Plug-in Estimation in High-Dimensional Linear Inverse Problems: A Rigorous Analysis

Alyson K. Fletcher
Dept. Statistics
UC Los Angeles
akfletcher@ucla.edu

Parthe Pandit
Dept. ECE
UC Los Angeles
parthepandit@ucla.edu

Sundeep Rangan
Dept. ECE
NYU
srangan@nyu.edu

Subrata Sarkar
Dept. ECE
The Ohio State Univ.
sarkar.51@osu.edu

Philip Schniter
Dept. ECE
The Ohio State Univ.
schniter.1@osu.edu

Abstract

Estimating a vector $\mathbf{x}$ from noisy linear measurements $\mathbf{A}\mathbf{x} + \mathbf{w}$ often requires use of prior knowledge or structural constraints on $\mathbf{x}$ for accurate reconstruction. Several recent works have considered combining linear least-squares estimation with a generic or "plug-in" denoiser function that can be designed in a modular manner based on the prior knowledge about $\mathbf{x}$. While these methods have shown excellent performance, it has been difficult to obtain rigorous performance guarantees. This work considers plug-in denoising combined with the recently-developed Vector Approximate Message Passing (VAMP) algorithm, which is itself derived via Expectation Propagation techniques. It shown that the mean squared error of this "plug-and-play" VAMP can be exactly predicted for high-dimensional right-rotationally invariant random $\mathbf{A}$ and Lipschitz denoisers. The method is demonstrated on applications in image recovery and parametric bilinear estimation.

1 Introduction

The estimation of an unknown vector $\mathbf{x}^0 \in \mathbb{R}^N$ from noisy linear measurements $\mathbf{y}$ of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w} \in \mathbb{R}^M, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a known transform and $\mathbf{w}$ is disturbance, arises in a wide-range of learning and inverse problems. In many high-dimensional situations, such as when the measurements are fewer than the unknown parameters (i.e., $M \ll N$), it is essential to incorporate known structure on $\mathbf{x}^0$ in the estimation process. A fundamental challenge is how to perform structured estimation of $\mathbf{x}^0$ while maintaining computational efficiency and a tractable analysis.

Approximate message passing (AMP), originally proposed in [1], refers to a powerful class of algorithms that can be applied to reconstruction of $\mathbf{x}^0$ from (1) that can easily incorporate a wide class of statistical priors. In this work, we restrict our attention to $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1}\mathbf{I})$, noting that AMP was extended to non-Gaussian measurements in [2, 3, 4]. AMP is computationally efficient, in that it generates a sequence of estimates $\{\widehat{\mathbf{x}}_k\}_{k=0}^{\infty}$ by iterating the steps

$$\widehat{\mathbf{x}}_k = \mathbf{g}(\mathbf{r}_k, \gamma_k) \tag{2a}$$

$$\mathbf{v}_k = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}_k + \tfrac{N}{M}\langle \nabla \mathbf{g}(\mathbf{r}_k, \gamma_k)\rangle \mathbf{v}_{k-1} \tag{2b}$$

$$\mathbf{r}_{k+1} = \widehat{\mathbf{x}}_k + \mathbf{A}^{\mathsf{T}}\mathbf{v}_k, \quad \gamma_{k+1} = M/\|\mathbf{v}_k\|^2, \tag{2c}$$

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

initialized with $\mathbf{r}_0 = \mathbf{A}^\mathsf{T}\mathbf{y}$, $\gamma_0 = M/\|\mathbf{y}\|^2$, $\mathbf{v}_{-1} = \mathbf{0}$, and assuming $\mathbf{A}$ is scaled so that $\|\mathbf{A}\|_F^2 \approx N$. In (2), $\mathbf{g} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ is an estimation function chosen based on prior knowledge about $\mathbf{x}^0$, and $\langle \nabla \mathbf{g}(\mathbf{r}, \gamma) \rangle := \frac{1}{N} \sum_{n=1}^N \frac{\partial g_n(\mathbf{r},\gamma)}{\partial r_n}$ denotes the divergence of $\mathbf{g}(\mathbf{r}, \gamma)$. For example, if $\mathbf{x}^0$ is known to be sparse, then it is common to choose $\mathbf{g}(\cdot)$ to be the componentwise soft-thresholding function, in which case AMP iteratively solves the LASSO [5] problem.

Importantly, for large, i.i.d., sub-Gaussian random matrices $\mathbf{A}$ and Lipschitz denoisers $\mathbf{g}(\cdot)$, the performance of AMP can be exactly predicted by a scalar *state evolution* (SE), which also provides testable conditions for optimality [6, 7, 8]. The initial work [6, 7] focused on the case where $\mathbf{g}(\cdot)$ is a separable function with identical components (i.e., $[\mathbf{g}(\mathbf{r}, \gamma)]_n = g(r_n, \gamma)\ \forall n$), while the later work [8] allowed non-separable $\mathbf{g}(\cdot)$. Interestingly, these SE analyses establish the fact that

$$\mathbf{r}_k = \mathbf{x}^0 + \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_k), \tag{3}$$

leading to the important interpretation that $\mathbf{g}(\cdot)$ acts as a *denoiser*. This interpretation provides guidance on how to choose $\mathbf{g}(\cdot)$. For example, if $\mathbf{x}$ is i.i.d. with a known prior, then (3) suggests to choose a separable $\mathbf{g}(\cdot)$ composed of minimum mean-squared error (MMSE) scalar denoisers $g(r_n, \gamma) = \mathbb{E}(x_n | r_n = x_n + \mathcal{N}(0, 1/\gamma))$. In this case, [6, 7] established that, whenever the SE has a unique fixed point, the estimates $\widehat{\mathbf{x}}_k$ generated by AMP converge to the Bayes optimal estimate of $\mathbf{x}^0$ from $\mathbf{y}$. As another example, if $\mathbf{x}$ is a natural image, for which an analytical prior is lacking, then (3) suggests to choose $\mathbf{g}(\cdot)$ as a sophisticated image-denoising algorithm like BM3D [9] or DnCNN [10], as proposed in [11]. Many other examples of structured estimators $\mathbf{g}(\cdot)$ can be considered; we refer the reader to [8] and Section 5. Prior to [8], AMP SE results were established for special cases of $\mathbf{g}(\cdot)$ in [12, 13]. Plug-in denoisers have been combined in related algorithms [14, 15, 16].

An important limitation of AMP's SE is that it holds only for large, i.i.d., sub-Gaussian $\mathbf{A}$. AMP itself often fails to converge with small deviations from i.i.d. sub-Gaussian $\mathbf{A}$, such as when $\mathbf{A}$ is mildly ill-conditioned or non-zero-mean [4, 17, 18]. Recently, a robust alternative to AMP called *vector AMP* (VAMP) was proposed and analyzed in [19], based closely on expectation propagation [20]—see also [21, 22, 23]. There it was established that, if $\mathbf{A}$ is a large right-rotationally invariant random matrix and $\mathbf{g}(\cdot)$ is a separable Lipschitz denoiser, then VAMP's performance can be exactly predicted by a scalar SE, which also provides testable conditions for optimality. Importantly, VAMP applies to arbitrarily conditioned matrices $\mathbf{A}$, which is a significant benefit over AMP, since it is known that ill-conditioning is one of AMP's main failure mechanisms [4, 17, 18].

Unfortunately, the SE analyses of VAMP in [24] and its extension in [25] are limited to separable denoisers. This limitation prevents a full understanding of VAMP's behavior when used with non-separable denoisers, such as state-of-the-art image-denoising methods as recently suggested in [26]. The main contribution of this work is to show that the SE analysis of VAMP can be extended to a large class of non-separable denoisers that are Lipschitz continuous and satisfy a certain convergence property. The conditions are similar to those used in the analysis of AMP with non-separable denoisers in [8]. We show that there are several interesting non-separable denoisers that satisfy these conditions, including group-structured and convolutional neural network based denoisers.

An extended version with all proofs and other details are provided in [27].

2 Review of Vector AMP

The steps of VAMP algorithm of [19] are shown in Algorithm 1. Each iteration has two parts: A denoiser step and a Linear MMSE (LMMSE) step. These are characterized by *estimation functions* $\mathbf{g}_1(\cdot)$ and $\mathbf{g}_2(\cdot)$ producing estimates $\widehat{\mathbf{x}}_{1k}$ and $\widehat{\mathbf{x}}_{2k}$. The estimation functions take inputs $\mathbf{r}_{1k}$ and $\mathbf{r}_{2k}$ that we call *partial estimates*. The LMMSE estimation function is given by,

$$\mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k}) := \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1} \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{y} + \gamma_{2k}\mathbf{r}_{2k}\right), \tag{4}$$

where $\gamma_w > 0$ is a parameter representing an estimate of the precision (inverse variance) of the noise $\mathbf{w}$ in (1). The estimate $\widehat{\mathbf{x}}_{2k}$ is thus an MMSE estimator, treating the $\mathbf{x}$ as having a Gaussian prior with mean given by the partial estimate $\mathbf{r}_{2k}$. The estimation function $\mathbf{g}_1(\cdot)$ is called the *denoiser* and can be designed identically to the denoiser $\mathbf{g}(\cdot)$ in the AMP iterations (2). In particular, the denoiser is used to incorporate the structural or prior information on $\mathbf{x}$. As in AMP, in lines 5 and 11, $\langle \nabla \mathbf{g}_i \rangle$ denotes the normalized divergence.

Algorithm 1 Vector AMP (LMMSE form)

```
Require: LMMSE estimator $\mathbf{g}_2(\cdot,\gamma_{2k})$ from (4), denoiser $\mathbf{g}_1(\cdot,\gamma_{1k})$, and number of iterations $K_{\rm it}$.
Select initial $\mathbf{r}_{10}$ and $\gamma_{10} \geq 0$.
for $k = 0, 1, \ldots, K_{\rm it}$ do
    // Denoising
    $\widehat{\mathbf{x}}_{1k} = \mathbf{g}_1(\mathbf{r}_{1k}, \gamma_{1k})$
    $\alpha_{1k} = \langle \nabla \mathbf{g}_1(\mathbf{r}_{1k}, \gamma_{1k}) \rangle$
    $\eta_{1k} = \gamma_{1k}/\alpha_{1k}$, $\gamma_{2k} = \eta_{1k} - \gamma_{1k}$
    $\mathbf{r}_{2k} = (\eta_{1k}\widehat{\mathbf{x}}_{1k} - \gamma_{1k}\mathbf{r}_{1k})/\gamma_{2k}$

    // LMMSE estimation
    $\widehat{\mathbf{x}}_{2k} = \mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k})$
    $\alpha_{2k} = \langle \nabla \mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k}) \rangle$
    $\eta_{2k} = \gamma_{2k}/\alpha_{2k}$, $\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$
    $\mathbf{r}_{1,k+1} = (\eta_{2k}\widehat{\mathbf{x}}_{2k} - \gamma_{2k}\mathbf{r}_{2k})/\gamma_{1,k+1}$
end for
Return $\widehat{\mathbf{x}}_{1K_{\rm it}}$.
```

The main result of [24] is that, under suitable conditions, VAMP admits a state evolution (SE) analysis that precisely describes the mean squared error (MSE) of the estimates $\widehat{\mathbf{x}}_{1k}$ and $\widehat{\mathbf{x}}_{2k}$ in a certain large system limit (LSL). Importantly, VAMP's SE analysis applies to arbitrary right rotationally invariant $\mathbf{A}$. This class is considerably larger than the set of sub-Gaussian i.i.d. matrices for which AMP applies. However, the SE analysis in [24] is restricted separable Lipschitz denoisers that can be described as follows: Let $g_{1n}(\mathbf{r}_1, \gamma_1)$ be the n-th component of the output of $\mathbf{g}_1(\mathbf{r}_1, \gamma_1)$. Then, it is assumed that,

$$\widehat{x}_{1n} = g_{1n}(\mathbf{r}_1, \gamma_1) = \phi(r_{1n}, \gamma_1), \tag{5}$$

for some function scalar-output function $\phi(\cdot)$ that does not depend on the component index n. Thus, the estimator is separable in the sense that the n-th component of the estimate, $\widehat{x}_{1n}$ depends only on the n-th component of the input r_{1n} as well as the precision level γ_1. In addition, it is assumed that $\phi(r_1, \gamma_1)$ satisfies a certain Lipschitz condition. The separability assumption precludes the analysis of more general denoisers mentioned in the Introduction.

3 Extending the Analysis to Non-Separable Denoisers

The main contribution of the paper is to extend the state evolution analysis of VAMP to a class of denoisers that we call *uniformly Lipschitz* and *convergent under Gaussian noise*. This class is significantly larger than separable Lipschitz denoisers used in [24]. To state these conditions precisely, consider a sequence of estimation problems, indexed by a vector dimension N. For each N, suppose there is some "true" vector $\mathbf{u} = \mathbf{u}(N) \in \mathbb{R}^N$ that we wish to estimate from noisy measurements of the form, $\mathbf{r} = \mathbf{u} + \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^N$ is Gaussian noise. Let $\widehat{\mathbf{u}} = \mathbf{g}(\mathbf{r}, \gamma)$ be some estimator, parameterized by γ.

Definition 1. *The sequence of estimators* $\mathbf{g}(\cdot)$ *are said to be* *uniformly Lipschitz continuous* *if there exists constants* A*,* B *and* $C > 0$*, such that*

$$\|\mathbf{g}(\mathbf{r}_2, \gamma_2) - \mathbf{g}(\mathbf{r}_1, \gamma_1)\| \leq (A + B|\gamma_2 - \gamma_1|)\|\mathbf{r}_2 - \mathbf{r}_1\| + C\sqrt{N}|\gamma_2 - \gamma_1|, \tag{6}$$

for any $\mathbf{r}_1, \mathbf{r}_2, \gamma_1, \gamma_2$ *and* N*.*

Definition 2. *The sequence of random vectors* $\mathbf{u}$ *and estimators* $\mathbf{g}(\cdot)$ *are said to be* *convergent under Gaussian noise* *if the following condition holds: Let* $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^N$ *be two sequences where* (z_{1n}, z_{2n}) *are i.i.d. with* $(z_{1n}, z_{2n}) = \mathcal{N}(0, \mathbf{S})$ *for some positive definite covariance* $\mathbf{S} \in \mathbb{R}^{2\times 2}$*. Then, all the following limits exist almost surely:*

$$\lim_{N\rightarrow\infty} \frac{1}{N}\mathbf{g}(\mathbf{u}+\mathbf{z}_1, \gamma_1)^{\mathsf{T}}\mathbf{g}(\mathbf{u}+\mathbf{z}_2, \gamma_2), \quad \lim_{N\rightarrow\infty} \frac{1}{N}\mathbf{g}(\mathbf{u}+\mathbf{z}_1, \gamma_1)^{\mathsf{T}}\mathbf{u}, \tag{7a}$$

$$\lim_{N\rightarrow\infty} \frac{1}{N}\mathbf{u}^{\mathsf{T}}\mathbf{z}_1, \quad \lim_{N\rightarrow\infty} \frac{1}{N}\|\mathbf{u}\|^2 \tag{7b}$$

$$\lim_{N\rightarrow\infty} \langle \nabla \mathbf{g}(\mathbf{u}+\mathbf{z}_1, \gamma_1) \rangle = \frac{1}{NS_{12}}\mathbf{g}(\mathbf{u}+\mathbf{z}_1, \gamma_1)^{\mathsf{T}}\mathbf{z}_2, \tag{7c}$$

for all γ_1, γ_2 and covariance matrices $\mathbf{S}$. Moreover, the values of the limits are continuous in $\mathbf{S}$, γ_1 and γ_2.

With these definitions, we make the following key assumption on the denoiser.

Assumption 1. *For each N, suppose that we have a "true" random vector $\mathbf{x}^0 \in \mathbb{R}^N$ and a denoiser $\mathbf{g}_1(\mathbf{r}_1, \gamma_1)$ acting on signals $\mathbf{r}_1 \in \mathbb{R}^N$. Following Definition 1, we assume the sequence of denoiser functions indexed by N, is uniformly Lipschitz continuous. In addition, the sequence of true vectors $\mathbf{x}^0$ and denoiser functions are convergent under Gaussian noise following Definition 2.*

The first part of Assumption 1 is relatively standard: Lipschitz and uniform Lipschitz continuity of the denoiser is assumed several AMP-type analyses including [6, 28, 24] What is new is the assumption in Definition 2. This assumption relates to the behavior of the denoiser $\mathbf{g}_1(\mathbf{r}_1, \gamma_1)$ in the case when the input is of the form, $\mathbf{r}_1 = \mathbf{x}^0 + \mathbf{z}$. That is, the input is the true signal with a Gaussian noise perturbation. In this setting, we will be requiring that certain correlations converge. Before continuing our analysis, we briefly show that separable denoisers as well as several interesting non-separable denoisers satisfy these conditions.

Separable Denoisers. We first show that the class of denoisers satisfying Assumption 1 includes the separable Lipschitz denoisers studied in most AMP analyses such as [6]. Specifically, suppose that the true vector $\mathbf{x}^0$ has i.i.d. components with bounded second moments and the denoiser $\mathbf{g}_1(\cdot)$ is separable in that it is of the form (5). Under a certain uniform Lipschitz condition, it is shown in the extended version of this paper [27] that the denoiser satisfies Assumption 1.

Group-Based Denoisers. As a first non-separable example, let us suppose that the vector $\mathbf{x}^0$ can be represented as an $L \times K$ matrix. Let $\mathbf{x}^0_\ell \in \mathbb{R}^K$ denote the ℓ-th row and assume that the rows are i.i.d. Each row can represent a *group*. Suppose that the denoiser $\mathbf{g}_1(\cdot)$ is *groupwise separable*. That is, if we denote by $\mathbf{g}_{1\ell}(\mathbf{r}, \ell)$ the ℓ-th row of the output of the denoiser, we assume that

$$\mathbf{g}_{1\ell}(\mathbf{r}, \gamma) = \phi(\mathbf{r}_\ell, \gamma) \in \mathbb{R}^K, \tag{8}$$

for a vector-valued function $\phi(\cdot)$ that is the same for all rows. Thus, the ℓ-th row output $\mathbf{g}_\ell(\cdot)$ depends only on the ℓ-th row input. Such groupwise denoisers have been used in AMP and EP-type methods for group LASSO and other structured estimation problems [29, 30, 31]. Now, consider the limit where the group size K is fixed, and the number of groups $L \to \infty$. Then, under suitable Lipschitz continuity conditions, the extended version of this paper [27] shows that groupwise separable denoiser also satisfies Assumption 1.

Convolutional Denoisers. As another non-separable denoiser, suppose that, for each N, $\mathbf{x}^0$ is an N sample segment of a stationary, ergodic process with bounded second moments. Suppose that the denoiser is given by a linear convolution,

$$\mathbf{g}_1(\mathbf{r}_1) := T_N(\mathbf{h} * \mathbf{r}_1), \tag{9}$$

where $\mathbf{h}$ is a finite length filter and $T_N(\cdot)$ truncates the signal to its first N samples. For simplicity, we assume there is no dependence on γ_1. Convolutional denoising arises in many standard linear estimation operations on wide sense stationary processes such as Weiner filtering and smoothing [32]. If we assume that $\mathbf{h}$ remains constant and $N \to \infty$, the extended version of this paper [27] shows that the sequence of random vectors $\mathbf{x}^0$ and convolutional denoisers $\mathbf{g}_1(\cdot)$ satisfies Assumption 1.

Convolutional Neural Networks. In recent years, there has been considerable interest in using trained deep convolutional neural networks for image denoising [33, 34]. As a simple model for such a denoiser, suppose that the denoiser is a composition of maps,

$$\mathbf{g}_1(\mathbf{r}_1) = (F_L \circ F_{L-1} \circ \cdots \circ F_1)(\mathbf{r}_1), \tag{10}$$

where $F_\ell(\cdot)$ is a sequence of layer maps where each layer is either a multi-channel convolutional operator or Lipschitz separable activation function, such as sigmoid or ReLU. Under mild assumptions on the maps, it is shown in the extended version of this paper [27] that the estimator sequence $\mathbf{g}_1(\cdot)$ can also satisfy Assumption 1.

Singular-Value Thresholding (SVT) Denoiser. Consider the estimation of a low-rank matrix $\mathbf{X}^0$ from linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}^0)$, where $\mathcal{A}$ is some linear operator [35]. Writing the SVD of $\mathbf{R}$ as $\mathbf{R} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$, the SVT denoiser is defined as

$$\mathbf{g}_1(\mathbf{R}, \gamma) := \sum_i (\sigma_i - \gamma)_+ \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}, \tag{11}$$

where $(x)_+ := \max\{0, x\}$. In the extended version of this paper [27], we show that $\mathbf{g}_1(\cdot)$ satisfies Assumption 1.

4 Large System Limit Analysis

4.1 System Model

Our main theoretical contribution is to show that the SE analysis of VAMP in [19] can be extended to the non-separable case. We consider a sequence of problems indexed by the vector dimension N. For each N, we assume that there is a “true" random vector $\mathbf{x}^0 \in \mathbb{R}^N$ observed through measurements $\mathbf{y} \in \mathbb{R}^M$ of the form in (1) where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_{w0}^{-1}\mathbf{I})$. We use γ_{w0} to denote the “true" noise precision to distinguish this from the postulated precision, γ_w, used in the LMMSE estimator (4). Without loss of generality (see below), we assume that $M = N$. We assume that $\mathbf{A}$ has an SVD,

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}, \quad \mathbf{S} = \mathrm{diag}(\mathbf{s}), \quad \mathbf{s} = (s_1, \ldots, s_N), \tag{12}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{S}$ is non-negative and diagonal. The matrix $\mathbf{U}$ is arbitrary, $\mathbf{s}$ is an i.i.d. random vector with components $s_i \in [0, s_{max}]$ almost surely. Importantly, we assume that $\mathbf{V}$ is Haar distributed, meaning that it is uniform on the $N \times N$ orthogonal matrices. This implies that $\mathbf{A}$ is *right rotationally invariant* meaning that $\mathbf{A} \overset{d}{=} \mathbf{A}\mathbf{V}_0$ for any orthogonal matrix $\mathbf{V}_0$. We also assume that $\mathbf{w}$, $\mathbf{x}^0$, $\mathbf{s}$ and $\mathbf{V}$ are all independent. As in [19], we can handle the case of rectangular $\mathbf{V}$ by zero padding $\mathbf{s}$.

These assumptions are similar to those in [19]. The key new assumption is Assumption 1. Given such a denoiser and postulated variance γ_w, we run the VAMP algorithm, Algorithm 1. We assume that the initial condition is given by,

$$\mathbf{r} = \mathbf{x}^0 + \mathcal{N}(\mathbf{0}, \tau_{10}\mathbf{I}), \tag{13}$$

for some initial error variance τ_{10}. In addition, we assume

$$\lim_{N\to\infty} \gamma_{10} = \overline{\gamma}_{10}, \tag{14}$$

almost surely for some $\overline{\gamma}_{10} \geq 0$.

Analogous to [24], we define two key functions: *error functions* and *sensitivity functions*. The error functions characterize the MSEs of the denoiser and LMMSE estimator under AWGN measurements. For the denoiser $\mathbf{g}_1(\cdot, \gamma_1)$, we define the error function as

$$\mathcal{E}_1(\gamma_1, \tau_1) := \lim_{N\to\infty} \frac{1}{N} \|\mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}, \gamma_1) - \mathbf{x}^0\|^2, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \tau_1\mathbf{I}), \tag{15}$$

and, for the LMMSE estimator, as

$$\begin{aligned}\mathcal{E}_2(\gamma_2, \tau_2) &:= \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\|\mathbf{g}_2(\mathbf{r}_2, \gamma_2) - \mathbf{x}^0\|^2, \\ &\mathbf{r}_2 = \mathbf{x}^0 + \mathcal{N}(0, \tau_2\mathbf{I}), \quad \mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathcal{N}(0, \gamma_{w0}^{-1}\mathbf{I}).\end{aligned} \tag{16}$$

The limit (15) exists almost surely due to the assumption of $\mathbf{g}_1(\cdot)$ being convergent under Gaussian noise. Although $\mathcal{E}_2(\gamma_2, \tau_2)$ implicitly depends on the precisions γ_{w0} and γ_w, we omit this dependence to simplify the notation. We also define the *sensitivity functions* as

$$\mathcal{A}_i(\gamma_i, \tau_i) := \lim_{N\to\infty} \langle \nabla \mathbf{g}_i(\mathbf{x}^0 + \mathbf{z}_i, \gamma_i) \rangle, \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \tau_i\mathbf{I}). \tag{17}$$

The LMMSE error function (16) and sensitivity functions (17) are identical to those in the VAMP analysis [19]. The denoiser error function (15) generalizes the error function in [19] for non-separable denoisers.

4.2 State Evolution of VAMP

We now show that the VAMP algorithm with a non-separable denoiser follows the identical state evolution equations as the separable case given in [19]. Define the error vectors,

$$\mathbf{p}_k := \mathbf{r}_{1k} - \mathbf{x}^0, \quad \mathbf{q}_k := \mathbf{V}^{\mathsf{T}}(\mathbf{r}_{2k} - \mathbf{x}^0). \tag{18}$$

Thus, $\mathbf{p}_k$ represents the error between the partial estimate $\mathbf{r}_{1k}$ and the true vector $\mathbf{x}^0$. The error vector $\mathbf{q}_k$ represents the transformed error $\mathbf{r}_{2k} - \mathbf{x}^0$. The SE analysis will show that these errors are asymptotically Gaussian. In addition, the analysis will exactly predict the variance on the partial estimate errors (18) and estimate errors, $\widehat{\mathbf{x}}_i - \mathbf{x}^0$. These variances are computed recursively through what we will call the *state evolution* equations:

$$\overline{\alpha}_{1k} = \mathcal{A}_1(\overline{\gamma}_{1k}, \tau_{1k}), \quad \overline{\eta}_{1k} = \frac{\overline{\gamma}_{1k}}{\overline{\alpha}_{1k}}, \quad \overline{\gamma}_{2k} = \overline{\eta}_{1k} - \overline{\gamma}_{1k} \tag{19a}$$

$$\tau_{2k} = \frac{1}{(1-\overline{\alpha}_{1k})^2}\left[\mathcal{E}_1(\overline{\gamma}_{1k}, \tau_{1k}) - \overline{\alpha}_{1k}^2 \tau_{1k}\right], \tag{19b}$$

$$\overline{\alpha}_{2k} = \mathcal{A}_2(\overline{\gamma}_{2k}, \tau_{2k}), \quad \overline{\eta}_{2k} = \frac{\overline{\gamma}_{2k}}{\overline{\alpha}_{2k}}, \quad \overline{\gamma}_{1,k+1} = \overline{\eta}_{2k} - \overline{\gamma}_{2k} \tag{19c}$$

$$\tau_{1,k+1} = \frac{1}{(1-\overline{\alpha}_{2k})^2}\left[\mathcal{E}_2(\overline{\gamma}_{2k}, \tau_{2k}) - \overline{\alpha}_{2k}^2 \tau_{2k}\right], \tag{19d}$$

which are initialized with $k = 0$, τ_{10} in (13) and $\overline{\gamma}_{10}$ defined from the limit (14). The SE equations in (19) are identical to those in [19] with the new error and sensitivity functions for the non-separable denoisers. We can now state our main result, which is proven in the extended version of this paper [27].

Theorem 1. *Under the above assumptions and definitions, assume that the sequence of true random vectors $\mathbf{x}^0$ and denoisers $\mathbf{g}_1(\mathbf{r}_1, \gamma_1)$ satisfy Assumption 1. Assume additionally that, for all iterations k, the solution $\overline{\alpha}_{1k}$ from the SE equations* (19) *satisfies $\overline{\alpha}_{1k} \in (0,1)$ and $\overline{\gamma}_{ik} > 0$. Then,*

(a) For any k, the error vectors on the partial estimates, $\mathbf{p}_k$ and $\mathbf{q}_k$ in (18) *can be written as,*

$$\mathbf{p}_k = \widetilde{\mathbf{p}}_k + O(\tfrac{1}{\sqrt{N}}), \quad \mathbf{q}_k = \widetilde{\mathbf{q}}_k + O(\tfrac{1}{\sqrt{N}}), \tag{20}$$

where, $\widetilde{\mathbf{p}}_k$ and $\widetilde{\mathbf{q}}_k \in \mathbb{R}^N$ are each i.i.d. Gaussian random vectors with zero mean and per component variance τ_{1k} and τ_{2k}, respectively.

(b) For any fixed iteration $k \geq 0$, and $i = 1, 2$, we have, almost surely

$$\lim_{N\to\infty} \frac{1}{N}\|\widehat{\mathbf{x}}_i - \mathbf{x}^0\|^2 = \frac{1}{\overline{\eta}_{ik}}, \quad \lim_{N\to\infty}(\alpha_{ik}, \eta_{ik}, \gamma_{ik}) = (\overline{\alpha}_{ik}, \overline{\eta}_{ik}, \overline{\gamma}_{ik}). \tag{21}$$

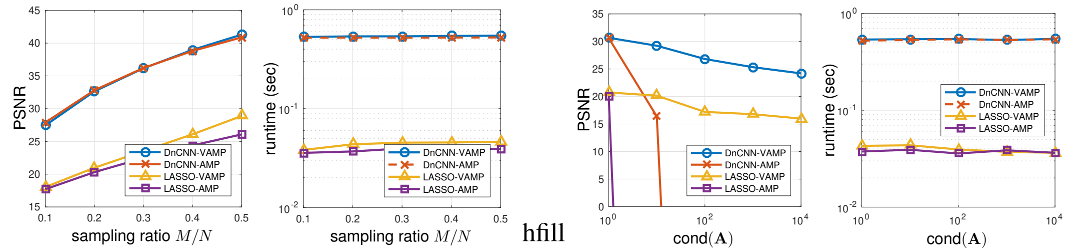
In (20), we have used the notation, that when $\mathbf{u}, \widetilde{\mathbf{u}} \in \mathbb{R}^N$ are sequences of random vectors, $\mathbf{u} = \widetilde{\mathbf{u}} + O(\frac{1}{\sqrt{N}})$ means $\lim_{N\to\infty} \frac{1}{N}\|\mathbf{u} - \widetilde{\mathbf{u}}\|^2 = 0$ almost surely. Part (a) of Theorem 1 thus shows that the error vectors $\mathbf{p}_k$ and $\mathbf{q}_k$ in (18) are approximately i.i.d. Gaussian. The result is a natural extension to the main result on separable denoisers in [19]. Moreover, the variance on the variance on the errors, along with the mean squared error (MSE) of the estimates $\widehat{\mathbf{x}}_{ik}$ can be exactly predicted by the same SE equations as the separable case. The result thus provides an asymptotically exact analysis of VAMP extended to non-separable denoisers.

5 Numerical Experiments

5.1 Compressive Image Recovery

We first consider the problem of compressive image recovery, where the goal is to recover an image $\mathbf{x}^0 \in \mathbb{R}^N$ from measurements $\mathbf{y} \in \mathbb{R}^M$ of the form (1) with $M \ll N$. This problem arises in many imaging applications, such as magnetic resonance imaging, radar imaging, computed tomography, etc., although the details of $\mathbf{A}$ and $\mathbf{x}^0$ change in each case.

One of the most popular approaches to image recovery is to exploit sparsity in the wavelet transform coefficients $\mathbf{c} := \mathbf{\Psi}\mathbf{x}^0$, where $\mathbf{\Psi}$ is a suitable orthonormal wavelet transform. Rewriting (1) as



hfill

(a) Average PSNR and runtime with vs. M/N with well-conditioned $\mathbf{A}$ and no noise after 12 iterations.

(b) Average PSNR and runtime versus cond$(\mathbf{A})$ at $M/N = 0.2$ and no noise after 10 iterations.

Figure 1: Compressive image recovery: PSNR and runtime vs. rate M/N and cond$(\mathbf{A})$

$\mathbf{y} = \mathbf{A}\mathbf{\Psi}\mathbf{c} + \mathbf{w}$, the idea is to first estimate $\mathbf{c}$ from $\mathbf{y}$ (e.g., using LASSO) and then form the image estimate via $\widehat{\mathbf{x}} = \mathbf{\Psi}^\mathsf{T}\widehat{\mathbf{c}}$. Although many algorithms exist to solve the LASSO problem, the AMP algorithms are among the fastest (see, e.g., [36, Fig.1]). As an alternative to the sparsity-based approach, it was recently suggested in [11] to recover $\mathbf{x}^0$ directly using AMP (2) by choosing the estimation function $\mathbf{g}$ as a sophisticated image-denoising algorithm like BM3D [9] or DnCNN [10].

Figure 1a compares the LASSO- and DnCNN-based versions of AMP and VAMP for 128×128 image recovery under well-conditioned $\mathbf{A}$ and no noise. Here, $\mathbf{A} = \mathbf{JPHD}$, where $\mathbf{D}$ is a diagonal matrix with random ± 1 entries, $\mathbf{H}$ is a discrete Hadamard transform (DHT), $\mathbf{P}$ is a random permutation matrix, and $\mathbf{J}$ contains the first M rows of $\mathbf{I}_N$. The results average over the well-known *lena*, *barbara*, *boat*, *house*, and *peppers* images using 10 random draws of $\mathbf{A}$ for each. The figure shows that AMP and VAMP have very similar runtimes and PSNRs when $\mathbf{A}$ is well-conditioned, and that the DnCNN approach is about 10 dB more accurate, but 10× as slow, as the LASSO approach. Figure 2 shows the state-evolution prediction of VAMP's PSNR on the *barbara* image at $M/N = 0.5$, averaged over 50 draws of $\mathbf{A}$. The state-evolution accurately predicts the PSNR of VAMP.

To test the robustness to the condition number of $\mathbf{A}$, we repeated the experiment from Fig. 1a using $\mathbf{A} = \mathbf{J}\text{Diag}(\mathbf{s})\mathbf{PHD}$, where $\text{Diag}(\mathbf{s})$ is a diagonal matrix of singular values. The singular values were geometrically spaced, i.e., $s_m/s_{m-1} = \rho\ \forall m$, with ρ chosen to achieve a desired cond$(\mathbf{A}) := s_1/s_M$. The sampling rate was fixed at $M/N = 0.2$, and the measurements were noiseless, as before. The results, shown in Fig. 1b, show that AMP diverged when cond$(\mathbf{A}) \geq 10$, while VAMP exhibited only a mild PSNR degradation due to ill-conditioned $\mathbf{A}$. The original images and example image recoveries are included in the extended version of this paper.

5.2 Bilinear Estimation via Lifting

We now use the structured linear estimation model (1) to tackle problems in *bilinear* estimation through a technique known as "lifting" [37, 38, 39, 40]. In doing so, we are motivated by applications like blind deconvolution [41], self-calibration [39], compressed sensing (CS) with matrix uncertainty [42], and joint channel-symbol estimation [43]. All cases yield measurements $\mathbf{y}$ of the form

$$\mathbf{y} = \left(\textstyle\sum_{l=1}^{L} b_l \mathbf{\Phi}_l\right)\mathbf{c} + \mathbf{w} \in \mathbb{R}^M, \tag{22}$$

where $\{\mathbf{\Phi}_l\}_{l=1}^{L}$ are known, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_w)$, and the objective is to recover both $\mathbf{b} := [b_1, \dots, b_L]^\mathsf{T}$ and $\mathbf{c} \in \mathbb{R}^P$. This bilinear problem can be "lifted" into a linear problem of the form (1) by setting

$$\mathbf{A} = [\mathbf{\Phi}_1 \quad \mathbf{\Phi}_2 \quad \cdots \quad \mathbf{\Phi}_L] \in \mathbb{R}^{M \times LP} \text{ and } \mathbf{x} = \text{vec}(\mathbf{c}\mathbf{b}^\mathsf{T}) \in \mathbb{R}^{LP}, \tag{23}$$

where vec$(\mathbf{X})$ vectorizes $\mathbf{X}$ by concatenating its columns. When $\mathbf{b}$ and $\mathbf{c}$ are i.i.d. with known priors, the MMSE denoiser $\mathbf{g}(\mathbf{r}, \gamma) = \mathbb{E}(\mathbf{x}|\mathbf{r} = \mathbf{x} + \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma))$ can be implemented near-optimally by the rank-one AMP algorithm from [44] (see also [45, 46, 47]), with divergence estimated as in [11].

We first consider *CS with matrix uncertainty* [42], where b_1 is known. For these experiments, we generated the unknown $\{b_l\}_{l=2}^{L}$ as i.i.d. $\mathcal{N}(0, 1)$ and the unknown $\mathbf{c} \in \mathbb{R}^P$ as K-sparse with $\mathcal{N}(0, 1)$ nonzero entries. Fig. 2 shows that the MSE on $\mathbf{x}$ of lifted VAMP is very close to its SE prediction when $K = 12$. We then compared lifted VAMP to PBiGAMP from [48], which applies AMP directly to the (non-lifted) bilinear problem, and to WSS-TLS from [42], which uses non-convex optimization. We also compared to MMSE estimation of $\mathbf{b}$ under oracle knowledge of $\mathbf{c}$, and MMSE

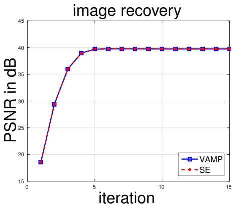


Figure 2: SE prediction & VAMP for image recovery and CS with matrix uncertainty

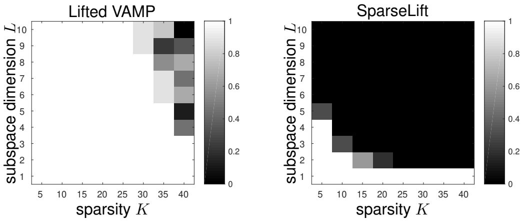
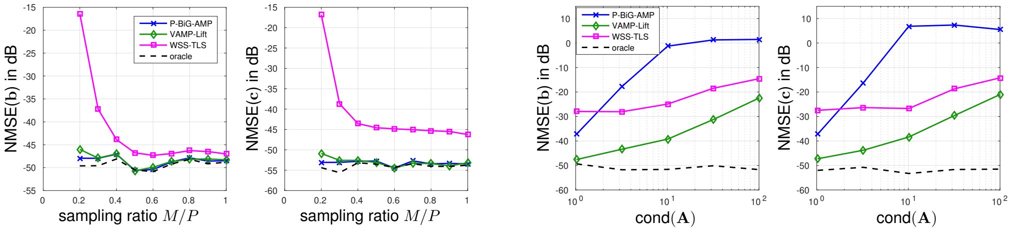


Figure 3: Self-calibration: Success rate vs. sparsity K and subspace dimension L



(a) NMSE vs. M/P with i.i.d. $\mathcal{N}(0,1)$ $\mathbf{A}$.

(b) NMSE vs. cond$(\mathbf{A})$ at $M/P = 0.6$.

Figure 4: Compressive sensing with matrix uncertainty

estimation of $\mathbf{c}$ under oracle knowledge of support$(\mathbf{c})$ and $\mathbf{b}$. For $b_1 = \sqrt{20}$, $L = 11$, $P = 256$, $K = 10$, i.i.d. $\mathcal{N}(0,1)$ matrix $\mathbf{A}$, and SNR $= 40$ dB, Fig. 4a shows the normalized MSE on $\mathbf{b}$ (i.e., $\mathsf{NMSE}(\mathbf{b}) := \mathbb{E}\|\widehat{\mathbf{b}} - \mathbf{b}^0\|^2/\mathbb{E}\|\mathbf{b}^0\|^2$) and $\mathbf{c}$ versus sampling ratio M/P. This figure demonstrates that lifted VAMP and PBiGAMP perform close to the oracles and much better than WSS-TLS.

Although lifted VAMP performs similarly to PBiGAMP in Fig. 4a, its advantage over PBiGAMP becomes apparent with non-i.i.d. $\mathbf{A}$. For illustration, we repeated the previous experiment, but with $\mathbf{A}$ constructed using the SVD $\mathbf{A} = \mathbf{U}\text{Diag}(\mathbf{s})\mathbf{V}^{\mathsf{T}}$ with Haar distributed $\mathbf{U}$ and $\mathbf{V}$ and geometrically spaced $\mathbf{s}$. Also, to make the problem more difficult, we set $b_1 = 1$. Figure 4b shows the normalized MSE on $\mathbf{b}$ and $\mathbf{c}$ versus cond$(\mathbf{A})$ at $M/P = 0.6$. There it can be seen that lifted VAMP is much more robust than PBiGAMP to the conditioning of $\mathbf{A}$.

We next consider the *self-calibration* problem [39], where the measurements take the form

$$\mathbf{y} = \text{Diag}(\mathbf{H}\mathbf{b})\boldsymbol{\Psi}\mathbf{c} + \mathbf{w} \in \mathbb{R}^M. \tag{24}$$

Here the matrices $\mathbf{H} \in \mathbb{R}^{M\times L}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{M\times P}$ are known and the objective is to recover the unknown vectors $\mathbf{b}$ and $\mathbf{c}$. Physically, the vector $\mathbf{H}\mathbf{b}$ represents unknown calibration gains that lie in a known subspace, specified by $\mathbf{H}$. Note that (24) is an instance of (22) with $\boldsymbol{\Phi}_l = \text{Diag}(\mathbf{h}_l)\boldsymbol{\Psi}$, where $\mathbf{h}_l$ denotes the lth column of $\mathbf{H}$. Different from "CS with matrix uncertainty," all elements in $\mathbf{b}$ are now unknown, and so WSS-TLS [42] cannot be applied. Instead, we compare lifted VAMP to the SparseLift approach from [39], which is based on convex relaxation and has provable guarantees. For our experiment, we generated $\boldsymbol{\Psi}$ and $\mathbf{b} \in \mathbb{R}^L$ as i.i.d. $\mathcal{N}(0,1)$; $\mathbf{c}$ as K-sparse with $\mathcal{N}(0,1)$ nonzero entries; $\mathbf{H}$ as randomly chosen columns of a Hadamard matrix; and $\mathbf{w} = \mathbf{0}$. Figure 3 plots the success rate versus L and K, where "success" is defined as $\mathbb{E}\|\widehat{\mathbf{c}\mathbf{b}}^{\mathsf{T}} - \mathbf{c}^0(\mathbf{b}^0)^{\mathsf{T}}\|_F^2/\mathbb{E}\|\mathbf{c}^0(\mathbf{b}^0)^{\mathsf{T}}\|_F^2 < -60$ dB. The figure shows that, relative to SparseLift, lifted VAMP gives successful recoveries for a wider range of L and K.

6 Conclusions

We have extended the analysis of the method in [24] to a class of non-separable denoisers. The method provides a computational efficient method for reconstruction where structural information and constraints on the unknown vector can be incorporated in a modular manner. Importantly, the method admits a rigorous analysis that can provide precise predictions on the performance in high-dimensional random settings.

Acknowledgments

A. K. Fletcher and P. Pandit were supported in part by the National Science Foundation under Grants 1738285 and 1738286 and the Office of Naval Research under Grant N00014-15-1-2677. S. Rangan was supported in part by the National Science Foundation under Grants 1116589, 1302336, and 1547332, and the industrial affiliates of NYU WIRELESS. P. Schniter and S. Sarkar were supported in part by the National Science Foundation under Grant CCF-1716388.

References

[1] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[2] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE ISIT*, 2011, pp. 2174–2178.

[3] S. Rangan, P. Schniter, E. Riegler, A. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," in *Proc. IEEE ISIT*, Jul. 2013, pp. 664–668.

[4] S. Rangan, P. Schniter, and A. K. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. IEEE ISIT*, Jul. 2014, pp. 236–240.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[7] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference*, vol. 2, no. 2, pp. 115–144, 2013.

[8] R. Berthier, A. Montanari, and P.-M. Nguyen, "State evolution for approximate message passing with non-separable functions," *arXiv preprint arXiv:1708.03950*, 2017.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[10] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[11] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Trans. Info. Thy.*, vol. 62, no. 9, pp. 5117–5144, 2016.

[12] D. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *IEEE Trans. Info. Thy.*, vol. 59, no. 6, pp. 3396–3433, 2013.

[13] Y. Ma, C. Rush, and D. Baron, "Analysis of approximate message passing with a class of non-separable denoisers," in *Proc. ISIT*, 2017, pp. 231–235.

[14] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2013, pp. 945–948.

[15] S. Chen, C. Luo, B. Deng, Y. Qin, H. Wang, and Z. Zhuang, "BM3D vector approximate message passing for radar coded-aperture imaging," in *PIERS-FALL*, 2017, pp. 2035–2038.

[16] X. Wang and S. H. Chan, "Parameter-free plug-and-play ADMM for image restoration," in *Proc. IEEE Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1323–1327.

[17] F. Caltagirone, L. Zdeborová, and F. Krzakala, "On convergence of approximate message passing," in *Proc. IEEE ISIT*, Jul. 2014, pp. 1812–1816.

[18] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE ICASSP*, 2015, pp. 2021–2025.

[19] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," in *Proc. IEEE ISIT*, 2017, pp. 1588–1592.

[20] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learning Res.*, vol. 1, pp. 2177–2204, 2005.

[21] A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Expectation consistent approximate inference: Generalizations and convergence," in *Proc. IEEE ISIT*, 2016, pp. 190–194.

[22] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[23] K. Takeuchi, “Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements,” in *Proc. ISIT*, 2017, pp. 501–505.

[24] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” *arXiv:1610.03082*, 2016.

[25] A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, “Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems,” in *Proc. NIPS*, 2017, pp. 2542–2551.

[26] P. Schniter, A. K. Fletcher, and S. Rangan, “Denoising-based vector AMP,” in *Proc. Intl. Biomedical and Astronomical Signal Process. (BASP) Workshop*, 2017, p. 77.

[27] A. K. Fletcher, P. Pandit, S. Rangan, S. Sarkar, and P. Schniter, “Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis,” *arxiv preprint 1806.10466*, 2018.

[28] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, “Approximate message passing with consistent parameter estimation and applications to sparse learning,” *IEEE Trans. Info. Theory*, vol. 60, no. 5, pp. 2969–2985, Apr. 2014.

[29] A. Taeb, A. Maleki, C. Studer, and R. Baraniuk, “Maximin analysis of message passing algorithms for recovering block sparse signals,” *arXiv preprint arXiv:1303.2389*, 2013.

[30] M. R. Andersen, O. Winther, and L. K. Hansen, “Bayesian inference for structured spike and slab priors,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1745–1753.

[31] S. Rangan, A. K. Fletcher, V. K. Goyal, E. Byrne, and P. Schniter, “Hybrid approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4577–4592, Sept 2017.

[32] L. L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Reading, MA, 1991, vol. 63.

[33] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” in *Advances in Neural Information Processing Systems*, 2012, pp. 341–349.

[34] L. Xu, J. S. Ren, C. Liu, and J. Jia, “Deep convolutional neural network for image deconvolution,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[35] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[36] M. Borgerding, P. Schniter, and S. Rangan, “AMP-inspired deep networks for sparse linear inverse problems,” *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, 2017.

[37] E. J. Candès, T. Strohmer, and V. Voroninski, “PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[38] A. Ahmed, B. Recht, and J. Romberg, “Blind deconvolution using convex programming,” *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[39] S. Ling and T. Strohmer, “Self-calibration and biconvex compressive sensing,” *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[40] M. A. Davenport and J. Romberg, “An overview of low-rank matrix recovery from incomplete observations,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 608–622, 2016.

[41] S. S. Haykin, Ed., *Blind Deconvolution*. Upper Saddle River, NJ: Prentice-Hall, 1994.

[42] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compressive sampling,” *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2002–2016, 2011.

[43] P. Sun, Z. Wang, and P. Schniter, “Joint channel-estimation and equalization of single-carrier systems via bilinear AMP,” *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2772–2785, 2018.

[44] S. Rangan and A. K. Fletcher, “Iterative estimation of constrained rank-one matrices in noise,” in *Proc. IEEE ISIT*, Cambridge, MA, Jul. 2012, pp. 1246–1250.

[45] Y. Deshpande and A. Montanari, “Information-theoretically optimal sparse PCA,” in *Proc. ISIT*, 2014, pp. 2197–2201.

[46] R. Matsushita and T. Tanaka, “Low-rank matrix reconstruction and clustering via approximate message passing,” in *Proc. NIPS*, 2013, pp. 917–925.

[47] T. Lesieur, F. Krzakala, and L. Zdeborova, “Phase transitions in sparse PCA,” in *Proc. IEEE ISIT*, 2015, pp. 1635–1639.

[48] J. Parker and P. Schniter, “Parametric bilinear generalized approximate message passing,” *IEEE J. Sel. Topics Signal Proc.*, vol. 10, no. 4, pp. 795–808, 2016.

[49] E. J. Candes, C. A. Sing-Long, and J. D. Trzasko, “Unbiased risk estimates for singular value thresholding and spectral estimators,” *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4643–4657, 2013.

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

Supplementary Material

A Details on Example Denoisers

In this section, we provide more details on the denoiser examples in Section 3. We also provide conditions under which these denoisers satisfy Assumption 1.

Separable Denoisers. Assume that $\phi(\cdot)$ satisfies a Lipschitz condition,

$$|\phi(r_2,\gamma_2) - \phi(r_1,\gamma_1)| \leq (A + B|\gamma_2 - \gamma_1|)|r_2 - r_1| + C|\gamma_2 - \gamma_1|, \tag{25}$$

for some constants A, B and $C > 0$. Applying the triangle inequality to (25) shows that $\mathbf{g}_1(\cdot)$ satisfies (6). Therefore, $\mathbf{g}_1(\cdot)$ satisfies the condition in Definition 1. Also, the first limit in (7) is given by,

$$\frac{1}{N}\sum_{n=1}^{N} \phi(x_n^0 + z_{1n},\gamma_1)\phi(x_n^0 + z_{2n},\gamma_2) = \mathbb{E}\left[\phi(x_n^0 + z_{1n},\gamma_1)\phi(x_n^0 + z_{2n},\gamma_2)\right],$$

which follows from the Strong Law of Large Numbers and the fact that we have assumed that the components of $\mathbf{x}^0$ are i.i.d. The remaining limits in(7) can be shown to similar converge. In particular,

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}_1,\gamma_1)^{\mathsf{T}}\mathbf{z}_2 = \mathbb{E}\left[\phi(x_n^0 + z_{1n},\gamma_1)z_{2n}\right],$$
$$\lim_{N\to\infty} \langle \nabla \mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}_1,\gamma_1)\rangle = \mathbb{E}\left[\phi'(x_n^0 + z_{1n},\gamma_1)\right],$$

where $\phi'(\cdot)$ is the derivative with respect to the first argument. Moreover, from Stein's lemma,

$$\mathbb{E}\left[\phi(x_n^0 + z_{1n},\gamma_1)z_{2n}\right] = \mathbb{E}\left[\phi'(x_n^0 + z_{1n},\gamma_1)\right]\mathbb{E}\left[z_{1n}z_{2n}\right] = \mathbb{E}\left[\phi'(x_n^0 + z_{1n},\gamma_1)\right]S_{12},$$

which shows that equality of the two limits in (7c). This shows that separable, uniform Lipschitz denoisers $\mathbf{g}_1(\cdot)$ with i.i.d. true signal $\mathbf{x}^0$ satisfy Assumption 1.

Group-based Denoisers. For the groupwise denoiser case, assume that $\boldsymbol{\phi}(\cdot)$ in (8) satisfies a Lipschitz condition,

$$\|\boldsymbol{\phi}(\mathbf{r}_2,\gamma_2) - \phi(\mathbf{r}_1,\gamma_1)\| \leq (A + B|\gamma_2 - \gamma_1|)\|\mathbf{r}_2 - \mathbf{r}_1\| + C|\gamma_2 - \gamma_1|, \tag{26}$$

for constants $A, B, C > 0$ and any $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^K$. Then, it is easily verified that $\mathbf{g}_1(\cdot)$ is uniformly Lipschitz according to Definition 1. To prove that the denoiser satisfies the convergent conditions in Definition 2, let $\mathbf{z}_1$ and $\mathbf{z}_2$ be two sequence of vectors as in Definition 2. Each $\mathbf{z}_i$ can be viewed as a $L \times K$ matrix. We let $\mathbf{z}_{i\ell}$ be the ℓ-th row of $\mathbf{z}_i$. With these definitions, the first sum in (7) is given by,

$$\frac{1}{LK}\sum_{\ell=1}^{L} \boldsymbol{\phi}(\mathbf{x}_\ell^0 + \mathbf{z}_{1\ell})\boldsymbol{\phi}(\mathbf{x}_\ell^0 + \mathbf{z}_{2\ell}),$$

which is a sum of i.i.d. terms. Hence, the sum converges as $L \to \infty$. The convergence of the other sums can be proven similarly.

Convolutional Denoisers. To prove that $\mathbf{g}_1(\mathbf{r}_1)$ in (9) satisfies Assumption 1, first observe that since $\mathbf{g}_1(\mathbf{r})$ is linear. Moreover, since it is realized from a truncated linear filter, its norm is given by,

$$\|\mathbf{g}_1(\mathbf{r})\| \leq A\|\mathbf{r}\|, \quad A := \arg\max_{\theta\in[0,2\pi]} |\widehat{H}(e^{i\theta})|,$$

where $\widehat{H}(e^{i\theta})$ is the discrete-time Fourier transform of the filter $\mathbf{h}$. The bound here holds for all N. Since there is no dependence on γ, the sequence $\mathbf{g}_1(\cdot)$ is uniformly Lipschitz and satisfies Definition 1. To prove that $\mathbf{x}^0$ and $\mathbf{g}_1(\cdot)$ satisfy Definition 2, consider two sequences $\mathbf{z}_1$ and $\mathbf{z}_2$ be as in Definition 2. Let $\mathbf{y}_i$ be the outputs of the convolution (without truncation),

$$\mathbf{y}_i = \mathbf{h} * (\mathbf{x}^0 + \mathbf{z}_i), \quad i = 1, 2.$$

Let $\mathbf{y}$ and $\mathbf{z}$ denote the vector-valued process with components (y_{1n}, y_{2n}) and (z_{1n}, z_{2n}). By assumption, $\mathbf{z}$ is i.i.d. Gaussian. Since $\mathbf{x}^0$ is stationary and ergodic and $\mathbf{z}$ is i.i.d. and $\mathbf{h}$ is a finite length filter, $\mathbf{y}$ is a stationary and ergodic. In addition, $\mathbf{y}$ will have bounded second moments. Now,

$$\mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}_i) = T_N(\mathbf{y}_i),$$

which is the first N samples of $\mathbf{y}_i$. Hence,

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}_1)^{\mathsf{T}}\mathbf{g}_1(\mathbf{x}^0 + \mathbf{z}_2) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} y_{1n}y_{2n}$$

and this limit converges almost surely due to the ergodicity of $\mathbf{y}$. The other limits in Definition 2 can be similarly proven to converge.

Convolutional Neural Networks. As a simple model for a convolutional neural network denoiser, suppose that the true signal, $\mathbf{x}^0$, arises from N time samples of a stationary and ergodic multi-variate process $\mathbf{x}^0$. Let $\mathbf{x}_n^0 \in R^{d_0}$ denote the n-th sample of the process and d_0 denote the dimension of the input. Given an N-sample input $\mathbf{r}$, if we let

$$\mathbf{z}_{\ell+1} = F_\ell(\mathbf{z}_\ell), \quad \mathbf{z}_0 = \mathbf{r},$$

then the $\mathbf{z}_L = \mathbf{g}_1(\mathbf{r})$. Assume that each layer output $\mathbf{z}_\ell$ has N time samples with dimension d_ℓ at each time sample. Also, assume that each layer $F_\ell(\cdot)$ of the denoiser in (10) is one of two possibilities:

- *Convolutional layer:* In this case, the layer mapping $z_{\ell+1} = F_\ell(z_\ell)$ is given by a linear multi-channel convolution,
$$\mathbf{z}_{\ell+1,n} = \sum_{k=0}^{K_\ell - 1} \mathbf{H}_{\ell,k}\mathbf{z}_{\ell,n-k}, \quad n = 0, \ldots, N-1,$$
where $\mathbf{H}_{\ell,k}$ are the matrix coefficients in a convolution kernel. We assume the convolution filter are fixed with finite length.
- *Separable activation:* In this case, the layer mapping is given by
$$\mathbf{z}_{\ell+1} = \phi_\ell(\mathbf{z}_\ell),$$
where $\phi_\ell(\cdot)$ is separable and Lipschitz. This model would include most common activation functions including sigmoids and ReLUs.

Since the convolutional kernels are finite in length, the convolution layers are Lipschitz. In fact, the Lipschitz constant is given by the spectral norm,

$$\|F_\ell(\mathbf{z}_\ell)\| \leq A_\ell \|\mathbf{z}_\ell\|, \quad A_\ell := \max_{\theta \in [0, 2\pi]} \sigma_{\max}(\widehat{\mathbf{H}}_\ell(e^{i\theta})),$$

where $\widehat{\mathbf{H}}_\ell(e^{i\theta})$ is the discrete-time multivariable Fourier transform of the convolution kernel $\mathbf{H}_\ell$ and $\sigma_{\max}(\cdot)$ is the maximum singular value. By assumption, the activation layers are also Lipschitz. Since the composition of Lipschitz functions is Lipschitz, the mapping $\mathbf{g}_1(\cdot)$ in (10) is Lipschitz and satisfies Definition 1.

Also, since $\mathbf{x}^0$ is a multi-variate stationary and ergodic random process, similar arguments as in the convolutional example can be used to show that the limits in (7) hold almost surely. Thus, the $\mathbf{g}_1(\cdot)$ satisfies Assumption 1.

Singular-Value Thresholding (SVT) Denoiser. To show that $\mathbf{g}_1$ in (11) is uniformly pseudo-Lipschitz, we first note that $\mathbf{g}_1$ is the proximal operator of the nuclear norm $\|\cdot\|_*$, i.e.,

$$\mathbf{g}_1(\mathbf{r}, \gamma) = \underset{\mathbf{x} \in \mathbb{R}^{N_1 \times N_2}}{\arg\min} \; \gamma \|\mathbf{x}\|_* + \frac{1}{2}\|\mathbf{x} - \mathbf{r}\|_F^2.$$

From [49], we have that $\mathbf{g}_1$ is non-expansive because the nuclear norm is convex and proper, i.e.,

$$\begin{aligned} &\|\mathbf{g}_1(\mathbf{r}_1, \gamma) - \mathbf{g}_1(\mathbf{r}_2, \gamma)\|_F^2 \leq (\mathbf{r}_1 - \mathbf{r}_2)^{\mathsf{T}}(\mathbf{g}_1(\mathbf{r}_1, \gamma) - \mathbf{g}_1(\mathbf{r}_2, \gamma)) \\ \Rightarrow &\|\mathbf{g}_1(\mathbf{r}_1, \gamma) - \mathbf{g}_1(\mathbf{r}_2, \gamma)\|_F \leq \|\mathbf{r}_1 - \mathbf{r}_2\|_F. \end{aligned} \tag{27}$$

Let the SVD of $\mathbf{r}_2 \in \mathbb{R}^{N_1 \times N_2}$ be $\sum_{i=1}^{\min\{N_1,N_2\}} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$. We can generalize the Lipschitz condition in (27) into

$$\begin{aligned} \|\mathbf{g}_1(\mathbf{r}_1, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_2)\|_F &= \|\mathbf{g}_1(\mathbf{r}_1, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_1) + \mathbf{g}_1(\mathbf{r}_2, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_2)\|_F \\ &\leq \|\mathbf{g}_1(\mathbf{r}_1, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_1)\|_F + \|\mathbf{g}_1(\mathbf{r}_2, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_2)\|_F \\ &\leq \|\mathbf{r}_1 - \mathbf{r}_2\|_F + \|\mathbf{g}_1(\mathbf{r}_2, \gamma_1) - \mathbf{g}_1(\mathbf{r}_2, \gamma_2)\|_F \\ &\overset{\text{(a)}}{=} \|\mathbf{r}_1 - \mathbf{r}_2\|_F + \left\| \sum_{i=1}^{\min\{N_1,N_2\}} ((\sigma_i - \gamma_1)_+ - (\sigma_i - \gamma_2)_+)\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}} \right\|_F \\ &\leq \|\mathbf{r}_1 - \mathbf{r}_2\|_F + \sum_{i=1}^{\min\{N_1,N_2\}} |(\sigma_i - \gamma_1)_+ - (\sigma_i - \gamma_2)_+| \\ &\leq \|\mathbf{r}_1 - \mathbf{r}_2\|_F + \min\{N_1, N_2\}|\gamma_1 - \gamma_2| \\ &\overset{\text{(b)}}{\leq} \|\mathbf{r}_1 - \mathbf{r}_2\|_F + \sqrt{N}|\gamma_1 - \gamma_2|, \end{aligned}$$

where in (a) we have used the the definition of $\mathbf{g}_1$ from (11) and the SVD of $\mathbf{r}_2$, and in (b) we used $\min\{N_1, N_2\} \leq \sqrt{N_1 N_2} = \sqrt{N}$. Next, we show that $\mathbf{g}_1$ also satisfies the convergence conditions in Definition 2. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be two sequences constructed according to Definition 1 and let $\mathbf{x}^0$ be the true signal. Assume that

$$\lim_{N\to\infty} \frac{1}{N}\|\mathbf{x}^0\|_F^2 \text{ and } \lim_{N\to\infty} \frac{1}{N}\mathbf{z}_1^{\mathsf{T}}\mathbf{x}^0 \text{ exist almost surely.} \tag{28}$$

If we write $\mathbf{g}_1(\mathbf{r},\gamma) = [g_1(\mathbf{r},\gamma),\dots,g_N(\mathbf{r},\gamma)]^{\mathsf T}$, then the following series converges because it is bounded:

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}|g_i(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)g_i(\mathbf{x}^0+\mathbf{z}_2,\gamma_2)| &\le \lim_{N\to\infty}\frac{1}{N}\|\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)\|_F\|\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_2,\gamma_2)\|_F \\ &\le \lim_{N\to\infty}\sqrt{\frac{1}{N}\|\mathbf{x}^0+\mathbf{z}_1\|_F^2}\sqrt{\frac{1}{N}\|\mathbf{x}^0+\mathbf{z}_2\|_F^2} \\ &\overset{\text{(a)}}{<} \infty,\end{aligned}$$

where (a) follows from the assumption (28). Since absolute convergence implies convergence, the following series converges:

$$\lim_{N\to\infty}\frac{1}{N}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_2,\gamma_2) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}g_i(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)g_i(\mathbf{x}^0+\mathbf{z}_2,\gamma_2). \qquad (29)$$

If we choose the covariance matrix in Definition 1 to be $\mathbf{S} = \left[\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right]$, then we get

$$\lim_{N\to\infty}\frac{1}{N}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)^{\mathsf T}\mathbf{x}^0 = \lim_{N\to\infty}\frac{1}{N}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_2,0). \qquad (30)$$

Thus, (30) also converges since it is a special case of (29).

It can be easily shown that $\frac{1}{N}\mathbf{z}_2^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)$ is uniformly Lipschitz. Using [8, Lemma 23] and Stein's Lemma [50], we get

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\mathbf{z}_2^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1) &\overset{\text{P}}{\simeq} \lim_{N\to\infty}\frac{1}{N}\mathbb{E}[\mathbf{z}_2^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)] \\ &= \lim_{N\to\infty}\frac{S_{12}}{N}\mathbb{E}[\nabla\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)], \qquad (31)\end{aligned}$$

where $\overset{\text{P}}{\simeq}$ denotes convergence in probability. To show the final convergence condition in Definition 2, let us assume that $\langle\nabla\mathbf{g}_1(\mathbf{r},\gamma)\rangle$ is uniformly Lipschitz. (We are as yet unable to prove this claim.) Then we have $\lim_{N\to\infty}\langle\nabla\mathbf{g}_1(\mathbf{r},\gamma)\rangle \overset{\text{P}}{\simeq} \lim_{N\to\infty}\mathbb{E}[\langle\nabla\mathbf{g}_1(\mathbf{r},\gamma)\rangle]$ using [8, Lemma 23]. Thus, together with (31), we get the desired result

$$\lim_{N\to\infty}\frac{1}{N}\langle\nabla\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1)\rangle = \lim_{N\to\infty}\frac{1}{S_{12}N}\mathbf{z}_2^{\mathsf T}\mathbf{g}_1(\mathbf{x}^0+\mathbf{z}_1,\gamma_1). \qquad (32)$$

B Preliminary Results

Since our proof will follow that of [24], we review a few key results from that work that will be used here as well. The most important provides a characterization of a Haar-distributed matrix $\mathbf{V}$ under linear constraints. A similar result was key to the original analysis of Gaussian matrices in the Bayati-Montanari work [6]. Let $\mathbf{V}\in\mathbb{R}^{N\times N}$ be Haar-distributed and suppose we wish to find the conditional distribution of $\mathbf{V}$ under the event that it satisfies linear constraints

$$\mathbf{A} = \mathbf{V}\mathbf{B}, \qquad (33)$$

for some matrices $\mathbf{A},\mathbf{B}\in\mathbb{R}^{N\times s}$ for some s. Assume $\mathbf{A}$ and $\mathbf{B}$ are full column rank (hence $s\le N$). Let

$$\mathbf{U}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^{\mathsf T}\mathbf{A})^{-1/2}, \quad \mathbf{U}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}^{\mathsf T}\mathbf{B})^{-1/2}. \qquad (34)$$

Also, let $\mathbf{U}_{\mathbf{A}^\perp}$ and $\mathbf{U}_{\mathbf{B}^\perp}$ be any $N\times(N-s)$ matrices whose columns are an orthonormal bases for $\mathrm{Range}(\mathbf{A})^\perp$ and $\mathrm{Range}(\mathbf{B})^\perp$, respectively.

Lemma 1. *[24, Lemma 4] Let $\mathbf{V}\in\mathbb{R}^{N\times N}$ be a random matrix that is Haar distributed. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are deterministic and G is the event that $\mathbf{V}$ satisfies linear constraints* (33)*. Then, the can write $\mathbf{V}$ as,*

$$\mathbf{V} = \mathbf{A}(\mathbf{A}^{\mathsf T}\mathbf{A})^{-1}\mathbf{B}^{\mathsf T} + \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp}^{\mathsf T},$$

where $\widetilde{\mathbf{V}}$ is a Haar distributed matrix independent of G.

Lemma 1 is used in conjunction with the following result.

Lemma 2. *Fix a dimension $s\ge 0$, and suppose that we have sequences $\mathbf{x}=\mathbf{x}(N)$ and $\mathbf{U}=\mathbf{U}(N)$ are sequences such that for each N,*

(i) $\mathbf{U}=\mathbf{U}(N)\in\mathbb{R}^{N\times(N-s)}$ is a random matrix with $\mathbf{U}^{\mathsf T}\mathbf{U}=\mathbf{I}$;

(ii) $\mathbf{x} = \mathbf{x}(N) \in \mathbb{R}^{N-s}$ *a random vector whose mean squared magnitude converges almost surely as*

$$\lim_{N\to\infty} \frac{1}{N}\|\mathbf{x}\|^2 = \tau,$$

for some $\tau > 0$.

(iii) $\mathbf{V} = \mathbf{V}(N) \in \mathbb{R}^{(N-s)\times(N-s)}$ *is a Haar distributed, independent of* $\mathbf{U}$ *and* $\mathbf{x}$.

Then, if we define $\mathbf{y} = \mathbf{U}\mathbf{V}\mathbf{x}$*, we have that the components of* $\mathbf{y}$ *are approximately Gaussian in that,*

$$\mathbf{y} = \widetilde{\mathbf{y}} + \boldsymbol{\eta}, \tag{35}$$

where $\widetilde{\mathbf{y}} \sim \mathcal{N}(0, \tau\mathbf{I})$ *and*

$$\lim_{N\to\infty} \frac{1}{N}\|\boldsymbol{\eta}\|^2 = 0,$$

almost surely.

Proof. This can be proven similar to that of [24, Lemma 5]. □

C A General Convergence Result

Similar to the proof in [19], we prove our main result, Theorem 1, by considering the following more general recursion. We are given a dimension N, an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$, an initial random vector $\mathbf{u}_0 \in \mathbb{R}^N$, along with random vectors $\mathbf{w}^p, \mathbf{w}^q \in \mathbb{R}^N$. Then, we generate a sequence of iterates by the following recursion:

$$\mathbf{p}_k = \mathbf{V}\mathbf{u}_k \tag{36a}$$
$$\alpha_{1k} = \langle \nabla \mathbf{f}_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k}) \rangle, \quad \gamma_{2k} = \Gamma_1(\gamma_{1k}, \alpha_{1k}) \tag{36b}$$
$$\mathbf{v}_k = C_1(\alpha_{1k}) \left[\mathbf{f}_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k}) - \alpha_{1k}\mathbf{p}_k \right] \tag{36c}$$
$$\mathbf{q}_k = \mathbf{V}^{\mathsf{T}}\mathbf{v}_k \tag{36d}$$
$$\alpha_{2k} = \langle \nabla \mathbf{f}_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) \rangle, \quad \gamma_{1,k+1} = \Gamma_2(\gamma_{2k}, \alpha_{2k}) \tag{36e}$$
$$\mathbf{u}_{k+1} = C_2(\alpha_{2k}) \left[\mathbf{f}_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) - \alpha_{2k}\mathbf{q}_k \right], \tag{36f}$$

which is initialized with $\mathbf{u}_0$ and a scalar γ_{10}. We index the recursions by N. We assume that the initial constant and norm of the initial vector converges as

$$\lim_{N\to\infty} \gamma_{10} = \overline{\gamma}_{10}, \quad \lim_{N\to\infty} \frac{1}{N}\|\mathbf{u}_0\|^2 = \tau_{10}, \tag{37}$$

for some constants $\overline{\gamma}_{10}$ and τ_{10}. The matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$ is assumed to be uniformly distributed on the set of orthogonal matrices independent of $\mathbf{u}_0$, $\mathbf{w}^p$ and $\mathbf{w}^q$. For the functions $\mathbf{f}_p(\cdot)$ and $\mathbf{f}_q(\cdot)$ we need a slight generalization of Definitions 1 and 2.

Definition 3. *For each N, suppose that* $\mathbf{u} \in \mathbb{R}^N$ *is a random vector and* $\mathbf{f}(\mathbf{z}, \mathbf{u}, \gamma) \in \mathbb{R}^N$ *is a function on* $\mathbf{z} \in \mathbb{R}^N$*,* $\mathbf{u} \in \mathbb{R}^N$ *and* $\gamma \in \mathbb{R}$*. Let G be some closed, convex set of values* γ*. We say the sequence is* uniformly Lipschitz continuous *if there exists constants A, B and C* > 0*, such that*

$$\limsup_{N\to\infty} \lim_{N\to\infty} \frac{1}{\sqrt{N}} \|\mathbf{f}(\mathbf{z}_2, \mathbf{u}, \gamma_2) - \mathbf{f}(\mathbf{z}_1, \mathbf{u}, \gamma_1)\| \tag{38}$$
$$\leq \limsup_{N\to\infty} \frac{A + B|\gamma_2 - \gamma_1|}{\sqrt{N}} \|\mathbf{z}_2 - \mathbf{z}_1\| + C|\gamma_2 - \gamma_1|, \tag{39}$$

almost surely for any $\mathbf{z}_1, \mathbf{z}_2$ *and* $\gamma_1, \gamma_2 \in G$.

Definition 4. *Let* $\mathbf{u}$*,* $\mathbf{f}(\cdot)$ *and G be as in Definition 3. The sequence* $\mathbf{u}$ *and* $\mathbf{f}(\cdot)$ *are said to be* convergent under Gaussian noise *if the following condition holds: Let* $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^N$ *be two sequences where* (z_{n1}, z_{n2}) *are i.i.d. with* $(z_{n1}, z_{n2}) = \mathcal{N}(0, \mathbf{S})$ *for some positive definite covariance* $\mathbf{S} \in \mathbb{R}^{2\times 2}$*. Then, the following limits exists almost surely,*

$$\mathcal{M}(\mathbf{S}, \gamma_1, \gamma_2) := \lim_{N\to\infty} \mathbf{f}(\mathbf{u}, \mathbf{z}_1, \gamma_1)^{\mathsf{T}} \mathbf{f}(\mathbf{u}, \mathbf{z}_2, \gamma_2) \tag{40}$$

$$\mathcal{A}(S_{11}, \gamma_1) := \lim_{N\to\infty} \langle \nabla \mathbf{f}(\mathbf{u}, \mathbf{z}_1, \gamma_1) \rangle = \frac{1}{NS_{12}} \mathbf{f}(\mathbf{u}, \mathbf{z}_1, \gamma_1)^{\mathsf{T}} \mathbf{z}_2, \tag{41}$$

for all $\gamma_1, \gamma_2 \in G$ *and covariance matrices* $\mathbf{S}$*. Moreover, the functions* $\mathcal{M}(\cdot)$ *and* $\mathcal{A}(\cdot)$ *are continuous in* $\mathbf{S}$*,* γ_1 *and* γ_2.

Our critical assumption is that, following Definitions 3 and 4, the sequence of random vectors $\mathbf{w}^p$ and functions $\mathbf{f}_p(\mathbf{p},\mathbf{w}^p,\gamma_1)$ (as indexed by N) are uniformly Lipschitz continuous and convergent under Gaussian noise for $\gamma_1 \in G_1$ for some closed, convex set G_1. Similarly, the sequence $\mathbf{w}^q$ and function $\mathbf{f}_q(\mathbf{p},\mathbf{w}^q,\gamma_2)$ is also uniformly Lipschitz continuous and convergent under Gaussian noise for $\gamma_2 \in G_2$ for some closed, convex set G_2. In this case, we can define the second moments,

$$\mathcal{M}_p(\tau_1,\gamma_1) := \lim_{N\to\infty} \frac{1}{N}\|\mathbf{f}_p(\mathbf{p},\mathbf{w}^p,\gamma_1)\|^2, \quad \mathbf{p}\sim\mathcal{N}(0,\tau_1\mathbf{I}), \tag{42a}$$

$$\mathcal{M}_q(\tau_2,\gamma_2) := \lim_{N\to\infty} \frac{1}{N}\|\mathbf{f}_q(\mathbf{q},\mathbf{w}^q,\gamma_2)\|^2, \quad \mathbf{q}\sim\mathcal{N}(0,\tau_2\mathbf{I}), \tag{42b}$$

as well as the sensitivity functions,

$$\mathcal{A}_p(\tau_1,\gamma_1) := \lim_{N\to\infty} \langle\nabla\mathbf{f}_p(\mathbf{p},\mathbf{w}^p,\gamma_1)\rangle), \quad \mathbf{p}\sim\mathcal{N}(0,\tau_1\mathbf{I}), \tag{43a}$$

$$\mathcal{A}_q(\tau_2,\gamma_2) := \lim_{N\to\infty} \langle\nabla\mathbf{f}_q(\mathbf{q},\mathbf{w}^q,\gamma_2)\rangle, \quad \mathbf{q}\sim\mathcal{N}(0,\tau_2\mathbf{I}). \tag{43b}$$

The limits exist due to the assumption of $\mathbf{f}_p$ and $\mathbf{f}_q$ being convergent under Gaussian noise. In addition, Definition 4 shows that the sensitivity functions are also given by,

$$\mathcal{A}_p(\tau_1,\gamma_1) == \lim_{N\to\infty} \frac{1}{N\tau_1}\mathbf{p}^{\mathsf{T}}\mathbf{f}_p(\mathbf{p},\mathbf{w}^p,\gamma_1), \quad \mathbf{p}\sim\mathcal{N}(0,\tau_1\mathbf{I}), \tag{44a}$$

$$\mathcal{A}_q(\tau_2,\gamma_2) = \lim_{N\to\infty} \frac{1}{N\tau_2}\mathbf{q}^{\mathsf{T}}\mathbf{f}_q(\mathbf{q},\mathbf{w}^q,\gamma_2) \quad \mathbf{q}\sim\mathcal{N}(0,\tau_2\mathbf{I}). \tag{44b}$$

Under the above assumptions, define the SE equations,

$$\overline{\alpha}_{1k} = \mathcal{A}_p(\tau_{1k},\overline{\gamma}_{1k}) \tag{45a}$$

$$\tau_{2k} = C_1^2(\overline{\alpha}_{1k})\left\{\mathcal{M}_p(\tau_{1k},\overline{\gamma}_{1k}) - \overline{\alpha}_{1k}^2\tau_{1k}\right\} \tag{45b}$$

$$\overline{\gamma}_{2k} = \Gamma_1(\overline{\gamma}_{1k},\overline{\alpha}_{1k}) \tag{45c}$$

$$\overline{\alpha}_{2k} = \mathcal{A}_q(\tau_{2k},\overline{\gamma}_{2k}) \tag{45d}$$

$$\tau_{1,k+1} = C_2^2(\overline{\alpha}_{2k})\left\{\mathcal{M}_p(\tau_{2k},\overline{\gamma}_{2k}) - \overline{\alpha}_{2k}^2\tau_{2k}\right\} \tag{45e}$$

$$\gamma_{1,k+1} = \Gamma_2(\overline{\gamma}_{2k},\overline{\alpha}_{2k}), \tag{45f}$$

which are initialized with $\overline{\gamma}_{10}$ and τ_{10} in (37).

For the sequel, we will use the notation that, if $\mathbf{x}=\mathbf{x}(N)$ and $\mathbf{y}=\mathbf{y}(N)\in\mathbb{R}^N$ are two sequences of random vectors that scale with N,

$$\mathbf{x} = \mathbf{y} + O(\tfrac{1}{\sqrt{N}}) \Longleftrightarrow \lim_{N\to\infty}\frac{1}{N}\|\mathbf{x}-\mathbf{y}\|^2 \text{ almost surely.} \tag{46}$$

With this definition, we have the following result.

Theorem 2. *Consider the recursions* (36) *and SE equations* (45) *under the above assumptions. Assume additionally that, for all k and $i=1,2$, the functions $C_i(\alpha_i)$ and $\Gamma_i(\gamma_i,\alpha_i)$ are continuous at the points $(\gamma_i,\alpha_i)=(\overline{\gamma}_{ik},\overline{\alpha}_{ik})$ from the SE equations. Also, assume that $\overline{\gamma}_{ik}\in G_i$ for all i. Then,*

(a) For each k, we can write $\mathbf{p}_k = \widetilde{\mathbf{p}}_k + O(\frac{1}{\sqrt{N}})$ such that the matrix,

$$\widetilde{\mathbf{P}}_k = [\widetilde{\mathbf{p}}_0,\cdots,\widetilde{\mathbf{p}}_k]\in\mathbb{R}^{N\times k+1}, \tag{47}$$

is independent of $\mathbf{w}^p$ and has i.i.d. rows, $(\widetilde{p}_{n0},\cdots,\widetilde{p}_{nk})$, that are zero mean, $k+1$-dimensional Gaussian random vectors. In addition, we have that

$$E\widetilde{p}_{nk}^2 = \tau_{1k}, \quad \lim_{N\to\infty}(\alpha_{1k},\gamma_{2k}) = (\overline{\alpha}_{1k},\overline{\gamma}_{2k}), \tag{48}$$

where the limit holds almost surely.

(b) For each k, we can write $\mathbf{q}_k = \widetilde{\mathbf{q}}_k + O(\frac{1}{\sqrt{N}})$ such that the matrix,

$$\widetilde{\mathbf{Q}}_k = [\widetilde{\mathbf{q}}_0,\cdots,\widetilde{\mathbf{q}}_k]\in\mathbb{R}^{N\times k+1}, \tag{49}$$

is independent of $\mathbf{w}^q$ and has i.i.d. rows, $(\widetilde{q}_{n0},\cdots,\widetilde{q}_{nk})$, that are zero mean, $k+1$-dimensional Gaussian random vectors. In addition, we have that

$$E\widetilde{q}_{nk}^2 = \tau_{2k}, \quad \lim_{N\to\infty}(\alpha_{2k},\gamma_{1,k+1}) = (\overline{\alpha}_{2k},\overline{\gamma}_{1,k+1}), \tag{50}$$

where the limit holds almost surely.

Proof. We will prove this in the next Appendix, Appendix D. □

D Proof of Theorem 2

D.1 Induction Argument

The proof has a similar structure to the proof of the general convergence result in [24]. So, we will highlight only the key differences. Similar to [24], we use an induction argument. Given iterations $k, \ell \geq 0$, define the hypothesis, $H_{k,\ell}$ as the statement:

- Part (a) of Theorem 2 is true up to k; and
- Part (b) of Theorem 2 is true up to ℓ.

The induction argument will then follow by showing the following three facts:

- $H_{0,-1}$ is true;
- If $H_{k,k-1}$ is true, then so is $H_{k,k}$;
- If $H_{k,k}$ is true, then so is $H_{k+1,k}$.

D.2 Induction Initialization

We first show that the hypothesis $H_{0,-1}$ is true. That is, we must show that the rows of (47) are i.i.d. Gaussians and the limits in (48) hold for $k = 0$. This is a special case of Lemma 2. Specifically, for each N, let $\mathbf{U} = \mathbf{I}_N$, the $N \times N$ identity matrix, which trivially satisfies property (i) of Lemma 2 with $s = 0$. Also, $\mathbf{x} = \mathbf{u}_0$ satisfies property (ii) due to the assumption (37). Then, since $\mathbf{p}_0 = \mathbf{V}\mathbf{u}_0 = \mathbf{U}\mathbf{V}\mathbf{x}$ and $\mathbf{V}$ is Haar distributed independent of $\mathbf{u}_0$, we have that

$$\mathbf{p}_0 = \widetilde{\mathbf{p}}_0 + O(\tfrac{1}{N}), \quad \widetilde{\mathbf{p}}_0 \sim \mathcal{N}(0, \tau_{10}\mathbf{I}). \tag{51}$$

This proves the Gaussianity of the rows of (47) for $k = 0$. Also,

$$\begin{aligned}\lim_{N\to\infty} \alpha_{10} &\overset{(a)}{=} \lim_{N\to\infty} \langle \nabla \mathbf{f}_p(\mathbf{p}_0, \mathbf{w}^p, \gamma_{10}) \rangle \\ &\overset{(b)}{=} \lim_{N\to\infty} \langle \nabla \mathbf{f}_p(\widetilde{\mathbf{p}}_0, \mathbf{w}^p, \overline{\gamma}_{10}) \rangle \overset{(c)}{=} \mathcal{A}_p(\tau_{10}, \overline{\gamma}_{10}) \overset{(d)}{=} \overline{\alpha}_{10},\end{aligned} \tag{52}$$

where (a) follows from (36b); (b) follows from (37), (51) along with the Lipschitz continuity assumption of $\mathbf{f}_p(\cdot)$; (c) follows from the definition (43); and (d) follows from (45a). In addition,

$$\lim_{N\to\infty} \gamma_{10} \overset{(a)}{=} \lim_{N\to\infty} \Gamma_1(\gamma_{10}, \alpha_{10}) \overset{(b)}{=} \Gamma_1(\overline{\gamma}_{10}, \overline{\alpha}_{10}) \overset{(c)}{=} \overline{\gamma}_{20} \tag{53}$$

where (a) follows from (36b); (b) follows from (37), (52) and the continuity of $\Gamma_1(\cdot)$ and (c) follows from (45c). This proves (48).

D.3 The Induction Recursion

We next show the implication $H_{k,k-1} \Rightarrow H_{k,k}$. The implication $H_{k,k} \Rightarrow H_{k+1,k}$ is proven similarly. Hence, fix k and assume that $H_{k,k-1}$ holds. To show $H_{k,k}$, we need to show the Gaussianity of the rows of (49) and that the limits in (50) hold.

First, similar to the proof of (52), we have that

$$\lim_{N\to\infty} \alpha_{2k} = \overline{\alpha}_{2k}. \tag{54}$$

Also, by the induction hypothesis, $\gamma_{2k} \to \overline{\gamma}_{2k}$, and similar to the proof of (53),

$$\lim_{N\to\infty} \gamma_{1,k+1} = \overline{\gamma}_{1,k+1}. \tag{55}$$

This proves (50). We next need to compute various correlations.

Lemma 3. *Under the hypothesis $H_{k,k-1}$, then for any $i, j = 0, \ldots, k$ the following limits exist almost surely,*

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{p}_i^{\mathsf{T}}\mathbf{p}_j, \quad \lim_{N\to\infty} \frac{1}{N}\mathbf{v}_i^{\mathsf{T}}\mathbf{v}_j. \tag{56}$$

Also,

$$\lim_{N\to\infty} \frac{1}{N}\|\mathbf{v}_k\|^2 = \tau_{2k}, \quad \lim_{N\to\infty} \frac{1}{N}\mathbf{v}_i^{\mathsf{T}}\mathbf{p}_j = 0. \tag{57}$$

Proof. For the first part of (56),

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{p}_i^{\mathsf{T}}\mathbf{p}_j \overset{(a)}{=} \lim_{N\to\infty} \frac{1}{N}\widetilde{\mathbf{p}}_i^{\mathsf{T}}\widetilde{\mathbf{p}}_j \overset{(b)}{=} \mathbb{E}(\widetilde{p}_{in}\widetilde{p}_{jn}),$$

where (a) follows due to induction hypothesis that $\mathbf{p}_\ell = \widetilde{\mathbf{p}}_\ell + O(\frac{1}{N})$ for $\ell \leq k$ and (b) follows from the fact that $(\widetilde{p}_{in}, \widetilde{p}_{jn})$ are i.i.d., so the limit occurs almost surely by the Strong Law of Large Numbers. For the second part of (56),

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\mathbf{v}_i^{\mathsf{T}}\mathbf{v}_j \\
&\overset{(a)}{=} \lim_{N\to\infty} \frac{C_1(\alpha_{1i})C_1(\alpha_{1j})}{N}\left[\mathbf{f}_p(\mathbf{p}_i, \mathbf{w}^p, \gamma_{1i}) - \alpha_{1i}\mathbf{p}_i\right]^{\mathsf{T}}\left[\mathbf{f}_p(\mathbf{p}_j, \mathbf{w}^p, \gamma_{1j}) - \alpha_{1j}\mathbf{p}_j\right] \\
&\overset{(b)}{=} \lim_{N\to\infty} \frac{C_1(\overline{\alpha}_{1i})C_1(\overline{\alpha}_{1j})}{N}\left[\mathbf{f}_p(\widetilde{\mathbf{p}}_i, \mathbf{w}^p, \overline{\gamma}_{1i}) - \overline{\alpha}_{1i}\widetilde{\mathbf{p}}_i\right]^{\mathsf{T}}\left[\mathbf{f}_p(\widetilde{\mathbf{p}}_j, \mathbf{w}^p, \overline{\gamma}_{1j}) - \overline{\alpha}_{1j}\widetilde{\mathbf{p}}_j\right], \qquad (58)
\end{aligned}$$

where (a) follows from (36c); (b) follows from the fact that $\mathbf{p}_k = \widetilde{\mathbf{p}}_k + O(\frac{1}{\sqrt{N}})$, (50) and the continuity assumptions of $\mathbf{f}_p(\cdot)$ and $C_1(\cdot)$. We can expand this sum into four terms and use the fact that $\mathbf{f}_p(\cdot)$ is convergent under Gaussian noise to show that all the terms are converge almost surely. Hence, both the limits in (56) exist almost surely.

In the special case when $i = j = k$, we have that,

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N}\|\mathbf{v}_k\|^2 &= \lim_{N\to\infty} \frac{C_1^2(\overline{\alpha}_{1i})}{N}\|\mathbf{f}_p(\widetilde{\mathbf{p}}_k, \mathbf{w}^p, \overline{\gamma}_{ki}) - \overline{\alpha}_{1i}\widetilde{\mathbf{p}}_k\|^2 \\
&= \lim_{N\to\infty} \frac{C_1^2(\overline{\alpha}_{1k})}{N}\left[\|\mathbf{f}_p(\widetilde{\mathbf{p}}_k, \mathbf{w}^p, \overline{\gamma}_{1k})\|^2 - 2\overline{\alpha}_{1k}\widetilde{\mathbf{p}}_k^{\mathsf{T}}\mathbf{f}_p(\widetilde{\mathbf{p}}_k, \mathbf{w}^p, \overline{\gamma}_{1k}) + \overline{\alpha}_{1k}^2\|\widetilde{\mathbf{p}}_k\|^2\right] \\
&\overset{(a)}{=} C_1^2(\overline{\alpha}_{1k})\left(\mathcal{M}_p(\tau_{1k}, \overline{\gamma}_{1k}) - 2\overline{\alpha}_{1k}^2\tau_{1k} + \overline{\alpha}_{1k}^2\tau_{1k}\right) \\
&= C_1^2(\overline{\alpha}_{1k})\left(\mathcal{M}_p(\tau_{1k}, \overline{\gamma}_{1k}) - \overline{\alpha}_{1k}^2\tau_{1k}\right) \overset{(b)}{=} \tau_{2k}, \qquad (59)
\end{aligned}$$

where (a) follows from the limits in (42) and (44); and (b) follows from (45b). This proves the first relation in (57). For the second relation,

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N}\mathbf{v}_i^{\mathsf{T}}\mathbf{p}_j &\overset{(a)}{=} \lim_{N\to\infty} \frac{C_1(\alpha_{1i})}{N}\left(\mathbf{f}_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k}) - \alpha_{1i}\mathbf{p}_i\right)^{\mathsf{T}}\mathbf{p}_j \\
&\overset{(b)}{=} \lim_{N\to\infty} \frac{C_1(\overline{\alpha}_{1i})}{N}\left(\mathbf{f}_p(\widetilde{\mathbf{p}}_k, \mathbf{w}^p, \overline{\gamma}_{1k}) - \overline{\alpha}_{1i}\widetilde{\mathbf{p}}_i\right)^{\mathsf{T}}\widetilde{\mathbf{p}}_j \\
&\overset{(c)}{=} (\mathcal{A}_p(\tau_{1i}, \overline{\gamma}_{1i}) - \overline{\alpha}_{1i})\mathrm{cov}(\widetilde{p}_{ni}, \widetilde{p}_{jn}) \overset{(d)}{=} 0, \qquad (60)
\end{aligned}$$

where (a) follows from (36c); (b) follows from the fact that $\mathbf{p}_k = \widetilde{\mathbf{p}}_k + O(\frac{1}{\sqrt{N}})$; (c) follows from the assumption that $\mathbf{f}_p(\cdot)$ is convergent under Gaussian noise as given in Definition 2; and (d) follows from (45a). □

The remainder of the proof now follows a very similar structure to that in [24]. First, let

$$\mathbf{U}_k := [\mathbf{u}_0 \cdots \mathbf{u}_k] \in \mathbb{R}^{N\times(k+1)},$$

represent the first $k+1$ values of the vector $\mathbf{u}_\ell$. Define the matrices $\mathbf{V}_k$, $\mathbf{Q}_k$ and $\mathbf{P}_k$ similarly. Let G_k be the set of random vectors,

$$G_k := \{\mathbf{U}_k, \mathbf{P}_k, \mathbf{V}_k, \mathbf{Q}_{k-1}\}. \qquad (61)$$

With some abuse of notation, we will also use G_k to denote the sigma-algebra generated by these variables. The set (61) contains all the outputs of the algorithm (36) immediately *before* (36d) in iteration k.

Now, the actions of the matrix $\mathbf{V}$ in the recursions (36) are through the matrix-vector multiplications (36a) and (36d). Hence, if we define the matrices,

$$\mathbf{A}_k := [\mathbf{P}_k\ \mathbf{V}_{k-1}], \quad \mathbf{B}_k := [\mathbf{U}_k\ \mathbf{Q}_{k-1}], \qquad (62)$$

the output of the recursions in the set G_k will be unchanged for all matrices $\mathbf{V}$ satisfying the linear constraints

$$\mathbf{A}_k = \mathbf{V}\mathbf{B}_k. \qquad (63)$$

Hence, the conditional distribution of $\mathbf{V}$ given G_k is precisely the uniform distribution on the set of orthogonal matrices satisfying (63). The matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ are of dimensions $N \times s$ where $s = 2k+1$. From Lemma 1,

$$\mathbf{V} = \mathbf{A}_k(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{B}_k^{\mathsf{T}} + \mathbf{U}_{\mathbf{A}_k^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}_k^\perp}^{\mathsf{T}}, \qquad (64)$$

where $\mathbf{U}_{\mathbf{A}_k^\perp}$ and $\mathbf{U}_{\mathbf{B}_k^\perp}$ are $N \times (N-s)$ matrices whose columns are an orthonormal basis for $\mathrm{Range}(\mathbf{A}_k)^\perp$ and $\mathrm{Range}(\mathbf{B}_k)^\perp$. The matrix $\widetilde{\mathbf{V}}$ is Haar distributed on the set of $(N-s) \times (N-s)$ orthogonal matrices and independent of G_k.

Next, similar to the proof in [24], we use (64) to write $\mathbf{q}_k$ in (36d) as a sum of two terms

$$\mathbf{q}_k = \mathbf{V}^{\mathsf{T}}\mathbf{v}_k = \mathbf{q}_k^{\mathrm{det}} + \mathbf{q}_k^{\mathrm{ran}}, \tag{65}$$

where $\mathbf{q}_k^{\mathrm{det}}$ is what we will call the *deterministic* part:

$$\mathbf{q}_k^{\mathrm{det}} = \mathbf{B}_k(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{A}_k^{\mathsf{T}}\mathbf{v}_k, \tag{66}$$

and $\mathbf{q}_k^{\mathrm{ran}}$ is what we will call the *random* part:

$$\mathbf{q}_k^{\mathrm{ran}} = \mathbf{U}_{\mathbf{B}_k^\perp}\widetilde{\mathbf{V}}^{\mathsf{T}}\mathbf{U}_{\mathbf{A}_k^\perp}^{\mathsf{T}}\mathbf{v}_k. \tag{67}$$

The next two lemmas evaluate the asymptotic distributions of the two terms in (65) and are similar to those in the proof in [24].

Lemma 4. *Under the induction hypothesis $H_{k,k-1}$, there exists constants $\beta_{k,0}, \ldots, \beta_{k,k-1}$ such that*

$$\mathbf{q}_k^{\mathrm{det}} = \beta_{k0}\widetilde{\mathbf{q}}_0 + \cdots + \beta_{k,k-1}\widetilde{\mathbf{q}}_{k-1} + O(\tfrac{1}{\sqrt{N}}). \tag{68}$$

Proof. From Lemma 3, these exists almost surely. We evaluate the asymptotic values of various terms in (66). Using the definition of $\mathbf{A}_k$ in (62),

$$\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k = \begin{bmatrix} \mathbf{P}_k^{\mathsf{T}}\mathbf{P}_k & \mathbf{P}_k^{\mathsf{T}}\mathbf{V}_{k-1} \\ \mathbf{V}_{k-1}^{\mathsf{T}}\mathbf{P}_k & \mathbf{V}_{k-1}^{\mathsf{T}}\mathbf{V}_{k-1} \end{bmatrix}$$

For $i, j \leq k$, define

$$Q_{ij}^p := \lim_{N\to\infty} \frac{1}{N}\mathbf{p}_i^{\mathsf{T}}\mathbf{p}_j, \quad Q_{ij}^v := \lim_{N\to\infty} \frac{1}{N}\mathbf{v}_i^{\mathsf{T}}\mathbf{v}_j.$$

From Lemma 3, these limits exists almost surely. Let $\mathbf{Q}^p$ be the matrix with components Q_{ij}^p for $i, j \leq k$ and let $\mathbf{Q}^v$ be the matrix with components Q_{ij}^v for $i, j < k$. Then, since $\mathbf{p}_i$ and $\mathbf{p}_j$ are the i-th and j-th column of $\mathbf{P}_k$, the (i,j)-th component of the matrix $\mathbf{P}_k^{\mathsf{T}}\mathbf{P}_k$ is given by

$$\lim_{N\to\infty} \frac{1}{N}\left[\mathbf{P}_k^{\mathsf{T}}\mathbf{P}_k\right]_{ij} = \lim_{N\to\infty} \frac{1}{N}\mathbf{p}_i^{\mathsf{T}}\mathbf{p}_j = Q_{ij}^p.$$

Similarly,

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{V}_{k-1}^{\mathsf{T}}\mathbf{V}_{k-1} = \mathbf{Q}^v$$

almost surely. Also, from Lemma 3,

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{P}_k^{\mathsf{T}}\mathbf{V}_{k-1} = 0,$$

almost surely. The above calculations show that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k = \begin{bmatrix} \mathbf{Q}^p & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^v \end{bmatrix}. \tag{69}$$

A similar calculation shows that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{A}_k^{\mathsf{T}}\mathbf{v}_k = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}^v \end{bmatrix}, \tag{70}$$

where $\mathbf{b}^v$ is the vector of correlations

$$\mathbf{b}^v = \begin{bmatrix} Q_{0k}^v \; Q_{1k}^v \; \cdots \; Q_{k-1,k}^v \end{bmatrix}^{\mathsf{T}}. \tag{71}$$

Combining (69) and (70) shows that

$$\lim_{N\to\infty} (\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{A}_k^{\mathsf{T}}\mathbf{v}_k = \begin{bmatrix} \mathbf{0} \\ \beta_k \end{bmatrix}, \tag{72}$$

where

$$\beta_k := [\mathbf{Q}^v]^{-1}\mathbf{b}^v.$$

Therefore,

$$\mathbf{q}_k^{\mathrm{det}} = \mathbf{B}_k(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{A}_k^{\mathsf{T}}\mathbf{v}_k = [\mathbf{U}_k \; \mathbf{Q}_{k-1}] \begin{bmatrix} \mathbf{0} \\ \beta_k \end{bmatrix} + O(\tfrac{1}{\sqrt{N}}) = \sum_{\ell=0}^{k-1} \beta_{k\ell}\widetilde{\mathbf{q}}_\ell + O(\tfrac{1}{\sqrt{N}}). \tag{73}$$

This completes the proof of the lemma. □

Lemma 5. *Under the induction hypothesis $H_{k,k\text{-}1}$, the following limit holds almost surely*

$$\lim_{N\to\infty} \frac{1}{N} \|\mathbf{U}^{\mathsf{T}}_{\mathbf{A}_k^{\perp}} \mathbf{v}_k\|^2 = \rho_k, \tag{74}$$

for some constant $\rho_k \geq 0$.

Proof. From (62), the matrix $\mathbf{A}_k$ has $s = 2k+1$ columns. From Lemma 1, $\mathbf{U}_{\mathbf{A}_k^{\perp}}$ is an orthonormal basis of $N-s$ in the $\mathrm{Range}(\mathbf{A}_k)^{\perp}$. Hence, the energy $\|\mathbf{U}_{\mathbf{A}_k^{\perp}} \mathbf{v}_k\|^2$ is precisely

$$\|\mathbf{U}_{\mathbf{A}_k^{\perp}} \mathbf{s}_k\|^2 = \mathbf{v}_k^{\mathsf{T}} \mathbf{v}_k - \mathbf{v}_k^{\mathsf{T}} \mathbf{A}_k (\mathbf{A}_k^{\mathsf{T}} \mathbf{A}_k)^{-1} \mathbf{A}_k^{\mathsf{T}} \mathbf{v}_k.$$

Using similar calculations as the previous lemma, we have

$$\lim_{N\to\infty} \frac{1}{N} \|\mathbf{U}_{\mathbf{A}_k} \mathbf{s}_k\|^2 = \tau_{2k} - (\mathbf{b}^v)^{\mathsf{T}} \left[\mathbf{Q}^v\right]^{-1} \mathbf{b}^v.$$

Hence, the lemma is proven if we define ρ_k as the right hand side of this equation. □

Lemma 6. *Under the induction hypothesis $H_{k,k\text{-}1}$, the "random" part $\mathbf{q}_k^{\mathrm{ran}}$ is given by,*

$$\mathbf{q}_k^{\mathrm{ran}} = \mathbf{u}_k + O(\tfrac{1}{\sqrt{N}}), \tag{75}$$

where $\mathbf{u}_k$ is an i.i.d. zero mean Gaussian random vector independent of $\mathbf{w}^p$ and $\widetilde{\mathbf{q}}_j$, $j = 0, \ldots, k-1$.

Proof. This is a direct application of Lemma 2. Let $\mathbf{x} = \mathbf{U}^{\mathsf{T}}_{\mathbf{A}_k^{\perp}} \mathbf{v}_k$ so that

$$\mathbf{q}_k^{\mathrm{det}} = \mathbf{U}_{\mathbf{B}_k^{\perp}} \mathbf{V}^{\mathsf{T}} \mathbf{x}_k.$$

For each N, $\mathbf{U}_{\mathbf{B}_k^{\perp}} \in \mathbb{R}^{N \times (N-s)}$ is a matrix with orthonormal columns spanning $\mathrm{Range}(\mathbf{B}_k)^{\perp}$. Also, since $\widetilde{\mathbf{V}}$ is uniformly distributed on the set of $(N-s) \times (N-s)$ orthogonal matrices, and independent of G_k, it is independent of $\mathbf{x}$. Lemma 5 also shows that

$$\lim_{N\to\infty} \frac{1}{N} \|\mathbf{x}\|^2 = \rho_k,$$

almost surely. The limit (75) now follows from Lemma 2. □

Using the partition (65) and Lemmas 4 and 6, we have that

$$\mathbf{q}_k = \widetilde{\mathbf{q}}_k + O(\tfrac{1}{\sqrt{N}}), \quad \widetilde{\mathbf{q}}_k := \beta_{k0} \widetilde{\mathbf{q}}_0 + \cdots + \beta_{k,k\text{-}1} \widetilde{\mathbf{q}}_{k\text{-}1} + \mathbf{u}.$$

Now, by the induction by hypothesis, the matrix $\widetilde{\mathbf{Q}}_{k\text{-}1}$ are have i.i.d. rows that are jointly Gaussian. The matrix $\widetilde{\mathbf{Q}}_k$ is formed by adding the column $\widetilde{\mathbf{q}}_k$ to $\widetilde{\mathbf{Q}}_{k\text{-}1}$. Since $\mathbf{u}$ is Gaussian i.i.d. independent of $\widetilde{\mathbf{q}}_j$ for $j < k$, we have that the matrix $\widetilde{\mathbf{Q}}_k$ will have i.i.d. rows that are jointly Gaussian.

It remains to show all the limits in (50). First,

$$\begin{aligned} E[\widetilde{q}_{nk}^2] &\overset{(a)}{=} \lim_{N\to\infty} \frac{1}{N} \|\widetilde{\mathbf{q}}_k\|^2 \\ &\overset{(b)}{=} \lim_{N\to\infty} \frac{1}{N} \|\mathbf{q}_k\|^2 \overset{(d)}{=} \lim_{N\to\infty} \frac{1}{N} \|\mathbf{v}_k\|^2 \overset{(d)}{=} \tau_{2k}, \end{aligned}$$

where (a) follows from the Strong Law of Large Numbers and the fact that the components of $\widetilde{\mathbf{q}}_k$ are i.i.d.; (b) follows from the fact that $\mathbf{q}_k = \widetilde{\mathbf{q}}_k + O(\frac{1}{\sqrt{N}})$; (c) follows from (36d) and the fact that $\mathbf{V}$ is orthogonal; and (d) follows from Lemma 3. Now the function $\Gamma_1(\gamma_1, \alpha_1)$ is assumed to be continuous at $(\overline{\gamma}_{1k}, \overline{\alpha}_{1k})$. Also, the induction hypothesis assumes that $\alpha_{1k} \to \overline{\alpha}_{1k}$ and $\gamma_{1k} \to \overline{\gamma}_{1k}$ almost surely. Hence,

$$\lim_{N\to\infty} \gamma_{2k} = \lim_{N\to\infty} \Gamma_1(\gamma_{1k}, \alpha_{1k}) = \overline{\gamma}_{2k}. \tag{76}$$

In addition,

$$\begin{aligned} \lim_{N\to\infty} \alpha_{2k} &\overset{(a)}{=} \lim_{N\to\infty} \langle \nabla \mathbf{f}_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) \rangle \\ &\overset{(b)}{=} \lim_{N\to\infty} \langle \nabla \mathbf{f}_q(\widetilde{\mathbf{q}}_k, \mathbf{w}^q, \overline{\gamma}_{2k}) \rangle \overset{(c)}{=} \mathcal{A}_q(\tau_{2k}, \overline{\gamma}_{2k}) \overset{(d)}{=} \overline{\alpha}_{2k}, \end{aligned} \tag{77}$$

where (a) follows from (36e); (b) follows from the Lipschitz continuity assumptions of $\mathbf{f}_q(\cdot)$; (c) follows from (43) and (d) follows from (45d). The limits (76) and (77) prove (50). This completes the induction argument and the proof of the theorem.

E Proof of Theorem 1

The proof is virtually identical to that used in [24]. Specifically, we show that Theorem 1 is a special case of Theorem 2. As in [24], we need to simply rewrite the recursions in Algorithm 1 in the form (36) by defining the error terms

$$\mathbf{p}_k := \mathbf{r}_{1k} - \mathbf{x}^0, \quad \mathbf{v}_k := \mathbf{r}_{2k} - \mathbf{x}^0, \tag{78}$$

and their transforms,

$$\mathbf{u}_k := \mathbf{V}^{\mathsf{T}}\mathbf{p}_k, \quad \mathbf{q}_k := \mathbf{V}^{\mathsf{T}}\mathbf{v}_k. \tag{79}$$

Also, define the disturbance terms

$$\mathbf{w}^q := (\boldsymbol{\xi}, \mathbf{s}), \quad \mathbf{w}^p := \mathbf{x}^0, \quad \boldsymbol{\xi} := \mathbf{U}^{\mathsf{T}}\mathbf{w}. \tag{80}$$

Also, define the update functions,

$$\mathbf{f}_q(\mathbf{q}, (\boldsymbol{\xi}, \mathbf{s}), \gamma_2) := \frac{\gamma_w \mathbf{s}\boldsymbol{\xi} + \gamma_2 \mathbf{q}}{\gamma_w \mathbf{s}^2 + \gamma_2}, \tag{81a}$$

$$\mathbf{f}_p(\mathbf{p}, \mathbf{x}^0, \gamma_1) := \mathbf{g}_1(\mathbf{p} + \mathbf{x}^0, \gamma_1) - \mathbf{x}^0. \tag{81b}$$

In the definition of the function $\mathbf{f}_q(\cdot)$, the product $\mathbf{s}\boldsymbol{\xi}$ and the division are to be taken componentwise. Also, let

$$C_i(\alpha_i) := \frac{1}{1-\alpha_i}, \quad \Gamma_i(\gamma_i, \alpha_i) := \gamma_i \left[\frac{1}{\alpha_i} - 1\right].$$

Then, it is shown in [24] that the recursions in Algorithm 1 exactly match (36).

So, all we need to do is show that the update functions in (81) satisfy Definitions 3 and 4. These conditions are proven in the next two lemmas. By the assumption of Theorem 1, $\overline{\gamma}_{2k} > 0$ for all k. So, for any finite k, there exists a lower bound $\gamma_{2,min} > 0$ such that $\overline{\gamma}_{2\ell} \geq \gamma_{2,min}$ for all $\ell \leq k$. Let $G_2 = \{\gamma_2 | \gamma_2 \geq \gamma_{2,min}\}$.

Lemma 7. *The sequence of random vectors* $\mathbf{w}^q$ *in* (80)*, functions* $\mathbf{f}_q(\cdot)$ *in* (81a) *satisfy Definitions 3 and 4 for* $\gamma_2 \in G_2$.

Proof. First note that the function $\mathbf{f}_q(\cdot)$ in (81a) is separable meaning that its n-th output is given by,

$$[\mathbf{f}_q(\mathbf{q}, (\boldsymbol{\xi}, \mathbf{s}), \gamma_2)]_n = \phi(q, s, \xi, \gamma_2) := \frac{\gamma_w s\xi + \gamma_2 q}{\gamma_w s + \gamma_2}. \tag{82}$$

For any $\gamma_2 \in G_2$, we can bound the partial derivatives,

$$\begin{aligned}
\left|\frac{\partial\phi(q,s,\xi,\gamma_2)}{\partial q}\right| &= \left|\frac{\gamma_2}{\gamma_w s + \gamma_2}\right| \leq 1, \\
\left|\frac{\partial\phi(q,s,\xi,\gamma_2)}{\partial \gamma_2}\right| &= \left|\frac{q(\gamma_w s + \gamma_2) - \gamma_w s\xi - \gamma_2 q}{(\gamma_w s + \gamma_2)^2}\right| \\
&\leq [|q| + |\xi|]\frac{\gamma_w s}{(\gamma_w s + \gamma_2)^2} \leq [|q| + |\xi|]\frac{1}{\gamma_{2,min}^2}.
\end{aligned}$$

Therefore, if we let $A = 1$, $B = C = 1/\gamma_{2,min}^2$, we get that,

$$|\phi(q_2, s, \xi, \gamma_{22}) - \phi(q_1, s, \xi, \gamma_{21})| \leq (A + B|\gamma_{22} - \gamma_{21}|)|q_2 - q_1| + C|\xi||\gamma_{22} - \gamma_{21}|,$$

for and q_1, q_2 and $\gamma_{21}, \gamma_{22} \in G_2$. This implies that for any vectors $\mathbf{q}_1, \mathbf{q}_2$,

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\|\mathbf{f}_q(\mathbf{q}_2, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{22}) - \mathbf{f}_q(\mathbf{q}_1, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{21})\| \\
&\qquad \leq \frac{(A + B|\gamma_{22} - \gamma_{21}|)}{\sqrt{N}}\|\mathbf{q}_2 - \mathbf{q}_1\| + C\frac{\|\boldsymbol{\xi}\|}{\sqrt{N}}|\gamma_{22} - \gamma_{21}|.
\end{aligned}$$

Since $\boldsymbol{\xi} := \mathbf{U}^{\mathsf{T}}\mathbf{w}$ and $\mathbf{U}$ is orthogonal, $\|\boldsymbol{\xi}\| = \|\mathbf{w}\|$. Also, since $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_w)$,

$$\lim_{N\to\infty}\frac{1}{N}\|\boldsymbol{\xi}\|^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{w}\|^2 = \frac{1}{\gamma_w},$$

almost surely. Therefore,

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{\sqrt{N}}\|\mathbf{f}_q(\mathbf{q}_2, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{22}) - \mathbf{f}_q(\mathbf{q}_1, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{21})\| \\
&\qquad \leq \limsup_{N\to\infty}\frac{(A + B|\gamma_{22} - \gamma_{21}|)}{\sqrt{N}}\|\mathbf{q}_2 - \mathbf{q}_1\| + \frac{C}{\sqrt{\gamma_w}}|\gamma_{22} - \gamma_{21}|,
\end{aligned}$$

which proves that $\mathbf{f}_q(\cdot)$ satisfies the uniform Lipschitz condition in Definition 3.

We turn to the convergence properties in Definition 4. For each N, let $\mathbf{q}_1, \mathbf{q}_2$ be vectors with components (q_{1n}, q_{2n}) that are i.i.d. and Gaussian $(q_{1n}, q_{2n}) \sim \mathcal{N}(0, \mathbf{S})$ for some positive definite covariance matrix $\mathbf{S}$. Let $\gamma_{21}, \gamma_{22} > 0$. Since $\boldsymbol{\xi} := \mathbf{U}^{\mathsf{T}}\mathbf{w}$, $\mathbf{U}$ is orthogonal, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_w)$, we have that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_w)$. Hence, the components of $\boldsymbol{\xi}$ are i.i.d. Also, by assumption, $\mathbf{s}$ has i.i.d. components, independent of $\boldsymbol{\xi}$. Therefore,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\mathbf{f}_q(\mathbf{q}_2, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{22})^{\mathsf{T}}\mathbf{f}_q(\mathbf{q}_1, (\boldsymbol{\xi}, \mathbf{s}), \gamma_{21}) \\
&\quad \overset{(a)}{=} \lim_{N\to\infty} \frac{1}{N}\phi(q_{2n}, \xi_n, s_n, \gamma_{22})\phi(q_{1n}, \xi_n, s_n, \gamma_{21}) \\
&\quad \overset{(b)}{=} \lim_{N\to\infty} \mathbb{E}\left[\phi(q_{2n}, \xi_n, s_n, \gamma_{22})\phi(q_{1n}, \xi_n, s_n, \gamma_{21})\right],
\end{aligned}$$

where (a) follows from the separability of $\mathbf{f}_q(\cdot)$ in (82) and (b) follows from the fact that terms are i.i.d., so we can apply the Strong Law of Large Numbers. The convergence of the limit is almost sure. This proves (40). The limit (41) can be proven similarly. Hence, the sequences $\mathbf{w}^q$ and $\mathbf{f}_q(\cdot)$ satisfy Definition 4. □

Next, consider $\mathbf{f}_p(\cdot)$ in (81b).

Lemma 8. *The sequence of random vectors* $\mathbf{w}^p$ *in* (80)*, functions* $\mathbf{f}_p(\cdot)$ *in* (81b) *satisfy Definitions 3 and 4.*

Proof. For any vectors $\mathbf{p}_1$, $\mathbf{p}_2$ and γ_1, γ_2,

$$\begin{aligned}
\|\mathbf{f}_p(\mathbf{p}_2, \mathbf{x}^0, \gamma_2) - \mathbf{f}_p(\mathbf{p}_1, \mathbf{x}^0, \gamma_1)\| &= \|\mathbf{g}_1(\mathbf{p}_2 + \mathbf{x}^0, \gamma_2) - \mathbf{g}_1(\mathbf{p}_1 + \mathbf{x}^0, \gamma_1)\| \\
&\le (A + B|\gamma_2 - \gamma_1|)\|\mathbf{p}_2 - \mathbf{p}_1\|^2 + \sqrt{N}C|\gamma_2 - \gamma_1|,
\end{aligned}$$

where the last step follows from the fact that $\mathbf{g}_1(\cdot)$ is uniformly Lipschitz continuous as per Definition 1. This shows that $\mathbf{f}_p(\cdot)$ satisfies the uniform Lipschitz continuity assumption in Definition 3.

Now suppose that $\mathbf{p}_1, \mathbf{p}_2$ are Gaussian vectors such that the components, (p_{1n}, p_{2n}) are i.i.d. with $(p_{1n}, p_{2n}) \sim \mathcal{N}(0, \mathbf{S})$. Then,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\mathbf{f}_p(\mathbf{p}_1, \mathbf{x}^0, \gamma_1)^{\mathsf{T}}\mathbf{f}_p(\mathbf{p}_2, \mathbf{x}^0, \gamma_2) = \\
&\quad = \lim_{N\to\infty} \frac{1}{N}\left[\mathbf{g}_1(\mathbf{p}_1 + \mathbf{x}^0, \gamma_1)^{\mathsf{T}}\mathbf{g}_1(\mathbf{p}_2 + \mathbf{x}^0, \gamma_2) - 2(\mathbf{x}^0)^{\mathsf{T}}\mathbf{g}_1(\mathbf{p}_1 + \mathbf{x}^0, \gamma_1) + \|\mathbf{x}^0\|^2\right].
\end{aligned}$$

All three terms on the right-hand side of this equation converge due to the assumption that the limits in (7) converge. Moreover, the limits are continuous in $\mathbf{S}$, γ_1 and γ_2. The convergence of (41) can be proven similarly. Hence, the sequences $\mathbf{w}^p$ and $\mathbf{f}_p(\cdot)$ satisfy Definition 4. □

Lemmas 7 and 8 show that the vectors $\mathbf{w}^q$ and $\mathbf{w}^p$ and functions $\mathbf{f}_q(\cdot)$ and $\mathbf{f}_p(\cdot)$ satisfy the necessary conditions of Theorem 2, which completes the proof of Theorem 1.

F Example Image Recoveries

Figure 5 shows the original images and examples of recovered images for various algorithms after 12 iterations under sampling rate $M/N = 0.3$, cond$(\mathbf{A}) = 1$, and no noise. There we see that the quality of DnCNN-based recovery far exceeds that of LASSO. The figure also shows that, in all cases, LASSO-VAMP outperformed LASSO-AMP and that in all but one case DnCNN-VAMP outperformed DnCNN-AMP.

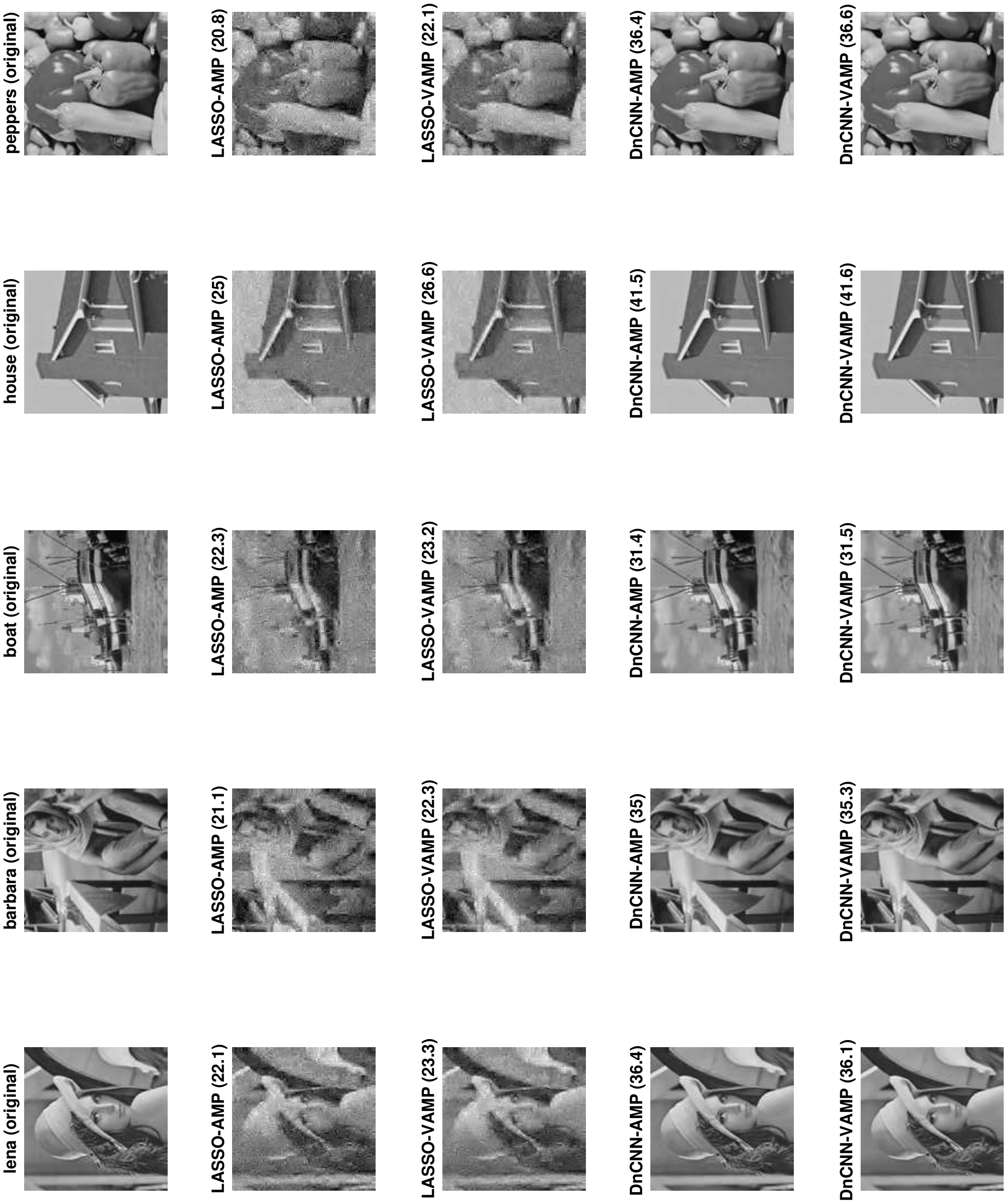


Figure 5: Compressive image recovery at $M/N = 0.3$: Original and recovered images (with PSNR)