, \ldots, \mathsf{b}_t\right) \in \mathbb{R}^{(t+1)\times(t+1)}$ and $\hat{\boldsymbol{H}}_{t-1} = \left[\hat{\boldsymbol{h}}^1 | \ldots | \hat{\boldsymbol{h}}^t\right]$, we have $\hat{\boldsymbol{H}}_{t-1} = \boldsymbol{A}\boldsymbol{Q}_{t-1} - \left[0|\boldsymbol{Q}_{t-2}\right]\mathsf{B}_{t-1}$. Then since $\boldsymbol{P}_{Q_{t-1}}\boldsymbol{q}^t = \boldsymbol{Q}_{t-1}\boldsymbol{\alpha}^t$,

$$\boldsymbol{A}\boldsymbol{q}^t = \boldsymbol{A}\boldsymbol{q}_\perp^t + \boldsymbol{A}\boldsymbol{Q}_{t-1}\boldsymbol{\alpha}^t \tag{119}$$

$$= \boldsymbol{A}\boldsymbol{q}_\perp^t + \left[0|\boldsymbol{Q}_{t-2}\right]\mathsf{B}_{t-1}\boldsymbol{\alpha}^t + \hat{\boldsymbol{H}}_{t-1}\boldsymbol{\alpha}^t. \tag{120}$$

This yields

$$\hat{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^{t+1} = \boldsymbol{P}_{Q_{t-1}} \boldsymbol{A} \boldsymbol{q}_\perp^t - \mathsf{b}_t \boldsymbol{q}^{t-1} + \left[0|\boldsymbol{Q}_{t-2}\right] \mathsf{B}_{t-1} \boldsymbol{\alpha}^t + \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right) \boldsymbol{\alpha}^t \tag{121}$$

$$= \boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \boldsymbol{Q}_{t-1}^\mathsf{T} A \boldsymbol{q}_\perp^t - \mathsf{b}_t \boldsymbol{q}^{t-1} + \left[0|\boldsymbol{Q}_{t-2}\right] \mathsf{B}_{t-1} \boldsymbol{\alpha}^t + \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right) \boldsymbol{\alpha}^t \tag{122}$$

$$\overset{(a)}{=} \boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \hat{\boldsymbol{H}}_{t-1}^\mathsf{T} \boldsymbol{q}_\perp^t - \mathsf{b}_t \boldsymbol{q}^{t-1} + \left[0|\boldsymbol{Q}_{t-2}\right] \mathsf{B}_{t-1} \boldsymbol{\alpha}^t + \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right) \boldsymbol{\alpha}^t \tag{123}$$

$$= \boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \boldsymbol{H}_{t-1}^\mathsf{T} \boldsymbol{q}_\perp^t - \mathsf{b}_t \boldsymbol{q}^{t-1} + \left[0|\boldsymbol{Q}_{t-2}\right] \mathsf{B}_{t-1} \boldsymbol{\alpha}^t \tag{124}$$

$$+ \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right) \boldsymbol{\alpha}^t + \boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right)^\mathsf{T} \boldsymbol{q}_\perp^t$$

$$= \sum_{s=1}^{t} c_s \boldsymbol{q}^{s-1} + \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right) \boldsymbol{\alpha}^t + \boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right)^\mathsf{T} \boldsymbol{q}_\perp^t, \tag{125}$$

where (a) holds because $\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{A} = \left(\boldsymbol{A} \boldsymbol{Q}_{t-1}\right)^\mathsf{T} = \hat{\boldsymbol{H}}_{t-1}^\mathsf{T} + \mathsf{B}_{t-1} \left[0|\boldsymbol{Q}_{t-2}\right]^\mathsf{T}$ and $\boldsymbol{Q}_{t-2}^\mathsf{T} \boldsymbol{P}_{Q_{t-1}}^\perp = 0$ and

$$c_s = \left[\left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \boldsymbol{H}_{t-1}^\mathsf{T} \boldsymbol{q}_\perp^t\right]_s - \mathsf{b}_s \left(-\alpha_{s+1}^t\right)^{\mathbb{I}_{s \neq t}}. \tag{126}$$

By the induction hypothesis,

$$\frac{1}{\sqrt{n}} \left\|\boldsymbol{H}_{t-1} - \hat{\boldsymbol{H}}_{t-1}\right\|_2 \leq \frac{1}{\sqrt{n}} \left\|\boldsymbol{H}_{t-1} - \hat{\boldsymbol{H}}_{t-1}\right\|_\mathrm{F} \xrightarrow[n \to \infty]{\mathrm{P}} 0. \tag{127}$$

By Lemmas 5 and C.5 and the non-degeneracy assumption, $\boldsymbol{\alpha}^t \xrightarrow[n \to \infty]{\mathrm{P}} \boldsymbol{\alpha}^{t,*}$ a constant vector in $\mathbb{R}^t$. Hence,

$$\frac{1}{\sqrt{n}} \left\|\left(\boldsymbol{H}_{t-1} - \hat{\boldsymbol{H}}_{t-1}\right) \boldsymbol{\alpha}^t\right\|_2 \leq \frac{1}{\sqrt{n}} \left\|\boldsymbol{H}_{t-1} - \hat{\boldsymbol{H}}_{t-1}\right\|_\mathrm{F} \left\|\boldsymbol{\alpha}^{t,*}\right\|_2 \xrightarrow[n \to \infty]{\mathrm{P}} 0. \tag{128}$$

By the non-degeneracy assumption,

$$\frac{1}{\sqrt{n}} \left\|\boldsymbol{Q}_{t-1} \left(\boldsymbol{Q}_{t-1}^\mathsf{T} \boldsymbol{Q}_{t-1}\right)^{-1} \left(\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right)^\mathsf{T} \boldsymbol{q}_\perp^t\right\|_2 \leq \frac{1}{\sqrt{n}} \left\|\hat{\boldsymbol{H}}_{t-1} - \boldsymbol{H}_{t-1}\right\|_\mathrm{F} \frac{1}{c_t \sqrt{n}} \left\|\boldsymbol{q}^t\right\|_2 \xrightarrow[n \to \infty]{\mathrm{P}} 0, \tag{129}$$

where $\frac{1}{\sqrt{n}} \left\|\boldsymbol{q}^t\right\|_2$ converges in probability to a finite constant by Lemma 5. We claim that $\frac{1}{\sqrt{n}} c_s \left\|\boldsymbol{q}^{s-1}\right\|_2 \xrightarrow[n \to \infty]{\mathrm{P}} 0$ for $s = 1, \ldots, t$. Then the thesis follows from this claim.

To prove the claim, denoting $R = \frac{1}{n}\left(\boldsymbol{Q}_{t-1}^{\mathsf{T}}\boldsymbol{Q}_{t-1}\right)^{-1}$ for brevity, we note that

$$c_s = \sum_{r=1}^{t} R_{s,r}\frac{1}{n}\left\langle \boldsymbol{h}^r, \boldsymbol{q}^t - \sum_{\ell=1}^{t}\alpha_\ell^t \boldsymbol{q}^{\ell-1}\right\rangle - \mathsf{b}_s\left(-\alpha_{s+1}^t\right)^{\mathbb{I}_{s\neq t}}. \tag{130}$$

We now analyze c_s. By Lemma 5,

$$\frac{1}{n}\langle \boldsymbol{h}^r, \boldsymbol{q}^0\rangle \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\frac{1}{n}\langle \boldsymbol{Z}^r, f_0(\boldsymbol{x}^0)\rangle\right] = 0 \tag{131}$$

since $\boldsymbol{Z}^r$ has zero mean. By Lemmas 5 and C.2, for $j = 2, \ldots, t-1$,

$$\frac{1}{n}\langle \boldsymbol{h}^r, \boldsymbol{q}^j\rangle \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\frac{1}{n}\left\langle \boldsymbol{Z}^r, f_j(\boldsymbol{Z}^j)\right\rangle\right] \tag{132}$$

$$= \mathrm{K}_{r,j}\mathbb{E}\left[\frac{1}{n}\mathrm{div}\ f_j\left(\boldsymbol{Z}^j\right)\right] \tag{133}$$

$$\overset{\mathrm{P}}{\simeq} \frac{1}{n}\left\langle \boldsymbol{q}^{r-1}, \boldsymbol{q}^{j-1}\right\rangle \mathsf{b}_j. \tag{134}$$

Therefore,

$$c_s \overset{\mathrm{P}}{\simeq} \left\{\sum_{r=1}^{t} R_{s,r}\frac{1}{n}\left\langle \boldsymbol{q}^{r-1}, \mathsf{b}_t\boldsymbol{q}^{t-1} - \sum_{\ell=2}^{t}\alpha_1^t \mathsf{b}_{\ell-1}\boldsymbol{q}^{\ell-2}\right\rangle - \mathsf{b}_s\left(-\alpha_{s+1}^t\right)^{\mathbb{I}_{s\neq t}}\right\}. \tag{135}$$

Identifying $\frac{1}{n}\left\langle \boldsymbol{q}^{r-1}, \boldsymbol{q}^{j-1}\right\rangle = \left(R^{-1}\right)_{r,j}$, we get

$$c_s \overset{\mathrm{P}}{\simeq} \left\{\mathsf{b}_t\mathbb{I}_{t=s} - \sum_{\ell=2}^{t}\alpha_\ell^t \mathsf{b}_{\ell-1}\mathbb{I}_{\ell-1=s} - \mathsf{b}_s\left(-\alpha_{s+1}^t\right)^{\mathbb{I}_{s\neq t}}\right\}, \tag{136}$$

i.e. $c_s \xrightarrow[n\to\infty]{\mathrm{P}} 0$. Finally, since $\frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^{s-1}\right\|_2$ converges in probability to a finite constant by Lemma 5, the claim is proven. □

Proof of Lemma 6 Let $\mathcal{H}_t$ be the statement $\frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^t - \boldsymbol{m}^t\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$ and $\frac{1}{\sqrt{n}}\left\|\boldsymbol{h}^{t+1} - \boldsymbol{x}^{t+1}\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$. We prove it by induction. The base case $\mathcal{H}_0$ is trivial because $\boldsymbol{q}^0 = \boldsymbol{m}^0$ and $\boldsymbol{h}^1 = \boldsymbol{x}^1$.

We now assume $\mathcal{H}_{t-1}$ is true and we show $\mathcal{H}_t$. We have

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^t - \boldsymbol{m}^t\right\|_2 = \frac{1}{\sqrt{n}}\left\|f_t\left(\boldsymbol{h}^t\right) - f_t\left(\boldsymbol{x}^t\right)\right\|_2 \leq L_t\frac{1}{\sqrt{n}}\left\|\boldsymbol{h}^t - \boldsymbol{x}^t\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0, \tag{137}$$

using that f_t is uniformly Lipschitz and the induction hypothesis $\mathcal{H}_{t-1}$. Further, we will prove that $\frac{1}{\sqrt{n}}\left\|\hat{\boldsymbol{h}}^{t+1}-\boldsymbol{x}^{t+1}\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$, which together with Lemma 13 yields $\mathcal{H}_t$. We have

$$\hat{\boldsymbol{h}}^{t+1}-\boldsymbol{x}^{t+1}=A(\boldsymbol{q}^t-\boldsymbol{m}^t)-\mathsf{b}_t\left(\boldsymbol{q}^{t-1}-\boldsymbol{m}^{t-1}\right), \tag{138}$$

thus by Theorem C.1 and $\mathcal{H}_{t-1}$,

$$\frac{1}{\sqrt{n}}\left\|\hat{\boldsymbol{h}}^{t+1}-\boldsymbol{x}^{t+1}\right\|_2 \leq \|\boldsymbol{A}\|_{\mathrm{op}}\frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^t-\boldsymbol{m}^t\right\|_2+\mathsf{b}_t\frac{1}{\sqrt{n}}\left\|\boldsymbol{q}^{t-1}-\boldsymbol{m}^{t-1}\right\|_2 \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{139}$$

This concludes the induction. □

5.5.4 *Proof of Lemma* 8 Let us first check assumption (A6) for the perturbed setting $\left\{\boldsymbol{x}^0, f_t^{\epsilon y}\right\}$. Consider $s,t\geq 1$, K a 2×2 covariance matrix and $(\boldsymbol{Z}^s,\boldsymbol{Z}^t)\in(\mathbb{R}^n)^2$, $(\boldsymbol{Z}^s,\boldsymbol{Z}^t)\sim\mathsf{N}\left(0,\mathrm{K}\otimes\boldsymbol{I}_n\right)$. Note that K is deterministic, not depending on the perturbation $\boldsymbol{y}$. We denote the expectation over $(\boldsymbol{Z}^s,\boldsymbol{Z}^t)$ as $\mathbb{E}_Z$. We have

$$\mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s^{\epsilon y}(\boldsymbol{Z}^s), f_t^{\epsilon y}(\boldsymbol{Z}^t)\right\rangle\right]=\mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s(\boldsymbol{Z}^s), f_t(\boldsymbol{Z}^t)\right\rangle\right]+\epsilon\mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s(\boldsymbol{Z}^s),\boldsymbol{y}^t\right\rangle\right] \tag{140}$$

$$+\epsilon\mathbb{E}_Z\left[\frac{1}{n}\left\langle \boldsymbol{y}^s, f_t(\boldsymbol{Z}^t)\right\rangle\right]+\epsilon^2\frac{1}{n}\left\langle\boldsymbol{y}^s,\boldsymbol{y}^t\right\rangle \tag{141}$$

$$=\mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s(\boldsymbol{Z}^s), f_t(\boldsymbol{Z}^t)\right\rangle\right]+\epsilon\frac{1}{n}\left\langle\mathbb{E}_Z\left[f_s(\boldsymbol{Z}^s)\right],\boldsymbol{y}^t\right\rangle \tag{142}$$

$$+\epsilon\frac{1}{n}\left\langle\boldsymbol{y}^s,\mathbb{E}_Z\left[f_t(\boldsymbol{Z}^t)\right]\right\rangle+\epsilon^2\frac{1}{n}\left\langle\boldsymbol{y}^s,\boldsymbol{y}^t\right\rangle. \tag{143}$$

- The first term does not depend on the perturbation and is thus deterministic. By assumption (A6) for the setting $\left\{\boldsymbol{x}^0, f_t\right\}$, $\mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s(\boldsymbol{Z}^s), f_t(\boldsymbol{Z}^t)\right\rangle\right]$ converges to a (deterministic) limit.
- The second term is Gaussian, with mean zero and variance

$$\frac{1}{n^2}\left\|\mathbb{E}\left[f_s(\boldsymbol{Z}^s)\right]\right\|_2^2\leq\frac{1}{n^2}\mathbb{E}\left[\left\|f_s(\boldsymbol{Z}^s)\right\|_2^2\right]\leq\frac{C}{n}, \tag{144}$$

for C a constant large enough, using again the assumption (A6) for the setting $\left\{\boldsymbol{x}^0, f_t\right\}$. Thus, $\frac{1}{n}\left\langle\mathbb{E}_Z\left[f_s(\boldsymbol{Z}^s)\right],\boldsymbol{y}^t\right\rangle$ is a Gaussian random variable, of standard deviation smaller than $\sqrt{C}/\sqrt{n}$. Then if $\eta>0$,

$$\Pr\left(\left|\frac{1}{n}\left\langle\mathbb{E}_Z\left[f_s(\boldsymbol{Z}^s)\right],\boldsymbol{y}^t\right\rangle\right|\geq\eta\right)\leq\Pr\left(|\mathsf{N}(0,1)|\geq\eta\frac{\sqrt{n}}{\sqrt{C}}\right)\leq\frac{\sqrt{C}}{\eta\sqrt{n}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{\eta^2 n}{C}\right), \tag{145}$$

which is summable. Using Borel–Cantelli's lemma, it is then easy to show that

$$\frac{1}{n}\left\langle \mathbb{E}_Z\left[f_s(\boldsymbol{Z}^s)\right], \boldsymbol{y}^t\right\rangle \xrightarrow[n\to\infty]{a.s} 0. \tag{146}$$

- The treatment of the third term is the same as for the second term.
- Using the law of large numbers, we get that

$$\frac{1}{n}\left\langle \boldsymbol{y}^s, \boldsymbol{y}^t\right\rangle \xrightarrow[n\to\infty]{a.s.} \mathbb{I}_{s=t}. \tag{147}$$

Putting things together, we get almost surely

$$\lim_{n\to\infty} \mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s^{\epsilon y}(\boldsymbol{Z}^s), f_t^{\epsilon y}(\boldsymbol{Z}^t)\right\rangle\right] = \lim_{n\to\infty} \mathbb{E}_Z\left[\frac{1}{n}\left\langle f_s(\boldsymbol{Z}^s), f_t(\boldsymbol{Z}^t)\right\rangle\right] + \epsilon^2 \mathbb{I}_{s=t}. \tag{148}$$

The proof of assumptions (A4) and (A5) are very similar, here we only state the resulting expressions: almost surely,

$$\lim_{n\to\infty} \frac{1}{n}\left\langle f_0^{\epsilon y}(\boldsymbol{x}^0), f_0^{\epsilon y}(\boldsymbol{x}^0)\right\rangle = \lim_{n\to\infty} \frac{1}{n}\left\langle f_0(\boldsymbol{x}^0), f_0(\boldsymbol{x}^0)\right\rangle + \epsilon^2, \tag{149}$$

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\left\langle f_0^{\epsilon y}(\boldsymbol{x}^0), f_t^{\epsilon y}(\boldsymbol{Z}^t)\right\rangle\right] = \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\left\langle f_0(\boldsymbol{x}^0), f_t(\boldsymbol{Z}^t)\right\rangle\right]. \tag{150}$$

Using Equations (148), (149) and (150), it is a simple induction that the state evolution for the perturbed setting $\{\boldsymbol{x}^0, f_t^{\epsilon y}\}$ is indeed non-random almost surely. □

5.5.5 *Proof of Lemma 9* By definition,

$$\boldsymbol{q}_\perp^{\epsilon y,t} = \boldsymbol{P}_{Q_{t-1}}^\perp f_t(\boldsymbol{h}^{\epsilon y,t}) + \epsilon \boldsymbol{P}_{Q_{t-1}}^\perp \boldsymbol{y}^t. \tag{151}$$

If we denote $\mathcal{F}_t$ as the σ-algebra generated by $\boldsymbol{h}^{\epsilon y,1}, \ldots, \boldsymbol{h}^{\epsilon y,t}, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^{t-1}$, it follows that

$$\boldsymbol{q}_\perp^{\epsilon y,t}|_{\mathcal{F}_t} \sim \mathsf{N}\left(\boldsymbol{P}_{Q_{t-1}}^\perp f_t(\boldsymbol{h}^{\epsilon y,t}), \epsilon^2 \boldsymbol{P}_{Q_{t-1}}^\perp\right). \tag{152}$$

When $n > t$, this conditional distribution is almost surely non-zero. Thus, when $n \geq t$, the matrix $\boldsymbol{Q}_{t-1}$ has full column rank.

To lower bound the minimum singular value of $\boldsymbol{Q}_{t-1}$, a more careful treatment is required. Using [5, Lemma 8], it is sufficient to check that there exists a constant c_ϵ such that almost surely, for n sufficiently large,

$$\frac{1}{n}\|\boldsymbol{q}_\perp^{\epsilon y,t}\|^2 \geq c_\epsilon. \tag{153}$$

We have

$$\Pr\left(\frac{1}{n}\left\|\boldsymbol{q}_{\perp}^{\epsilon y,t}\right\|^2 \le c_\epsilon \middle| \mathcal{F}_t\right) = \Pr\left(\left\|\mathsf{N}\left(\boldsymbol{P}_{Q_{t-1}}^{\perp} f_t(\boldsymbol{h}^{\epsilon y,t}), \epsilon^2 \boldsymbol{P}_{Q_{t-1}}^{\perp}\right)\right\|^2 \le c_\epsilon n \middle| \mathcal{F}_t\right) \tag{154}$$

$$\le \Pr\left(\left\|\mathsf{N}\left(0, \epsilon^2 \boldsymbol{P}_{Q_{t-1}}^{\perp}\right)\right\|^2 \le c_\epsilon n \middle| \mathcal{F}_t\right) \tag{155}$$

$$= \Pr\left(\chi_{n-t} \le \frac{c_\epsilon n}{\epsilon^2}\right) \tag{156}$$

$$= \Pr\left(\frac{\chi_{n-t}}{n-t} \le \frac{c_\epsilon}{\epsilon^2}\frac{n}{n-t}\right). \tag{157}$$

We can choose c_ϵ such that $c_\epsilon/\epsilon^2 = 1/4$ and consider only the case $n \ge 2t$ so that $n/(n-t) \le 2$. We then get

$$\Pr\left(\frac{1}{n}\left\|\boldsymbol{q}_{\perp}^{\epsilon y,t}\right\|^2 \le c_\epsilon \middle| \mathcal{F}_t\right) \le \Pr\left(\frac{\chi_{n-t}}{n-t} \le \frac{1}{2}\right). \tag{158}$$

Using concentration of the chi-squared variable, it is easy to show that $\Pr\left(\frac{\chi_{n-t}}{n-t} \le \frac{1}{2}\right)$ is summable over n. Taking expectation of the last inequality, we get

$$\sum_n \Pr\left(\frac{1}{n}\left\|\boldsymbol{q}_{\perp}^{\epsilon y,t}\right\|^2 \le c_\epsilon\right) < +\infty. \tag{159}$$

Then Borel–Cantelli's lemma concludes the proof. □

5.5.6 *Proof of Lemma* 10 Define k as the order of the sequence $\{\phi_n\}$ of uniformly pseudo-Lipschitz functions and L as its pseudo-Lipschitz constant. Under any coupling of $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Z}}$,

$$\left|\mathbb{E}\left[\phi_n(\boldsymbol{Z})\right] - \mathbb{E}\left[\phi_n\left(\tilde{\boldsymbol{Z}}\right)\right]\right| \le L\mathbb{E}\left[\left(1 + \left(\frac{\|\boldsymbol{Z}\|_2}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\tilde{\boldsymbol{Z}}\|_2}{\sqrt{n}}\right)^{k-1}\right)\frac{\|\boldsymbol{Z}-\tilde{\boldsymbol{Z}}\|_2}{\sqrt{n}}\right] \tag{160}$$

$$\le L\mathbb{E}\left[\left(1 + \left(\frac{\|\boldsymbol{Z}\|_2}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\tilde{\boldsymbol{Z}}\|_2}{\sqrt{n}}\right)^{k-1}\right)^2\right]^{1/2} \frac{1}{\sqrt{n}}\mathbb{E}\left[\|\boldsymbol{Z}-\tilde{\boldsymbol{Z}}\|_2^2\right]^{1/2}. \tag{161}$$

Taking the infimum over all possible coupling of $\boldsymbol{Z} \sim \mathsf{N}\left(0, \mathrm{K} \otimes \boldsymbol{I}_n\right)$ and $\tilde{\boldsymbol{Z}} \sim \mathsf{N}\left(0, \tilde{\mathrm{K}} \otimes \boldsymbol{I}_n\right)$, one gets a bound involving the Wasserstein distance W_2:

$$\left|\mathbb{E}\left[\phi_n(\boldsymbol{Z})\right] - \mathbb{E}\left[\phi_n(\tilde{\boldsymbol{Z}})\right]\right| \tag{162}$$

$$\le \sqrt{3}L\left(1 + \frac{\mathbb{E}\left[\|\boldsymbol{Z}\|_2^{2(k-1)}\right]}{n^{k-1}} + \frac{\mathbb{E}\left[\|\tilde{\boldsymbol{Z}}\|_2^{2(k-1)}\right]}{n^{k-1}}\right)^{1/2} \frac{1}{\sqrt{n}} W_2\left(\mathsf{N}\left(0, \mathrm{K}\otimes\boldsymbol{I}_n\right), \mathsf{N}\left(0, \tilde{\mathrm{K}}\otimes\boldsymbol{I}_n\right)\right). \tag{163}$$

We then use the two following identities for the Wasserstein distance:

$$W_2(\mu \otimes \nu, \mu' \otimes \nu')^2 = W_2(\mu, \mu')^2 + W_2(\nu, \nu')^2, \tag{164}$$

$$W_2\big(\mathsf{N}(0, \mathrm{K}), \mathsf{N}(0, \tilde{\mathrm{K}})\big)^2 = \mathrm{Tr}\Big(\mathrm{K} + \tilde{\mathrm{K}} - 2\big(\mathrm{K}^{1/2}\tilde{\mathrm{K}}\mathrm{K}^{1/2}\big)^{1/2}\Big). \tag{165}$$

For a proof of the second identity, see [20, Proposition 7]. It follows that

$$W_2\left(\mathsf{N}\left(0, \mathrm{K} \otimes \boldsymbol{I}_n\right), \mathsf{N}\big(0, \tilde{\mathrm{K}} \otimes \boldsymbol{I}_n\big)\right)^2 = n\mathrm{Tr}\left(\mathrm{K} + \tilde{\mathrm{K}} - 2\big(\mathrm{K}^{1/2}\tilde{\mathrm{K}}\mathrm{K}^{1/2}\big)^{1/2}\right). \tag{166}$$

Moreover, $Z \overset{\mathrm{d}}{=} \left(\mathrm{K}^{1/2} \otimes \boldsymbol{I}_n\right) X$ where $X \sim \mathsf{N}\left(0, \boldsymbol{I}_{nt}\right)$. Thus,

$$\mathbb{E}\left[\|Z\|_2^{2(k-1)}\right] \le \|\mathrm{K}^{1/2} \otimes \boldsymbol{I}_n\|_2^{2(k-1)} \mathbb{E}\left[\|X\|_2^{2(k-1)}\right] = \|\mathrm{K}\|_{\mathrm{op}}^{k-1} \mathbb{E}\left[\left(\chi^2_{nt}\right)^{k-1}\right]. \tag{167}$$

Using expressions for moments of chi-square variables, we get

$$\mathbb{E}\left[\left(\chi^2_{n\times t}\right)^{k-1}\right] = nt(nt+2)\ldots(nt+2(k-2)) \le n^{k-1}t^{k-1}(1+2(k-2))^{k-1} = C(k,t)n^{k-1} \tag{168}$$

for a constant $C(k, t)$ that depends only on k and t. Back to inequality (163),

$$\left|\mathbb{E}\left[\phi_n(\boldsymbol{Z})\right] - \mathbb{E}\left[\phi_n(\tilde{\boldsymbol{Z}})\right]\right| \tag{169}$$

$$\le \sqrt{3}L\left(1 + C(k,t)\left(\|\mathrm{K}\|_{\mathrm{op}}^{k-1} + \big\|\tilde{\mathrm{K}}\big\|_{\mathrm{op}}^{k-1}\right)\right)^{1/2}\left(\mathrm{Tr}\left(\mathrm{K} + \tilde{\mathrm{K}} - 2\big(\mathrm{K}^{1/2}\tilde{\mathrm{K}}\mathrm{K}^{1/2}\big)^{1/2}\right)\right)^{1/2}. \tag{170}$$

Notice that this bound is independent of n and converges to 0 as $\tilde{\mathrm{K}} \to \mathrm{K}$. □

5.5.7 *Proof of Lemma* 11 This lemma will be shown by induction.

Initialization. According to (149),

$$\mathrm{K}^{\epsilon}_{1,1} = \mathrm{K}_{1,1} + \epsilon^2 \xrightarrow[\epsilon \to 0]{} \mathrm{K}_{1,1}. \tag{171}$$

Induction. Let t be a non-negative integer. Assume that by the induction hypothesis, for any $r, s \le t$, $\mathrm{K}^{\epsilon}_{r,s} \to \mathrm{K}_{r,s}$. Then

$$\mathrm{K}^{\epsilon}_{s+1,t+1} \overset{\text{a.s.}}{=} \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\left\langle f^{\epsilon y}_s\left(\boldsymbol{Z}^{\epsilon,s}\right), f^{\epsilon y}_t\left(\boldsymbol{Z}^{\epsilon,t}\right)\right\rangle\right], \tag{172}$$

where $(\boldsymbol{Z}^{\epsilon,s}, \boldsymbol{Z}^{\epsilon,t})$ is a Gaussian vector, whose covariance is determined by $\mathrm{K}^{\epsilon}_{s,s}$, $\mathrm{K}^{\epsilon}_{t,t}$ and $\mathrm{K}^{\epsilon}_{s,t}$. Using (148), we have

$$\mathrm{K}^{\epsilon}_{s+1,t+1} \overset{\text{a.s.}}{=} \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\left\langle f_s\left(\boldsymbol{Z}^{\epsilon,s}\right), f_t\left(\boldsymbol{Z}^{\epsilon,t}\right)\right\rangle\right] + \epsilon^2 \mathbb{I}_{s=t}. \tag{173}$$

The sequence of functions $(z^s, z^t) \mapsto \frac{1}{n}\left\langle f_s\left(z^s\right), f_t\left(z^t\right)\right\rangle$ is uniformly pseudo-Lipschitz by Lemma C.5; thus, Lemma 10 and the induction hypothesis jointly ensure that

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\left\langle f_s\left(\boldsymbol{Z}^{\epsilon,s}\right), f_t\left(\boldsymbol{Z}^{\epsilon,t}\right)\right\rangle\right] = \lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\left\langle f_s\left(\boldsymbol{Z}^{s}\right), f_t\left(\boldsymbol{Z}^{t}\right)\right\rangle\right] = \mathrm{K}_{s+1,t+1}, \tag{174}$$

where $(\boldsymbol{Z}^s, \boldsymbol{Z}^t) \in (\mathbb{R}^n)^2$, $(\boldsymbol{Z}^s, \boldsymbol{Z}^t) \sim \mathsf{N}\left(0, \mathrm{K}\otimes \boldsymbol{I}_n\right)$. Thus, we indeed get

$$\mathrm{K}^{\epsilon}_{s+1,t+1} \xrightarrow[\epsilon\to 0]{} \mathrm{K}_{s+1,t+1}. \tag{175}$$

To finish the induction reasoning, one can check similarly that $\mathrm{K}^{\epsilon}_{1,t+1} \xrightarrow[\epsilon\to 0]{} \mathrm{K}_{1,t+1}$. □

5.5.8 *Proof of Lemma* 12 First, it is easy to check by induction that there exist constants $\tilde{C}_t$, $\tilde{C}'_t$ and $\tilde{C}''_t$ independent of n such that for all $\epsilon \le 1$, w.h.p.

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{m}^{\epsilon y,t}\right\|_2 \le \tilde{C}'_t, \tag{176}$$

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{x}^{\epsilon y,t+1}\right\|_2 \le \tilde{C}_t. \tag{177}$$

Indeed, one only needs to use that the functions involved are uniformly Lipschitz and Theorem C.1. Note that these inequalities hold for the original AMP iterates by taking $\epsilon = 0$.

We now prove our lemma by induction.

Initialization. We have

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{m}^{\epsilon y,0} - \boldsymbol{m}^0\right\|_2 = \epsilon\frac{\left\|\boldsymbol{y}^0\right\|_2}{\sqrt{n}} \le 2\epsilon \quad \text{w.h.p.}, \tag{178}$$

by the law of large numbers. Thus, we choose $h'_0(\epsilon) = 2\epsilon$. Furthermore,

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{x}^{\epsilon y,1} - \boldsymbol{x}^1\right\|_2 \le \|\boldsymbol{A}\|_{\mathrm{op}}\frac{1}{\sqrt{n}}\left\|\boldsymbol{m}^{\epsilon y,0} - \boldsymbol{m}^0\right\|_2 \le 6\epsilon \quad \text{w.h.p.}, \tag{179}$$

by Theorem C.1. Thus, we choose $h_0(\epsilon) = 6\epsilon$.

Induction. We assume here that $\mathrm{K}_{1,1}, \dots, \mathrm{K}_{t,t} > 0$. By induction hypothesis, we have already defined $h_0(\epsilon), h'_0(\epsilon), \dots, h_{t-1}(\epsilon), h'_{t-1}(\epsilon)$. We now choose $h_t(\epsilon)$ and $h'_t(\epsilon)$. We have

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{m}^{\epsilon y,t} - \boldsymbol{m}^t\right\|_2 \le \frac{1}{\sqrt{n}}\left\|f_t\left(\boldsymbol{x}^{\epsilon y,t}\right) - f_t\left(\boldsymbol{x}^t\right) + \epsilon\boldsymbol{y}^t\right\|_2 \tag{180}$$

$$\le L_t\frac{1}{\sqrt{n}}\left\|\boldsymbol{x}^{\epsilon y,t} - \boldsymbol{x}^t\right\|_2 + \epsilon\frac{\left\|\boldsymbol{y}^t\right\|_2}{\sqrt{n}} \le L_t h_{t-1}(\epsilon) + 2\epsilon \quad \text{w.h.p.} \tag{181}$$

using that f_t is uniformly Lipschitz with Lipschitz constant L_t. Thus, we choose $h_t'(\epsilon) = L_t h_{t-1}(\epsilon) + 2\epsilon$, which converges to zero as $\epsilon \to 0$. Furthermore,

$$\frac{1}{\sqrt{n}} \left\| \boldsymbol{x}^{\epsilon y,t+1} - \boldsymbol{x}^{t+1} \right\|_2 \le \|\boldsymbol{A}\|_{\text{op}} \frac{1}{\sqrt{n}} \left\| \boldsymbol{m}^{\epsilon y,t} - \boldsymbol{m}^t \right\|_2 + \frac{1}{\sqrt{n}} \left\| b_t^{\epsilon y} \boldsymbol{m}^{\epsilon y,t-1} - b_t \boldsymbol{m}^{t-1} \right\|_2 \tag{182}$$

$$\le 3h_t'(\epsilon) + \frac{1}{\sqrt{n}} \left\| b_t^{\epsilon y} \boldsymbol{m}^{\epsilon y,t-1} - b_t \boldsymbol{m}^{t-1} \right\|_2 \quad \text{w.h.p.} \tag{183}$$

by Theorem C.1. We have from (176) the following:

$$\frac{1}{\sqrt{n}} \left\| \mathsf{b}_t^{\epsilon y} \boldsymbol{m}^{\epsilon y,t-1} - \mathsf{b}_t \boldsymbol{m}^{t-1} \right\|_2 \le \left| \mathsf{b}_t^{\epsilon y} \right| \frac{1}{\sqrt{n}} \left\| \boldsymbol{m}^{\epsilon y,t-1} - \boldsymbol{m}^{t-1} \right\|_2 + \left| \mathsf{b}_t^{\epsilon y} - \mathsf{b}_t \right| \frac{1}{\sqrt{n}} \left\| \boldsymbol{m}^{t-1} \right\|_2 \tag{184}$$

$$\le L h_{t-1}'(\epsilon) + \left| \mathsf{b}_t^{\epsilon y} - \mathsf{b}_t \right| \tilde{C}_{t-1}'. \tag{185}$$

Since $\mathrm{K}_{t,t}^{\epsilon} \to \mathrm{K}_{t,t}$ when $\epsilon \to 0$ from Lemma 11 and $\mathrm{K}_{t,t} > 0$, we have $\mathrm{K}_{t,t}^{\epsilon} > 0$ for sufficiently small ϵ. Then using Lemma C.2, with $\boldsymbol{Z} \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$, we get

$$\left| \mathsf{b}_t^{\epsilon y} - \mathsf{b}_t \right| = \left| \mathbb{E}\left[\frac{1}{n} \text{div } f_t\left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} \boldsymbol{Z} \right) \right] - \mathbb{E}\left[\frac{1}{n} \text{div } f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) \right] \right| \tag{186}$$

$$= \left| \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} \mathbb{E}\left[\frac{1}{n} \left\langle \boldsymbol{Z}, f_t\left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} \boldsymbol{Z} \right) \right\rangle \right] - \frac{1}{\sqrt{\mathrm{K}_{t,t}}} \mathbb{E}\left[\frac{1}{n} \left\langle \boldsymbol{Z}, f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) \right\rangle \right] \right| \tag{187}$$

$$\le \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} \left| \mathbb{E}\left[\frac{1}{n} \left\langle \boldsymbol{Z}, f_t\left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} \boldsymbol{Z} \right) - f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) \right\rangle \right] \right|$$

$$+ \left| \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} - \frac{1}{\sqrt{\mathrm{K}_{t,t}}} \right| \mathbb{E}\left[\frac{1}{n} \left| \left\langle \boldsymbol{Z}, f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) \right\rangle \right| \right] \tag{188}$$

$$\le \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} \mathbb{E}\left[\frac{1}{n} \|\boldsymbol{Z}\|_2^2 \right]^{1/2} \mathbb{E}\left[\frac{1}{n} \left\| f_t\left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} \boldsymbol{Z} \right) - f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) \right\|_2^2 \right]^{1/2} \tag{189}$$

$$+ \left| \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} - \frac{1}{\sqrt{\mathrm{K}_{t,t}}} \right| \mathbb{E}\left[\left| \frac{1}{n} \left\langle \boldsymbol{Z}, f_t(0) \right\rangle \right| + \left| \frac{1}{n} \left\langle \boldsymbol{Z}, f_t\left(\sqrt{\mathrm{K}_{t,t}} \boldsymbol{Z} \right) - f_t(0) \right\rangle \right| \right] \tag{190}$$

$$\le \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} \mathbb{E}\left[\frac{1}{n} \|\boldsymbol{Z}\|_2^2 \right] L_t \left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} - \sqrt{\mathrm{K}_{t,t}} \right)$$

$$+ \left| \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} - \frac{1}{\sqrt{\mathrm{K}_{t,t}}} \right| \left(\frac{\left\| f_t(0) \right\|_2^2}{n} + \mathbb{E}\left[\frac{1}{n} \|\boldsymbol{Z}\|_2^2 \right] L_t \sqrt{\mathrm{K}_{t,t}} \right) \tag{191}$$

$$\le \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} L_t \left(\sqrt{\mathrm{K}_{t,t}^{\epsilon}} - \sqrt{\mathrm{K}_{t,t}} \right) + \left| \frac{1}{\sqrt{\mathrm{K}_{t,t}^{\epsilon}}} - \frac{1}{\sqrt{\mathrm{K}_{t,t}}} \right| \left(\frac{\left\| f_t(0) \right\|_2^2}{\sqrt{n}} + L_t \sqrt{\mathrm{K}_{t,t}} \right). \tag{192}$$

Since the quantity $\left\| f_t(0) \right\|_2^2 / n$ is upper bounded by a constant independent of n, we can plug (192) into (185) and choose correspondingly a function $h_t(\epsilon)$ such that $h_t(\epsilon) \to 0$ when $\epsilon \to 0$, enabled by the fact $\mathrm{K}_{t,t}^{\epsilon} \to \mathrm{K}_{t,t}$. □

6. Proof of Theorem 1 and Corollary 2 (asymmetric AMP)

Proof of Theorem 1 We reduce this case to the asymmetric case, as in [21]. Consider

$$\boldsymbol{A}_s = \sqrt{\frac{\delta}{\delta+1}} \begin{bmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^{\mathsf{T}} & \boldsymbol{C} \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}^0 = \begin{bmatrix} 0 \\ \boldsymbol{u}^0 \end{bmatrix},$$

where $\boldsymbol{B} \sim \mathrm{GOE}(m)$ and $\sqrt{\delta}\boldsymbol{C} \sim \mathrm{GOE}(n)$ are independent of each other and of $\boldsymbol{A}$. It is easy to see that $\boldsymbol{A}_s \sim \mathrm{GOE}(N)$, where $N = m + n$. We further let $f_t : \mathbb{R}^N \to \mathbb{R}^N$ be such that

$$f_{2t+1}(\boldsymbol{x}) = \sqrt{\frac{\delta+1}{\delta}} \begin{bmatrix} g_t(x_1, \dots, x_m) \\ 0 \end{bmatrix}, \quad f_{2t}(\boldsymbol{x}) = \sqrt{\frac{\delta+1}{\delta}} \begin{bmatrix} 0 \\ e_t(x_{m+1}, \dots, x_N) \end{bmatrix}$$

for any $\boldsymbol{x} \in \mathbb{R}^N$. We can define the symmetric AMP recursion $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$:

$$\boldsymbol{x}^{t+1} = \boldsymbol{A}_s \boldsymbol{m}^t - \mathsf{b}_t \boldsymbol{m}^{t-1}, \tag{193}$$

$$\boldsymbol{m}^t = f_t(\boldsymbol{x}^t), \tag{194}$$

$$\mathsf{b}_t = \mathbb{E}\left[\frac{1}{N} \mathrm{div}\, f_t(\boldsymbol{Z}^t)\right] \tag{195}$$

along with its state evolution $\{\mathrm{K}_{s,t} | f_t, \boldsymbol{x}^0\}$ (see Section 4 for a more complete definition of these quantities). Note that assumptions (A1)–(A6) are satisfied because of (B1)–(B6).

Note that here $\mathrm{K}_{2t,2t+1} = 0$. It is also easy to identify that

$$\boldsymbol{v}^t = \left(x_1^{2t+1}, \dots, x_m^{2t+1}\right), \tag{196}$$

$$\boldsymbol{u}^t = \left(x_{m+1}^{2t}, \dots, x_N^{2t}\right), \tag{197}$$

$$\Sigma_{s,t} = \mathrm{K}_{2s+1,2t+1}, \tag{198}$$

$$\mathrm{T}_{s,t} = \mathrm{K}_{2s,2t}. \tag{199}$$

Applying Theorem 3 to the AMP recursion $\{\boldsymbol{x}^t, \boldsymbol{m}^t | f_t, \boldsymbol{x}^0\}$ shows our theorem. □

Proof of Corollary 2 The proof is by induction over t. Let $\mathcal{H}_t$ be the claim that $\|\boldsymbol{u}^s - \hat{\boldsymbol{u}}^s\|_2/\sqrt{n} \overset{\mathrm{P}}{\simeq} 0$ for all $s \le t$ and $\|\boldsymbol{v}^s - \hat{\boldsymbol{v}}^s\|_2/\sqrt{n} \overset{\mathrm{P}}{\simeq} 0$ for all $s \le t-1$. The initial conditions imply immediately $\mathcal{H}_0$.

We now prove that $\mathcal{H}_t$ implies $\mathcal{H}_{t+1}$. Taking the difference of Equations (2) and (38) and using triangular inequality, we get

$$\|\boldsymbol{v}^t - \hat{\boldsymbol{v}}^t\|_2 \le \|\boldsymbol{A}\|_{\text{op}}\|e_t(\boldsymbol{u}^t) - e_t(\hat{\boldsymbol{u}}^t)\|_2 + |\mathsf{b}_t - \hat{\mathsf{b}}_t|\, \|g_{t-1}(\boldsymbol{v}^{t-1})\|_2 + |\hat{\mathsf{b}}_t|\, \|g_{t-1}(\hat{\boldsymbol{v}}^{t-1}) - g_{t-1}(\boldsymbol{v}^{t-1})\|_2 \tag{200}$$

$$\le C_0(\delta)L\|\boldsymbol{u}^t - \hat{\boldsymbol{u}}^t\|_2 + |\mathsf{b}_t - \hat{\mathsf{b}}_t|\, \|g_{t-1}(\boldsymbol{v}^{t-1})\|_2 + L|\hat{\mathsf{b}}_t|\, \|\hat{\boldsymbol{v}}^{t-1} - \boldsymbol{v}^{t-1}\|_2, \tag{201}$$

where L is the maximum Lipschitz constant of e_t and g_{t-1} and the second inequality holds with high probability by the Bai–Yin law [8]. Next notice that, with high probability, $\|g_{t-1}(\boldsymbol{v}^{t-1})\|_2/\sqrt{n} \le C$ for some constant C by Theorem 1 (together with Assumption (B6)) and that $|\hat{\mathsf{b}}_t| \le |\mathsf{b}_t| + |\hat{\mathsf{b}}_t - \mathsf{b}_t| \le L + 1$ with high probability by Assumption (39) and the Lipschitz continuity of e_t. Hence, for a suitable constant C_1, the following holds with high probability

$$\frac{1}{\sqrt{n}}\|\boldsymbol{v}^t - \hat{\boldsymbol{v}}^t\|_2 \le C_1 \left\{ \frac{1}{\sqrt{n}}\|\boldsymbol{u}^t - \hat{\boldsymbol{u}}^t\|_2 + |\mathsf{b}_t - \hat{\mathsf{b}}_t| + \frac{1}{\sqrt{n}}\|\hat{\boldsymbol{v}}^{t-1} - \boldsymbol{v}^{t-1}\|_2 \right\}. \tag{202}$$

We therefore have $\|\boldsymbol{v}^t - \hat{\boldsymbol{v}}^t\|_2/\sqrt{n} \overset{\text{P}}{\simeq} 0$ by Equation (39) and the induction hypothesis.

Taking the difference of Equations (2) and (38), we get

$$\|\boldsymbol{u}^{t+1} - \hat{\boldsymbol{u}}^{t+1}\|_2 \le \|\boldsymbol{A}\|_{\text{op}}\|g_t(\boldsymbol{v}^t) - g(\hat{\boldsymbol{v}}^t)\|_2 + |\mathsf{d}_t - \hat{\mathsf{d}}_t|\|e_t(\boldsymbol{u}^t)\|_2 + |\hat{\mathsf{d}}_t|\|e_t(\boldsymbol{u}^t) - e_t(\hat{\boldsymbol{u}}^t)\|_2 \tag{203}$$

$$\le C_0(\delta)L\|\boldsymbol{v}^t - \hat{\boldsymbol{v}}^t\|_2 + |\mathsf{d}_t - \hat{\mathsf{d}}_t|\|e_t(\boldsymbol{u}^t)\|_2 + L|\hat{\mathsf{d}}_t|\|\boldsymbol{u}^t - \hat{\boldsymbol{u}}^t\|_2, \tag{204}$$

and the proof is completed by the same argument as above. □

7. Application to general compressed sensing

In this section we discuss how the general theory of Section 3 applies to the problem of reconstructing an unknown signal $\boldsymbol{\theta}_0 \in \mathbb{R}^n$ from noisy linear measurements given by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta}_0 + \boldsymbol{w}. \tag{205}$$

Here, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the (known) sensing matrix, $\boldsymbol{y} \in \mathbb{R}^m$ is the measurement vector and $\boldsymbol{w}$ is a noise vector, independent of $\boldsymbol{A}$. We know $\boldsymbol{y}$ and $\boldsymbol{A}$, and we are required to reconstruct $\boldsymbol{\theta}_0$. As before, it is understood that we are really given a sequence of problems indexed by the dimensions n, with $m(n)/n \to \delta$.

If $m < n$, the problem becomes underdetermined. Reconstruction of $\boldsymbol{\theta}_0$ can be possible if we have some prior information. The prior knowledge can be encoded in a suitably chosen sequence of denoising function $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$, $t \in \mathbb{N}$ [16]. Given such a denoising function, we consider the following AMP algorithm:

$$\hat{\boldsymbol{\theta}}^{t+1} = \eta_t\left(\hat{\boldsymbol{\theta}}^t + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t\right), \tag{206}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\theta}}^t + \hat{\mathsf{b}}_t\boldsymbol{r}^{t-1}, \tag{207}$$

where the initialization is given by $\hat{\boldsymbol{\theta}}^0 = 0$ and $\eta_{-1}(\cdot) = 0$. We assume the Onsager coefficient $\hat{\mathsf{b}}_t$ to be a function of $\hat{\boldsymbol{\theta}}^0, \dots, \hat{\boldsymbol{\theta}}^t$ and $\boldsymbol{r}^0, \dots, \boldsymbol{r}^{t-1}$, but we will discuss concrete choices below.

7.1 *General theory*

We make the following assumptions (see Remarks 7.2 and 7.3 for a discussion):

- (**C1**) The sensing matrix $\boldsymbol{A}$ is Gaussian with i.i.d. entries, $(A_{ij})_{i\le m, j\le n} \sim \mathsf{N}(0, 1/m)$.
- (**C2**) For each t, the sequence (in n) of denoisers $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is uniformly Lipschitz.
- (**C3**) $\|\boldsymbol{\theta}_0\|_2/\sqrt{n}$ converges to a constant as $n \to \infty$.
- (**C4**) The limit $\sigma_w = \lim_{n\to\infty} \|\boldsymbol{w}\|_2/\sqrt{m} \in [0, \infty)$ exists.
- (**C5**) For any $t \in \mathbb{N}$ and any $\sigma \ge 0$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\langle \boldsymbol{\theta}_0, \eta_t(\boldsymbol{\theta}_0 + \boldsymbol{Z})\rangle\right], \tag{208}$$

  where $Z \sim \mathsf{N}(0, \sigma^2 \boldsymbol{I}_n)$.
- (**C6**) For any $s, t \in \mathbb{N}$ and any 2×2 covariance matrix $\boldsymbol{\Sigma}$, the following limit exists and is finite:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left\langle \eta_s(\boldsymbol{\theta}_0 + \boldsymbol{Z}), \eta_t\left(\boldsymbol{\theta}_0 + \boldsymbol{Z}'\right)\right\rangle\right], \tag{209}$$

  where $\left(\boldsymbol{Z}, \boldsymbol{Z}'\right) \sim \mathsf{N}(0, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$.

The technical Assumptions (C5) and (C6) ensure the existence of the limits in the following *state evolution* recursion:

$$\tau_0^2 = \sigma_w^2 + \lim_{n\to\infty} \frac{1}{\delta n}\|\boldsymbol{\theta}_0\|_2^2, \tag{210}$$

$$\tau_{t+1}^2 = \sigma_w^2 + \lim_{n\to\infty} \frac{1}{\delta n}\mathbb{E}\left[\left\|\eta_t(\boldsymbol{\theta}_0 + \tau_t Z) - \boldsymbol{\theta}_0\right\|_2^2\right], \tag{211}$$

where $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_n)$.

State evolution predicts the asymptotic behavior of the estimates $\hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \dots$ in terms of an iterative denoising process.

THEOREM 14 Under Assumptions (C1)–(C6), consider the recursion (206)–(207). Assume that $\hat{\mathsf{b}}_t(\hat{\boldsymbol{\theta}}^0, \boldsymbol{r}^0, \dots, \boldsymbol{r}^{t-1}, \hat{\boldsymbol{\theta}}^t)$ satisfies

$$\hat{\mathsf{b}}_t \overset{\mathrm{P}}{\simeq} \mathsf{b}_t \equiv \frac{1}{m}\mathbb{E}\left[\operatorname{div} \eta_{t-1}(\boldsymbol{\theta}_0 + \tau_{t-1}\boldsymbol{Z})\right], \quad \boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_n). \tag{212}$$

Further assume that the state evolution sequence satisfies $\tau_s > \sigma_w$ for all $s \le t$. Then, for any sequences $\phi_n : (\mathbb{R}^m)^2 \to \mathbb{R}$, $n \ge 1$ and $\psi_n : (\mathbb{R}^n)^2 \to \mathbb{R}$, $n \ge 1$, of uniformly pseudo-Lipschitz functions of order k

$$\phi_n(\boldsymbol{r}^t, \boldsymbol{w}) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{w} + \sqrt{\tau_t^2 - \sigma_w^2}\,\boldsymbol{Z}, \boldsymbol{w}\right)\right], \tag{213}$$

$$\psi_n\left(\hat{\boldsymbol{\theta}}^t + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t, \boldsymbol{\theta}_0\right) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\psi_n\left(\boldsymbol{\theta}_0 + \tau_t \boldsymbol{Z}', \boldsymbol{\theta}_0\right)\right], \tag{214}$$

where $\boldsymbol{Z} \sim \mathsf{N}\left(0, \boldsymbol{I}_m\right)$ and $\boldsymbol{Z}' \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$.

Proof. This is a special case of the asymmetric AMP of Equations (1) and (2), with

$$\boldsymbol{u}^{t+1} = \boldsymbol{\theta}_0 - \left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t + \hat{\boldsymbol{\theta}}^t\right), \tag{215}$$

$$\boldsymbol{v}^t = \boldsymbol{w} - \boldsymbol{r}^t, \tag{216}$$

$$e_t(\boldsymbol{u}) = \eta_{t-1}\left(\boldsymbol{\theta}_0 - \boldsymbol{u}\right) - \boldsymbol{\theta}_0, \tag{217}$$

$$g_t(\boldsymbol{v}) = \boldsymbol{v} - \boldsymbol{w}, \tag{218}$$

and the initialization $\boldsymbol{u}^0 = -\boldsymbol{\theta}_0$. Assumptions (B1)–(B6) are satisfied thanks to Assumptions (C1)–(C6). The claim follows from Theorem 1 and Corollary 2. □

REMARK 7.1 A special case of common interest is $\psi_n(\boldsymbol{x}, \boldsymbol{y}) = \|\eta_t(\boldsymbol{x}) - \boldsymbol{y}\|_2^2/n$, for which Theorem 14 yields

$$\frac{1}{n}\left\|\hat{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}_0\right\|_2^2 \overset{\mathrm{P}}{\simeq} \frac{1}{n}\mathbb{E}\left[\left\|\eta_t\left(\boldsymbol{\theta}_0 + \tau_t \boldsymbol{Z}'\right) - \boldsymbol{\theta}_0\right\|_2^2\right] \tag{219}$$

$$= \delta\left(\tau_{t+1}^2 - \sigma_w^2\right). \tag{220}$$

REMARK 7.2 Earlier work [5] assumes that η_t is separable and Lipschitz. The Lipschitz Assumption (C2) is much weaker: it trivially holds for the setting of [5], but comprises many interesting new cases. Among others, (C2) is satisfied by the non-separable denoisers introduced in Sections 1.1 and 1.2:

1. For any $\lambda \ge 0$, the singular value soft thresholding operator $\mathsf{S}(, \cdot\,; \lambda)$ is non-expansive (i.e. 1-Lipschitz). This follows immediately from the fact that it is a proximal operator,

$$\mathsf{S}(\boldsymbol{Y}; \lambda) = \underset{\boldsymbol{Y}'}{\mathrm{argmin}}\left\{\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Y}'\|_F^2 + \lambda\|\boldsymbol{Y}'\|_*\right\}, \tag{221}$$

   where $\|\boldsymbol{Y}'\|_1$ denotes the nuclear norm of $\boldsymbol{Y}'$ (the sum of the magnitudes of its singular values). Proximal operators are non-expansive.

2. The NLM denoiser η_t is Lipschitz, uniformly in n if the range R and the patch-size L remain bounded as n increases. This result is proved in Section B.

REMARK 7.3 Assumptions (C5) and (C6) are required to describe the joint behavior of $\boldsymbol{\theta}_0$ and η_t as $n \to \infty$. Let us discuss how they can be checked in specific settings.

First consider the setting of [5], namely when η_t is separable. Following [5] we assume that the empirical distribution $\hat{p}_{\theta_0}$ of the coordinates of $\boldsymbol{\theta}_0$ converges weakly to a probability distribution p as $n \to \infty$, and that $\mathbb{E}_{\hat{p}_{\theta_0}}[\Theta_0^2] \xrightarrow[n\to\infty]{} \mathbb{E}_p[\Theta_0^2]$. Since $\psi(\theta) = \theta\mathbb{E}_{Z\sim\mathsf{N}(0,\sigma^2)}[\eta_t(\theta+Z)]$ is pseudo-Lipschitz of order 2, the law of large numbers [5, Lemma 4] implies

$$\frac{1}{n}\mathbb{E}\left[\langle\boldsymbol{\theta}_0, \eta_t(\boldsymbol{\theta}_0+\boldsymbol{Z})\rangle\right] = \frac{1}{n}\sum_{i=1}^{n}\theta_{0,i}\mathbb{E}_{Z_i}[\eta_t(\theta_{0,i}+Z_i)] = \frac{1}{n}\sum_{i=1}^{n}\psi(\theta_{0,i}) \xrightarrow[n\to\infty]{} \mathbb{E}_{\Theta\sim p}[\psi(\Theta)], \tag{222}$$

and (C6) can be proven similarly.

Next consider the NLM denoiser. For any fixed range R, the denoised pixel at (i,j), $\eta(\boldsymbol{z})_{i,j}$ is only a function of the input $\boldsymbol{z}$ within a box of side $2(L+R)+1$ around (i,j). Assumptions (C5) and (C6) can be proved analogously to the above, provided the underlying image $\boldsymbol{\theta}_0$ is generated according to a random field with suitable mixing properties. Examples of mixing conditions that ensure the existence of the limits in Assumptions (C5) and (C6) are provided—for instance—in [22,35].

For the singular value thresholding denoiser $\mathsf{S}(\,\cdot\,;\lambda)$, we expect Assumptions (C5) and (C6) to hold if $\boldsymbol{\theta}_0(n) \in \mathbb{R}^{n_1\times n_2}$, $n_a = n_a(n)$, is a sequence of matrices indexed by $n = n_1 n_2$ with $n_1(n)/\sqrt{n} \to \rho \in (0,\infty)$, $n_2(n)/\sqrt{n} \to \rho^{-1} \in (0,\infty)$, and such that the empirical spectral distribution of $\boldsymbol{\theta}_0/\sqrt{n_1(n)}$ converges as $n \to \infty$, as suggested for instance by [15]. We leave an analysis of this example to future work.

Finally, it is often possible to avoid checking the existence of these limits by considering suitable subsequential limit. This strategy is illustrated in the next section.

REMARK 7.4 Two choices of the coefficient $\hat{\mathsf{b}}_t$ that satisfy the Assumption (212) are:

- The empirical mean

$$\hat{\mathsf{b}}_t = \frac{1}{m}\mathrm{div}\,\eta_{t-1}\left(\hat{\boldsymbol{\theta}}^{t-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^{t-1}\right). \tag{223}$$

  Using Theorem 14, this satisfies the assumptions by induction, provided $\boldsymbol{x} \mapsto \frac{1}{m}\mathrm{div}\eta_t(\boldsymbol{x})$ is uniformly Lipschitz for each t.

- If $\boldsymbol{x} \mapsto \frac{1}{m}\mathrm{div}\,\eta_t(\boldsymbol{x})$ is not uniformly Lipschitz, a smoothed version of Equation (223) achieves the same goal, namely,

$$\hat{\mathsf{b}}_t = \frac{1}{m}\mathbb{E}\left[\mathrm{div}\,\eta_{t-1}\left(\hat{\boldsymbol{\theta}}^{t-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^{t-1} + \varepsilon_n\boldsymbol{Z}\right)\right], \tag{224}$$

  where the expectation is with respect to $\boldsymbol{Z} \sim \mathsf{N}(0,\boldsymbol{I}_n)$, and ε_n is a deterministic sequence that converges to 0 sufficiently slowly. Adapting the arguments of Section 5.5.8, it is possible to show that this choice satisfies the Assumption (212).

We also note that, even if $\boldsymbol{x} \mapsto \frac{1}{m}\mathrm{div}\eta_t(\boldsymbol{x})$ is not uniformly Lipschitz, the choice (223) can still satisfy the Assumption (212). For instance, if $\eta_t(\,\cdot\,)$ is the soft thresholding denoiser (a case studied in [5,18]), then $\boldsymbol{x} \mapsto \frac{1}{m}\mathrm{div}\,\eta_t(\boldsymbol{x})$ is discontinuous, but nevertheless a standard weak convergence argument implies Equation (212).

7.2 *Denoising by convex projection*

An important feature of the theory developed in the previous section is that the denoiser η_t can be fairly general and not induced by an underlying optimization problem. Nevertheless, it is interesting to specialize the theory developed so far to cases with special additional structure.

One possible approach toward reconstruction from noisy measurements, cf. Equation (205), assumes that $\boldsymbol{\theta}_0$ belongs to a closed convex body $\mathcal{K} \subseteq \mathbb{R}^n$. The reconstruction method of choice solves the constrained least squares problem

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2, \tag{225}$$

$$\text{subject to} \quad \boldsymbol{\theta} \in \mathcal{K}. \tag{226}$$

Denoting by $\mathsf{P}_{\mathcal{K}}$ the projection onto the set $\mathcal{K}$ (which is a 1-Lipschitz denoiser), the corresponding AMP algorithm reads

$$\hat{\boldsymbol{\theta}}^{t+1} = \mathsf{P}_{\mathcal{K}}\left(\hat{\boldsymbol{\theta}}^{t} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t\right), \tag{227}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\theta}}^{t} + \hat{\mathsf{b}}_t\boldsymbol{r}^{t-1}, \tag{228}$$

where $\hat{\boldsymbol{\theta}}^{0} = 0$ and $\hat{\mathsf{b}}_t$ is an estimator of $\mathsf{b}_t = (1/m)\mathbb{E}\left[\mathrm{div}\mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \tau_t\boldsymbol{Z})\right]$. In many cases of interest, such estimator is simply given by $\hat{\mathsf{b}}_t = (1/m)\mathrm{div}\,\mathsf{P}_{\mathcal{K}}(\hat{\boldsymbol{\theta}}^{t} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{r}^t)$. It is possible to show that fixed points of this iteration are stationary points of the least square problem (225) and (226).

The constraint $\boldsymbol{\theta} \in \mathcal{K}$ is effective if $\mathcal{K}$ accurately captures the structure of the signal $\boldsymbol{\theta}_0$. We denote by $\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)$ the tangent cone of $\mathcal{K}$ at $\boldsymbol{\theta}_0$, i.e. the smallest convex cone containing $\mathcal{K} - \theta_0$. This can also be defined as

$$\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0) = \left\{\boldsymbol{v} \in \mathbb{R}^n : \quad \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} d(\boldsymbol{\theta}_0 + \varepsilon\,\boldsymbol{v}, \mathcal{K}) = 0\right\}, \tag{229}$$

with $d(\boldsymbol{x}, S) \equiv \inf\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 : \boldsymbol{y} \in S\}$ the Euclidean point-set distance. A highly structured signal $\boldsymbol{\theta}_0$ corresponds to a 'small' cone $\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)$. This can be quantified via its statistical dimension [**1**,**10**]

$$\Delta(\mathcal{C}) = \mathbb{E}\left\{\left\|\mathsf{P}_{\mathcal{C}}(\boldsymbol{Z})\right\|_2^2\right\}, \tag{230}$$

where expectation is with respect to $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_n)$. It turns out that the statistical dimension also controls the convergence of AMP. As for our general theory, we will consider a sequence of problems indexed by the dimension n.

THEOREM 15 Consider the AMP iteration (227) and (228), for a sequence of problems $(\boldsymbol{\theta}_0(n), \boldsymbol{A}(n), \mathcal{K}(n), \boldsymbol{w}(n))$ whereby $\boldsymbol{A} = \boldsymbol{A}(n) \in \mathbb{R}^{m\times n}$ is a matrix with i.i.d. Gaussian entries $(A_{ij})_{i\le m, j\le n} \sim_{iid}$

$\mathsf{N}(0, 1/m)$, $\mathcal{K} = \mathcal{K}(n) \subseteq \mathbb{R}^n$ is a closed convex set with $\limsup_{n\to\infty} \max_{\boldsymbol{x}\in\mathcal{K}(n)} \|\boldsymbol{x}\|_2/\sqrt{n} < \infty$, $\boldsymbol{\theta}_0 \in \mathcal{K}(n)$ and $\lim_{n\to\infty} \|\boldsymbol{w}(n)\|_2/\sqrt{m} = \sigma_w$. Assume $m/n \to \delta \in (0, \infty)$ and

$$\limsup_{n\to\infty} \frac{1}{m} \Delta\left(\mathcal{C}_{\mathcal{K}(n)}(\boldsymbol{\theta}_0(n))\right) \leq \rho \in [0, 1). \tag{231}$$

Then for any $t \geq 0$, letting $\mathsf{R}_0 \equiv \limsup_{n\to\infty} \|\boldsymbol{\theta}_0(n)\|_2/\sqrt{n}$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left\{\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0\|_2^2\right\} \leq \delta \mathsf{R}_0^2\, \rho^{t+1} + \delta\sigma_w^2 \frac{\rho - \rho^{t+1}}{1 - \rho}\,. \tag{232}$$

The proof of this statement is deferred to Appendix D.

This theorem establishes exponentially fast convergence (in the high-dimensional limit) in all the region $m \geq (1 + \eta)\Delta_n$, $\Delta_n = \Delta\left(\mathcal{C}_{\mathcal{K}(n)}(\boldsymbol{\theta}_0(n))\right)$, i.e. whenever exact reconstruction is possible in absence of noise [1]. Further, the convergence rate is precisely given by the ratio of the number of necessary measurements to the number of measurements Δ_n/m. For instance, it implies that, in order to achieve accuracy $\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0\|_2/\|\boldsymbol{\theta}_0\|_2 \leq \varepsilon$ in the noiseless case $\sigma_w = 0$, it is sufficient to run the AMP iteration (227) and (228) for approximately $\log(1/\varepsilon)/\log(m/\Delta_n)$ iterations.

The first result of this type (for separable soft thresholding denoising) was obtained in [18,19]. The only comparable result is obtained in recent work by Oymak *et al.* [33], which establishes exponential convergence of projected gradient descent, in a non-asymptotic sense, although at a slower rate.[5] In particular, in the noiseless case, ε accuracy requires $(n/m)\log(1/\varepsilon)$. It would be interesting to derive a non-asymptotic version of Theorem 15, which might be possible using the approach of [40].

Funding

This work was partially supported by the National Science Foundation (CCF-1319979, DMS-1613091 and CCF-1714305).

References

1. Amelunxen, D., Lotz, M., McCoy, M. B. & Tropp, J. A. (2014) Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, **3**, 224–294.
2. Buades, A., Coll, B. & Morel, J.-M. (2005) A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, **4**, 490–530.
3. Boucheron, S., Lugosi, G. & Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press.
4. Bayati, M., Lelarge, M. & Montanari, A. (2015) Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.*, **25**, 753–822.
5. Bayati, M. & Montanari, A. (2011) The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory*, **57**, 764–785.
6. Bayati, M. & Montanari, A. (2012) The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, **58**, 1997–2017.

[5] The same paper also proves convergence at a faster rate, but this requires $m > 2\Delta_n$, i.e. a number of measurements that is twice as large as the optimal one.

7. Bolthausen, E. (2014) An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.*, **325**, 333–366.
8. Bai, Z.-D. & Yin, Y.-Q. (1988) Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *Ann. Probab.*, **16**, 1729–1741.
9. Çakmak, B., Opper, M., Winther, O. & Fleury, B. H. (2017) Dynamical functional theory for compressed sensing. arXiv:1705.04284.
10. Chandrasekaran, V., Recht, B., Parrilo, P. A. & Willsky, A. S. (2012) The convex geometry of linear inverse problems. *Found. Comput. Math.*, **12**, 805–849.
11. Candes, E. J., Sing-Long, C. A. & Trzasko, J. D. (2013) Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.*, **61**, 4643–4657.
12. Deshpande, Y., Abbe, E. & Abbe, E. (2016) Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference*, **6**, 125–170.
13. Dabov, K., Foi, A., Katkovnik, V. & Egiazarian, K. (2007) Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, **16**, 2080–2095.
14. Donoho, D. L. & Gavish, M. (2014) Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.*, **42**, 2413–2440.
15. Donoho, D. L., Gavish, M. & Montanari, A. (2013a) The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 8405–8410.
16. Donoho, D. L., Johnstone, I. & Montanari, A. (2013b) Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory*, **59**, 3396–3433.
17. Donoho, D. L. & Montanari, A. (2016) High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields*, **166**, 935–969.
18. Donoho, D. L., Maleki, A. & Montanari, A. (2009) Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 18914–18919.
19. Donoho, D. L., Maleki, A. & Montanari, A. (2011) The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inform. Theory*, **57**, 6920–6941.
20. Givens, C. R. & Shortt, R. M. (1984) A class of Wasserstein metrics for probability distributions. *Michigan Math. J.*, **31**, 231–240.
21. Javanmard, A. & Montanari, A. (2013) State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference*, **2**, 115.
22. Jenish, N. & Prucha, I. R. (2009) Central limit theorems and uniform laws of large numbers for arrays of random fields. *J. Econometrics*, **150**, 86–98.
23. Koller, D. & Friedman, N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, Massachusetts: MIT Press.
24. Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A. & Zdeborová, L. (2016) Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Trans. Inform. Theory*, **62**, 4228–4265.
25. Kamilov, U. S., Rangan, S., Unser, M. & Fletcher, A. K. (2012) Approximate message passing with consistent parameter estimation and applications to sparse learning. *Adv. Neural Inform. Process. Syst.*, **2**, 2438–2446.
26. Metzler, C. A., Maleki, A. & Baraniuk, R. G. (2016) From denoising to compressed sensing. *IEEE Trans. Inform. Theory*, **62**, 5117–5144.
27. Montanari, A. & Nguyen, P.-M. (2017) In preparation.
28. Montanari, A. (2012) Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*. pp. 394–438. Cambridge: Cambridge University Press.
29. Ma, J. & Ping, L. (2017) Orthogonal AMP. *IEEE Access*, **5**, 2020–2033.
30. Ma, Y., Rush, C. & Baron, D. (2017) Analysis of approximate message passing with a class of non-separable denoisers. arXiv:1705.03126.

31. Ma, Y., Zhu, J. & Baron, D. (2016) Approximate message passing algorithm with universal denoising and gaussian mixture learning. *IEEE Trans. Signal Process.*, **64**, 5611–5622.
32. Opper, M., Cakmak, B. & Winther, O. (2016) A theory of solving tap equations for Ising models with general invariant random matrices. *J. Phys. A*, **49**, 114002.
33. Oymak, S., Recht, B. & Soltanolkotabi, M. (2015) Sharp time–data tradeoffs for linear inverse problems. arXiv:1507.04793.
34. Oymak, S., Thrampoulidis, C. & Hassibi, B. (2013) The squared-error of generalized Lasso: a precise analysis. *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, *IEEE*, pp. 1002–1009. Monticello, IL, USA: IEEE.
35. Peligrad, M. & Gut, A. (1999) Almost-sure results for a class of dependent random variables. *J. Theoret. Probab.*, **12**, 87–104.
36. Ramani, S., Blu, T. & Unser, M. (2008) Monte-carlo sure: a black-box optimization of regularization parameters for general denoising algorithms. *IEEE Trans. Image Process.*, **17**, 1540–1554.
37. Recht, B., Fazel, M. & Parrilo, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.
38. Romanov, E. & Gavish, M. (2017) Near-optimal matrix recovery from random linear measurements. arXiv:1705.09958.
39. Rangan, S., Schniter, P. & Fletcher, A. (2016) Vector approximate message passing. arXiv:1610.03082.
40. Rush, C. & Venkataramanan, R. (2016) Finite-sample analysis of approximate message passing. *IEEE International Symposium on Information Theory (ISIT), IEEE*, pp. 755–759. Barcelona, Spain: IEEE.
41. Schniter, P. (2010) Turbo reconstruction of structured sparse signals. *44th Annual Conference on Information Sciences and Systems (CISS)*. Princeton, NJ, USA: IEEE, pp. 1–6.
42. Schniter, P. & Rangan, S. (2015) Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.*, **63**, 1043–1055.
43. Som, S. & Schniter, P. (2012) Compressive imaging using approximate message passing and a Markov-tree prior. *IEEE Trans. Signal Process.*, **60**, 3439–3448.
44. Stein, C. (1972) *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.
45. Thouless, D. J., Anderson, P. W. & Palmer, R. G. (1977) Solution of 'solvable model of a spin glass'. *Philos. Mag.*, **35**, 593–601.
46. Tramel, E. W., Drémeau, A. & Krzakala, F. (2016) Approximate message passing with restricted Boltzmann machine priors. *J. Stat. Mech. Theory Exp.*, **2016**, 073401.

A. Technical aspects of the numerical simulations

A.1 *Matrix compressed sensing*

Here we state the formula for computing the divergence of the singular value soft thresholding operator. Recall that for a matrix $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$, with singular value decomposition

$$\boldsymbol{Y} = \sum_{i=1}^{n_1 \wedge n_2} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}, \tag{A.1}$$

the SVT operator with threshold λ yields

$$\mathsf{S}(\boldsymbol{Y}; \lambda) = \sum_{i=1}^{n_1 \wedge n_2} (\sigma_i - \lambda)_+ \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}. \tag{A.2}$$

As proved in [11], the divergence for this operator can be computed using the formula

$$\operatorname{div} \mathsf{S}(\boldsymbol{Y};\lambda) = \sum_{i=1}^{n_1 \wedge n_2} \left[\mathbf{1}_{\{\sigma_i > \lambda\}} + |m-n| \left(1 - \frac{\lambda}{\sigma_i}\right)_+ \right] + 2 \sum_{i \neq j, i, j=1}^{n_1 \wedge n_2} \frac{\sigma_i(\sigma_i - \lambda)_+}{\sigma_i^2 - \sigma_j^2}. \tag{A.3}$$

This expression should be understood in a weak sense as it is not defined on the negligible set where $\boldsymbol{Y}$ has repeated singular values.

A.2 *Compressed sensing with images*

In our simulation, to compute the state evolution iterates

$$\tau_0^2 = \sigma_w^2 + \lim_{n\to\infty} \frac{1}{\delta n} \|\boldsymbol{\theta}_0\|_2^2, \tag{A.4}$$

$$\tau_{t+1}^2 = \sigma_w^2 + \lim_{n\to\infty} \frac{1}{\delta n} \mathbb{E}\left[\left\| \eta_t\left(\boldsymbol{\theta}_0 + \tau_t \boldsymbol{Z}\right) - \boldsymbol{\theta}_0 \right\|_2^2 \right], \tag{A.5}$$

we approximated them by their non-asymptotic estimates:

$$\hat{\tau}_0^2 = \sigma_w^2 + \frac{1}{\delta n} \|\boldsymbol{\theta}_0\|_2^2, \tag{A.6}$$

$$\hat{\tau}_{t+1}^2 = \sigma_w^2 + \frac{1}{\delta n} \mathbb{E}\left[\left\| \eta_t\left(\boldsymbol{\theta}_0 + \hat{\tau}_t \boldsymbol{Z}\right) - \boldsymbol{\theta}_0 \right\|_2^2 \right]. \tag{A.7}$$

Here $n = 170 \times 170$ is the size of our image. However, we could not compute the expectation in equation (A.7) exactly. Thus, at each iteration we used a Monte Carlo method to approximate the expectation with the mean over 10 samples. Computing each sample amounts to adding gaussian noise of variance $\hat{\tau}_t^2$ over the Lena image, denoising with NLM, and computing the square error. The resulting state evolution is shown in Fig. 3.

B. Proof of the Lipschitz property for NLM

Proposition B.1 Let η be the NLM denoiser defined in (18) and (19). There exists a constant $C = C(L, R)$ that depends on the patch size L and the range R, but not on the dimension of the image (n_1, n_2) nor the precision parameter h, such that η is Lipschitz with constant C.

This section is dedicated to the proof of this Proposition. The letters $i = (i_1, i_2)$, $j = (j_1, j_2)$, $k = (k_1, k_2) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ will be used to denote positions in the image. To show that η is Lipschitz, we will bound the operator norm of the derivative $\boldsymbol{D}\eta(z) = \left(\partial \eta(z)_i / \partial z_k\right)_{i,k}$ at all $\boldsymbol{z}$. This will result from three lemmas.

Lemma B.2 There exists a number $N = N(L, R)$, that depends only on L and R, such that $\boldsymbol{D}\eta$ has at most $N(L, R)$ non-zero entries in each row and in each column.

Lemma B.3 There exists a constant $D = D(L, R)$, that depends only on L, R such that all entries of $\boldsymbol{D}\eta$ are bounded by D.

Lemma B.4 Let $\boldsymbol{A}$ be a matrix with at most N non-zero entries in each row and in each column and with each entry bounded by D. Then $\|\boldsymbol{A}\|_{\rm op} \leq ND$.

Lemmas B.2 and B.3 ensure that $\boldsymbol{D}\eta$ satisfies the assumptions of Lemma B.4, which gives $\|\boldsymbol{D}\eta\|_{\mathrm{op}} \leq N(L,R)D(L,R)$ and hence Proposition B.1 by denoting $C(L,R) = N(L,R)D(L,R)$.

Proof of Lemma B.5 Given formulas (18) and (19), $\partial\eta(z)_i/\partial z_k$ can be non-zero only if k is in the range of i, that is $\|k-i\| \leq R$, or if k is in the patch of size $L \times L$, centered in some k' in the range of i. When i is fixed, the number of indexes k satisfying one of these two conditions can be upper bound by a quantity depending only on R and L, and conversely when we invert i and k. Thus, we get the result. □

Proof of Lemma B.6 We first introduce some convenient notations. We denote $d_{ik} = \|P_k(z) - P_i(z)\|_2^2/(L^2h^2)$ the dissimilarity between patches i and k and $\varphi(d) = \exp(-d)$. Then (18) and (19) can be written as

$$\eta(z)_i = \frac{\sum_{l:\|l-i\|\leq R} \varphi(d_{il}) z_l}{\sum_{l:\|l-i\|\leq R} \varphi(d_{il})}. \tag{B.1}$$

We now compute the derivative: if $k \neq i$,

$$\frac{\partial \eta(z)_i}{\partial z_k} = \frac{\partial\left(\eta(z)_i - z_i\right)}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\frac{\sum_{l:\|l-i\|\leq R} \varphi(d_{il})(z_l - z_i)}{\sum_{l:\|l-i\|\leq R}\varphi(d_{il})}\right) \tag{B.2}$$

$$= \mathbf{1}_{\{\|k-i\|\leq R\}} \frac{\varphi(d_{ik})}{\sum_{l:\|l-i\|\leq R}\varphi(d_{il})} + \frac{\sum_{l:\|l-i\|\leq R}\varphi'(d_{il})\frac{\partial d_{il}}{\partial z_k}(z_l - z_i)}{\sum_{l:\|l-i\|\leq R}\varphi(d_{il})} \tag{B.3}$$

$$- \frac{\left(\sum_{l:\|l-i\|\leq R}\varphi(d_{il})(z_l - z_i)\right)\left(\sum_{l:\|l-i\|\leq R}\varphi'(d_{il})\frac{\partial d_{il}}{\partial z_k}\right)}{\left(\sum_{l:\|l-i\|\leq R}\varphi(d_{il})\right)^2}. \tag{B.4}$$

Here, we note that $\varphi'(\mathrm{d}) = -\varphi(d)$, that $\sum_{l:\|l-i\|\leq R}\varphi(d_{il}) \geq \varphi(d_{ii}) = \varphi(0) = 1$ and that

$$\frac{\partial d_{il}}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\frac{1}{L^2h^2}\sum_{\|j-l\|\leq R}(z_j - z_{i+(j-l)})^2\right) \tag{B.5}$$

$$= \frac{1}{L^2h^2}\left(2\mathbf{1}_{\{\|k-i\|\leq R\}}(z_{l+(k-i)} - z_k) - 2\mathbf{1}_{\{\|k-l\|\leq R\}}(z_k - z_{i+(k-l)})\right), \tag{B.6}$$

$$\left|\frac{\partial d_{il}}{\partial z_k}\right| \leq \frac{4}{L^2h^2}\max(|z_{l+(k-i)} - z_k|, |z_k - z_{i+(k-l)}|) \tag{B.7}$$

$$\leq \frac{4}{Lh}\sqrt{d_{il}}. \tag{B.8}$$

It follows that

$$\left|\frac{\partial\eta(z)_i}{\partial z_k}\right| \leq 1 + \sum_{l:\|l-i\|\leq R}\varphi(d_{il})\frac{4}{Lh}\sqrt{d_{il}}|z_l - z_i| \tag{B.9}$$

$$+ \left(\sum_{l:\|l-i\|\leq R}\varphi(d_{il})|z_l - z_i|\right)\left(\sum_{l:\|l-i\|\leq R}\varphi(d_{il})\frac{4}{Lh}\sqrt{d_{il}}\right), \tag{B.10}$$

and as $|z_l - z_i| \le \|P_l(z) - P_i(z)\|_2 = Lh\sqrt{d_{il}}$,

$$\left|\frac{\partial \eta(z)_i}{\partial z_k}\right| \le 1 + 4 \sum_{l:\|l-i\|\le R} \varphi(d_{il}) d_{il} + 4 \left(\sum_{l:\|l-i\|\le R} \varphi(d_{il})\sqrt{d_{il}} \right)^2 . \tag{B.11}$$

The functions $d \mapsto \varphi(d)d$ and $d \mapsto \varphi(d)\sqrt{d}$ being bounded by a universal constant C_2, we get

$$\left|\frac{\partial \eta(z)_i}{\partial z_k}\right| \le 1 + 4N(L,R)C_2 + 4\left(N(L,R)C_2\right)^2 ,$$

which proves the claim when $k \ne i$. The case $k = i$ is similar. □

Proof of Lemma B.7 The squared operator norm of $\boldsymbol{A}$ is the spectral radius of $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}$. Let λ be an eigenvalue of $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}$ and $\boldsymbol{x} \ne 0$ a corresponding eigenvector: $\lambda \boldsymbol{x} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}$. Denote j the index that maximizes the magnitude $|x_j|$. Then

$$|\lambda x_j| = \left| \sum_k \sum_l (A^{\mathsf{T}})_{jk} A_{kl} x_l \right| \le \sum_k |A_{kj}| \sum_l |A_{kl}||x_l| \le \left(\sum_k |A_{kj}| \sum_l |A_{kl}| \right) |x_j| .$$

As $\boldsymbol{x} \ne 0$, we must have $|x_j| > 0$ so we get

$$|\lambda| \le \sum_k |A_{kj}| \sum_l |A_{kl}| \le \sum_k |A_{kj}| ND \le (ND)^2,$$

using the sparsity and the boundedness of the matrix entries. Thus, the spectral radius of $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}$ is at most $(ND)^2$, which proves the Lemma. □

C. Some useful tools

We reminder the readers of three well-known results. The first concerns with the operator norm of $\boldsymbol{A} \in \mathrm{GOE}\,(n)$; see, e.g. [8] for a more general statement. The second is a simple consequence of Stein's lemma [44]. The last one is the Gaussian Poincaré inequality.

THEOREM C.1 Consider a sequence of matrices $\boldsymbol{A} \sim \mathrm{GOE}\,(n)$. Then $\|\boldsymbol{A}\|_{\mathrm{op}} \to 2$ almost surely as $n \to \infty$.

LEMMA C.2 (Stein's lemma [44]). For any 2×2 covariance matrix K and $\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right) \sim \mathsf{N}\left(0, \mathrm{K} \otimes \boldsymbol{I}_n\right)$ and any $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ such that $\frac{\partial \varphi_i}{\partial z_i}$ exists almost everywhere for $1 \le i \le n$, if $\mathbb{E}\left[\langle \boldsymbol{Z}_1, \varphi\left(\boldsymbol{Z}_2\right)\rangle\right]$ and $\mathbb{E}\left[\mathrm{div}\varphi\left(\boldsymbol{Z}_2\right)\right]$ exist, then

$$\mathbb{E}\left[\langle \boldsymbol{Z}_1, \varphi\left(\boldsymbol{Z}_2\right)\rangle\right] = \mathrm{K}_{1,2}\mathbb{E}\left[\mathrm{div}\varphi\left(\boldsymbol{Z}_2\right)\right] = \mathbb{E}\left[\frac{1}{n}\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2\rangle\right]\mathbb{E}\left[\mathrm{div}\varphi\left(\boldsymbol{Z}_2\right)\right]. \tag{C.1}$$

THEOREM C.3 (Gaussian Poincaré inequality [3]). Let $z \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$ and $\varphi : \mathbb{R}^n \to \mathbb{R}$ continuous, weakly differentiable. Then for some universal constant c,

$$\mathrm{Var}\left[\varphi\left(z\right)\right] \le c\,\mathbb{E}\left[\|\nabla\varphi\left(z\right)\|_2^2\right]. \tag{C.2}$$

We state some properties of the GOE matrices and provide proofs for completeness.

LEMMA C.4 Consider a sequence of matrices $\boldsymbol{A} \sim \mathrm{GOE}(n)$ and two sequences (in n) of (non-random vectors), $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, such that $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \sqrt{n}$.

(a) $\frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{u}\rangle \xrightarrow{\mathrm{P}} 0$.

(c) Let $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ be a sequence of projection matrices such that there exists a constant t that satisfies for all n, $\mathrm{rank}(\boldsymbol{P}) \le t$. Then $\frac{1}{n}\|\boldsymbol{P}\boldsymbol{A}\boldsymbol{u}\|_2^2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$.

(c) $\frac{1}{n}\|\boldsymbol{A}\boldsymbol{u}\|_2^2 \xrightarrow{\mathrm{P}} 1$.

(d) There exists a sequence (in n) of random vectors $\boldsymbol{z} \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$, defined on the same probability space, such that $\frac{1}{n}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{z}\|_2^2 \xrightarrow[n\to\infty]{\mathrm{P}} 0$.

Proof.

(a) Recall that $\boldsymbol{A} = \boldsymbol{G} + \boldsymbol{G}^{\mathsf{T}}$ where $G_{i,j}$ are i.i.d. $\mathsf{N}(0, 1/(2n))$ random variables, thus

$$\frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{u}\rangle = \frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u}\rangle + \frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{G}^{\mathsf{T}}\boldsymbol{u}\rangle. \tag{C.4}$$

The random variable $\frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u}\rangle$ is centered Gaussian with variance

$$\frac{1}{n^2}\sum_{i,j=1}^{n} v_i^2 u_j^2 \frac{1}{2n} = \frac{\|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2}{2n^3} = \frac{1}{2n} \longrightarrow 0. \tag{C.5}$$

Thus, $\frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u}\rangle$ converges in probability to 0. We can conclude as similarly, $\frac{1}{n}\langle \boldsymbol{v}, \boldsymbol{G}^{\mathsf{T}}\boldsymbol{u}\rangle$ also converges in probability to 0.

(b) Consider $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ an orthogonal basis of the image of $\boldsymbol{P}$, such that $\|\boldsymbol{v}_1\| = \cdots = \|\boldsymbol{v}_k\| = \sqrt{n}$. Note that k can depend on n, but k is uniformly bounded by t. Then, by point (a),

$$\frac{1}{n}\|\boldsymbol{P}\boldsymbol{A}\boldsymbol{u}\|_2^2 = \frac{1}{n}\sum_{j=1}^{k}\left(\frac{\langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{v}_j\rangle}{\|\boldsymbol{v}_j\|}\right)^2 = \sum_{j=1}^{k}\left(\frac{1}{n}\langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{v}_j\rangle\right)^2 \xrightarrow[n\to\infty]{} 0 \tag{C.6}$$

using that $k \le t$ for all n.

(c) This follows immediately from point (d) below.

(d) It is easy to check that $\boldsymbol{A}\boldsymbol{u}$ is a centered Gaussian vector with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_n + \frac{1}{n}\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$. Thus, there exists a Gaussian vector $\boldsymbol{z} \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$ such that $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} = \boldsymbol{z} + (\sqrt{2}-1)\frac{1}{n}\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z}$. We then have:

$$\frac{\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{z}\|_2}{\sqrt{n}} = \frac{\left\|\left(\boldsymbol{\Sigma}^{1/2} - \boldsymbol{I}_n\right)\boldsymbol{z}\right\|_2}{\sqrt{n}} = \frac{1}{n^{3/2}}(\sqrt{2}-1)\|\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z}\|_2 = (\sqrt{2}-1)\frac{1}{n}|\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z}| \xrightarrow[n\to\infty]{\mathrm{P}} 0, \tag{C.8}$$

where the last convergence follows from the fact that $\frac{1}{n}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{z}$ is a centered Gaussian random variable with variance $\|\boldsymbol{u}\|_2^2/n^2 = 1/n$. □

We state some useful properties of uniformly pseudo-Lipschitz functions. We omit the proofs, which are easy to verify.

LEMMA C.5 Let k be any positive integer. Consider two sequences $f : \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$ and $g : \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$ of uniformly pseudo-Lipschitz functions of order k. The sequence of functions $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $n \geq 1$, such that $\varphi(\boldsymbol{x}, \boldsymbol{y}) = \langle f(\boldsymbol{x}), g(\boldsymbol{y})\rangle$ is uniformly pseudo-Lipschitz of order $2k$.

LEMMA C.6 Let t, s and k be any three positive integers. Consider a sequence (in n) of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s \in \mathbb{R}^n$ such that $\frac{1}{\sqrt{n}} \left\|\boldsymbol{x}_j\right\| \leq c_j$ for some constant c_j independent of n, for $j = 1, \ldots, s$, and a sequence (in n) of order-k uniformly pseudo-Lipschitz functions $\varphi_n : (\mathbb{R}^n)^{t+s} \to \mathbb{R}$. The sequence of functions $\phi_n(\cdot) = \varphi_n\left(\cdot, x_1, \ldots, x_s\right)$ is also uniformly pseudo-Lipschitz of order k.

LEMMA C.7 Let t be any positive integer. Consider a sequence (in n) uniformly pseudo-Lipschitz functions $\varphi_n : (\mathbb{R}^n)^t \to \mathbb{R}$ of order k. The sequence of functions $\phi_n : (\mathbb{R}^n)^t \to \mathbb{R}$ such that $\phi_n\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\right) = \mathbb{E}\left[\varphi_n\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t + \boldsymbol{Z}\right)\right]$, in which $\boldsymbol{Z} \sim \mathsf{N}\left(0, a\boldsymbol{I}_n\right)$ and $a \geq 0$, is also uniformly pseudo-Lipschitz of order k.

Finally, we have the following result on the Gaussian concentration for pseudo-Lipschitz functions.

LEMMA C.8 Let $\boldsymbol{Z} \sim \mathsf{N}\left(0, \boldsymbol{I}_n\right)$. Let k be any positive integer and $L > 0$ a constant. Let $\varphi_n : \mathbb{R}^n \to \mathbb{R}$ be a sequence (in n) of random functions, independent of $\boldsymbol{Z}$, such that $\mathbb{P}\{\mathcal{E}_n\} \to 1$ as $n \to \infty$, where $\mathcal{E}_n$ is the event that φ_n is pseudo-Lipschitz of order k with pseudo-Lipschitz constant L. Then $\varphi(\boldsymbol{Z}) \overset{\mathrm{P}}{\simeq} \mathbb{E}[\varphi(\boldsymbol{Z})]$.

Proof. This is a straightforward application of Theorem C.3. In particular, by the definition of pseudo-Lipschitz functions of order k, under $\mathcal{E}_n$,

$$\mathbb{E}_{\boldsymbol{Z}}\left[\|\nabla\varphi(\boldsymbol{Z})\|_2^2\right] \leq \frac{L^2}{n}\mathbb{E}_{\boldsymbol{Z}}\left[\left(1 + 2\left(\frac{1}{\sqrt{n}}\|\boldsymbol{Z}\|_2\right)^{k-1}\right)^2\right] \leq \frac{L^2}{n}C(k), \tag{C.9}$$

for a constant $C(k)$ that only depends on k. Then by Theorem C.3, for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left\{\left|\varphi_n(\boldsymbol{Z}) - \mathbb{E}_{\boldsymbol{Z}}\left[\varphi_n(\boldsymbol{Z})\right]\right| > \epsilon\right\} &\leq \mathbb{E}\left\{\mathbb{P}_{\boldsymbol{Z}}\left\{\left|\varphi_n(\boldsymbol{Z}) - \mathbb{E}_{\boldsymbol{Z}}\left[\varphi_n(\boldsymbol{Z})\right]\right| > \epsilon\right\}\mathbb{I}_{\mathcal{E}_n}\right\} + \mathbb{P}\{\neg\mathcal{E}_n\} \\ &\leq \frac{L^2 C(k)}{n\epsilon^2} + \mathbb{P}\{\neg\mathcal{E}_n\}.\end{aligned}$$

This completes the proof. □

D. Proof of Theorem 15

By assumption,

$$\mathsf{R}_* \equiv 2\limsup_{n\to\infty}\max_{\boldsymbol{x}\in\mathcal{K}(n)}\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_2 < \infty. \tag{D.1}$$

Note for all $n \geq n_0$, $\|\hat{\boldsymbol{\theta}}^t\|_2, \|\boldsymbol{\theta}_0\|_2 \leq \mathsf{R}_*\sqrt{n}$ for all t.

Next fix $t \geq 0$ and denote by B_t the right-hand side of Equation (232). Assume by contradiction that $\limsup_{n\to\infty}\mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t(n) - \boldsymbol{\theta}_0(n)\|_2^2\}/n = B_t + \varepsilon > B_t$. We can then find a subsequence $\{n_1(\ell)\}_{\ell\geq 1}$ along which $\lim_{\ell\to\infty}\mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t(n_1(\ell)) - \boldsymbol{\theta}_0(n_1(\ell))\|_2^2\}/n_1(\ell) = B_t + \varepsilon$. We will prove that this subsequence can be further refined to $\{n_2(\ell)\}_{\ell\geq 1} \subseteq \{n_1(\ell)\}_{\ell\geq 1}$ such that $\lim_{\ell\to\infty}\mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t(n_2(\ell)) - \boldsymbol{\theta}_0(n_2(\ell))\|_2^2\}/n_2(\ell) \leq B_t$, thus leading to a contradiction.

To simplify the notation we can assume, without loss of generality, that the first subsequence is not needed, i.e. $\limsup_{n\to\infty} \mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t(n) - \boldsymbol{\theta}_0(n)\|_2^2\}/n = B_t + \varepsilon > B_t$. We then claim that we can find a subsequence $\{n_2(\ell)\}_{\ell\ge 1}$ along which Assumptions (C3), (C5) and (C6) hold, with $\eta_s(\,\cdot\,)$, $\eta_t(\,\cdot\,) = \mathsf{P}_{\mathcal{K}}(\,\cdot\,)$. Consider Assumption (C6). Let the functions $F_n : S_+^2 \to \mathbb{R}$ (with S_+^2 the cone of 2×2 positive semidefinite matrices) be defined by

$$F_n(\boldsymbol{\Sigma}) \equiv \frac{1}{n}\mathbb{E}\left[\left\langle \mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \boldsymbol{Z}), \mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \boldsymbol{Z}')\right\rangle\right], \tag{D.2}$$

where expectation is with respect to $(\boldsymbol{Z}, \boldsymbol{Z}') \sim \mathsf{N}(0, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$.

Note that the function $(\boldsymbol{Z}, \boldsymbol{Z}') \mapsto \left\langle \mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \boldsymbol{Z}), \mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \boldsymbol{Z}')\right\rangle/n$ is uniformly pseudo-Lipschitz of order 2. Hence, using Lemma 10, we have

$$\sup_{n\ge 1} \left|F_n(\boldsymbol{\Sigma}_1) - F_n(\boldsymbol{\Sigma}_2)\right| \le \xi(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2), \tag{D.3}$$

for some function ξ such that $\lim_{\boldsymbol{\Sigma}_1\to\boldsymbol{\Sigma}_2} \xi(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = 0$. Further, $\sup_{n\ge 1} |F_n(\boldsymbol{\Sigma})| \le \mathsf{R}_*^2$. Hence, by the Arzelà–Ascoli theorem, F_n converges uniformly on any compact set $\{\boldsymbol{\Sigma} : \|\boldsymbol{\Sigma}\|_F \le C\}$, thus satisfying condition (C6), along a certain subsequence $\{n_2'(\ell)\}_{\ell\ge 1}$. Assumption (C5) is established by the same argument, eventually refining the subsequence to $\{n_2''(\ell)\}_{\ell\ge 1}$. Finally, by taking a further subsequence $\{n_2(\ell)\}_{\ell\ge 1}$, we can assume that $\|\boldsymbol{\theta}_0(n_2(\ell))\|_2^2/\sqrt{n} \to \mathsf{R}_0$.

We can therefore apply Theorem 14 (and Remark 7.1) along this subsequence, to obtain $\|\hat{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}_0\|_2^2/n \overset{\mathrm{P}}{\simeq} \delta(\tau_{t+1}^2 - \sigma_w^2)$ and hence (since $\|\hat{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}_0\|_2^2/n \le \mathsf{R}_*^2$ is bounded uniformly)

$$\lim_{\ell\to\infty} \frac{1}{n}\mathbb{E}\left\{\|\hat{\boldsymbol{\theta}}^{t+1}(n_2(\ell)) - \boldsymbol{\theta}_0(n_2(\ell))\|_2^2\right\} = \delta(\tau_{t+1}^2 - \sigma_w^2). \tag{D.4}$$

Here τ_{t+1} is given recursively by Equation (211), namely, $\tau_0^2 = \mathsf{R}_0^2$ and

$$\tau_{s+1}^2 = \sigma_w^2 + G(\tau_s^2), \tag{D.5}$$

$$G(\tau^2) = \lim_{\substack{\ell\to\infty\\ n=n_2(\ell)}} \frac{1}{n\delta}\mathbb{E}\left[\left\|\mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \tau\boldsymbol{Z}) - \boldsymbol{\theta}_0\right\|_2^2\right], \tag{D.6}$$

where the limit exists by the existence of the limit of $F_n(\boldsymbol{\Sigma})$ above. Now, since $\mathcal{K} - \boldsymbol{\theta}_0 \subseteq \mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)$, we have

$$\left\|\mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \tau\boldsymbol{Z}) - \boldsymbol{\theta}_0\right\|_2^2 = \left\|\mathsf{P}_{\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)}\left[\mathsf{P}_{\mathcal{K}}(\boldsymbol{\theta}_0 + \tau\boldsymbol{Z}) - \boldsymbol{\theta}_0\right]\right\|_2^2 \le \left\|\mathsf{P}_{\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)}(\tau\boldsymbol{Z})\right\|_2^2. \tag{D.7}$$

Therefore,

$$G(\tau^2) \le \limsup_{n\to\infty} \frac{1}{m}\mathbb{E}\left\{\left\|\mathsf{P}_{\mathcal{C}_{\mathcal{K}}(\boldsymbol{\theta}_0)}(\boldsymbol{Z})\right\|_2^2\right\} \tau^2 \le \rho\,\tau^2. \tag{D.8}$$

We therefore get the recursion $\tau_{s+1}^2 \le \sigma_w^2 + \rho\tau_s^2$, which can be summed to yield

$$\tau_t^2 = \mathsf{R}_0^2\rho^t + \sigma_w^2\frac{1-\rho^t}{1-\rho}\,. \tag{D.9}$$

Therefore, using Equation (D.4), we get

$$\lim_{\substack{\ell\to\infty \\ n=n_2(\ell)}} \frac{1}{n}\mathbb{E}\left\{\|\hat{\boldsymbol{\theta}}^{t+1}(n) - \boldsymbol{\theta}_0(n)\|_2^2\right\} \leq B_t, \tag{D.10}$$

which yields the desired contradiction hence proving the theorem.