Vector Approximate Message Passing

Sundeep Rangan, *Fellow, IEEE*, Philip Schniter, *Fellow, IEEE*, and Alyson K. Fletcher, *Member, IEEE*.

***Abstract*—The standard linear regression (SLR) problem is to recover a vector $\mathbf{x}^0$ from noisy linear observations $\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w}$. The approximate message passing (AMP) algorithm recently proposed by Donoho, Maleki, and Montanari is a computationally efficient iterative approach to SLR that has a remarkable property: for large i.i.d. sub-Gaussian matrices $\mathbf{A}$, its per-iteration behavior is rigorously characterized by a scalar state-evolution whose fixed points, when unique, are Bayes optimal. AMP, however, is fragile in that even small deviations from the i.i.d. sub-Gaussian model can cause the algorithm to diverge. This paper considers a "vector AMP" (VAMP) algorithm and shows that VAMP has a rigorous scalar state-evolution that holds under a much broader class of large random matrices $\mathbf{A}$: those that are right-rotationally invariant. After performing an initial singular value decomposition (SVD) of $\mathbf{A}$, the per-iteration complexity of VAMP can be made similar to that of AMP. In addition, the fixed points of VAMP's state evolution are consistent with the replica prediction of the minimum mean-squared error recently derived by Tulino, Caire, Verdú, and Shamai. The effectiveness and state evolution predictions of VAMP are confirmed in numerical experiments.**

***Index Terms*—Belief propagation, message passing, variational inference, random matrices.**

I. Introduction

Consider the problem of recovering a vector $\mathbf{x}^0 \in \mathbb{R}^N$ from noisy linear measurements of the form

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

where $\mathbf{A}$ is a known matrix and $\mathbf{w}$ is an unknown, unstructured noise vector. In the statistics literature, this problem is known as *standard linear regression*, and in the signal processing literature this is known as *solving a linear inverse problem*, or as *compressive sensing* when $M \ll N$ and $\mathbf{x}^0$ is sparse.

A. Problem Formulations

One approach to recovering $\mathbf{x}^0$ is *regularized quadratic loss minimization*, where an estimate $\widehat{\mathbf{x}}$ of $\mathbf{x}^0$ is computed by solving an optimization problem of the form

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + f(\mathbf{x}). \tag{2}$$

Here, the penalty function or "regularization" $f(\mathbf{x})$ is chosen to promote a desired structure in $\widehat{\mathbf{x}}$. For example, the choice $f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ with $\lambda > 0$ promotes sparsity in $\widehat{\mathbf{x}}$.

Another approach is through the Bayesian methodology. Here, one presumes a prior density $p(\mathbf{x})$ and likelihood function $p(\mathbf{y}|\mathbf{x})$ and then aims to compute the posterior density

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}} \tag{3}$$

or, in practice, a summary of it [1]. Example summaries include the maximum *a posteriori* (MAP) estimate

$$\widehat{\mathbf{x}}_{\mathsf{MAP}} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}), \tag{4}$$

the minimum mean-squared error (MMSE) estimate

$$\widehat{\mathbf{x}}_{\mathsf{MMSE}} = \arg\min_{\widetilde{\mathbf{x}}} \int \|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 p(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x} = \mathbb{E}[\mathbf{x}|\mathbf{y}], \tag{5}$$

or the posterior marginal densities $\{p(x_n|\mathbf{y})\}_{n=1}^N$.

Note that, if the noise $\mathbf{w}$ is modeled as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1}\mathbf{I})$, i.e., additive white Gaussian noise (AWGN) with some precision $\gamma_w > 0$, then the regularized quadratic loss minimization problem (2) is equivalent to MAP estimation under the prior $p(\mathbf{x}) \propto \exp[-\gamma_w f(\mathbf{x})]$, where $\propto$ denotes equality up to a scaling that is independent of $\mathbf{x}$. Thus we focus on MAP, MMSE, and marginal posterior inference in the sequel.

B. Approximate Message Passing

Recently, the so-called *approximate message passing* (AMP) algorithm [2], [3] was proposed as an iterative method to recover $\mathbf{x}^0$ from measurements of the form (1). The AMP iterations are specified in Algorithm 1. There,[1] $\mathbf{g}_1(\cdot,\gamma_k) : \mathbb{R}^N \to \mathbb{R}^N$ is a *denoising* function parameterized by γ_k, and $\langle \mathbf{g}_1'(\mathbf{r}_k,\gamma_k)\rangle$ is its *divergence* at $\mathbf{r}_k$. In particular, $\mathbf{g}_1'(\mathbf{r}_k,\gamma_k) \in \mathbb{R}^N$ is the diagonal of the Jacobian,

$$\mathbf{g}_1'(\mathbf{r}_k,\gamma_k) = \mathrm{diag}\left[\frac{\partial \mathbf{g}_1(\mathbf{r}_k,\gamma_k)}{\partial \mathbf{r}_k}\right], \tag{6}$$

and $\langle\cdot\rangle$ is the empirical averaging operation

$$\langle \mathbf{u}\rangle := \frac{1}{N}\sum_{n=1}^N u_n. \tag{7}$$

When $\mathbf{A}$ is a large i.i.d. sub-Gaussian matrix, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_{w0}^{-1}\mathbf{I})$, and $\mathbf{g}(\cdot,\gamma_k)$ is *separable*, i.e.,

$$[\mathbf{g}_1(\mathbf{r}_k,\gamma_k)]_n = g_1(r_{kn},\gamma_k)\ \forall n, \tag{8}$$

with identical Lipschitz components $g_1(\cdot,\gamma_k) : \mathbb{R} \to \mathbb{R}$, AMP displays a remarkable behavior, which is that $\mathbf{r}_k$ behaves like

S. Rangan (email: srangan@nyu.edu) is with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY. His work was supported by the National Science Foundation under Grants 1302336, 1564142, and 1547332.

P. Schniter (email: schniter.1@osu.edu) is with the Department of Electrical and Computer Engineering, The Ohio State University. His work was supported in part by the National Science Foundation grant CCF-1527162.

A. K. Fletcher (email: akfletcher@ucla.edu) is with the Departments of Statistics, Mathematics, and Electrical Engineering, University of California, Los Angeles. Her work is supported in part by National Science Foundation grants 1254204 and 1564278 as well as the Office of Naval Research grant N00014-15-1-2677.

[1]The subscript "1" in $\mathbf{g}_1$ is used promote notational consistency with Vector AMP algorithm presented in the sequel.

Algorithm 1 AMP

Require: Matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, measurement vector $\mathbf{y}$, denoiser $\mathbf{g}_1(\cdot, \gamma_k)$, and number of iterations $K_{\rm it}$.

Set $\mathbf{v}_{-1} = \mathbf{0}$ and select initial $\mathbf{r}_0, \gamma_0$.
for $k = 0, 1, \ldots, K_{\rm it}$ **do**
  $\widehat{\mathbf{x}}_k = \mathbf{g}_1(\mathbf{r}_k, \gamma_k)$
  $\alpha_k = \langle \mathbf{g}_1'(\mathbf{r}_k, \gamma_k) \rangle$
  $\mathbf{v}_k = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}_k + \frac{N}{M}\alpha_{k-1}\mathbf{v}_{k-1}$
  $\mathbf{r}_{k+1} = \widehat{\mathbf{x}}_k + \mathbf{A}^{\mathsf{T}}\mathbf{v}_k$
  Select γ_{k+1}
end for
Return $\widehat{\mathbf{x}}_{K_{\rm it}}$.

a white-Gaussian-noise corrupted version of the true signal $\mathbf{x}^0$ [2]. That is,

$$\mathbf{r}_k = \mathbf{x}^0 + \mathcal{N}(\mathbf{0}, \tau_k \mathbf{I}), \tag{9}$$

for some variance $\tau_k > 0$. Moreover, the variance τ_k can be predicted through the following *state evolution* (SE):

$$\mathcal{E}(\gamma_k, \tau_k) = \frac{1}{N}\mathbb{E}\left[\left\|\mathbf{g}_1\big(\mathbf{x}^0 + \mathcal{N}(\mathbf{0}, \tau_k\mathbf{I}), \gamma_k\big) - \mathbf{x}^0\right\|^2\right] \tag{10a}$$

$$\tau_{k+1} = \gamma_{w0}^{-1} + \frac{N}{M}\mathcal{E}(\gamma_k, \tau_k), \tag{10b}$$

where $\mathcal{E}(\gamma_k, \tau_k)$ is the MSE on the AMP estimate $\widehat{\mathbf{x}}_k$.

The AMP SE (10) was rigorously established for i.i.d. Gaussian $\mathbf{A}$ in [4] and for i.i.d. sub-Gaussian $\mathbf{A}$ in [5] in the *large-system limit* (i.e., $N, M \to \infty$ and $N/M \to \delta \in (0, 1)$) under some mild regularity conditions. Because the SE (10) holds for generic $g_1(\cdot, \gamma_k)$ and generic γ_k-update rules, it can be used to characterize the application of AMP to many problems, as further discussed in Section II-A.

C. Limitations, Modifications, and Alternatives to AMP

An important limitation of AMP's SE is that it holds only under large i.i.d. sub-Gaussian $\mathbf{A}$. Although recent analysis [6] has rigorously analyzed AMP's performance under finite-sized i.i.d. Gaussian $\mathbf{A}$, there remains the important question of how AMP behaves with other $\mathbf{A}$.

Unfortunately, it turns out that the AMP Algorithm 1 is somewhat fragile with regard to the construction of $\mathbf{A}$. For example, AMP diverges with even mildly ill-conditioned or non-zero-mean $\mathbf{A}$ [7]–[9]. Although damping [7], [9], mean-removal [9], sequential updating [10], and direct free-energy minimization [11] all help to prevent AMP from diverging, such strategies are limited in effectiveness.

Many other algorithms for standard linear regression (1) have been designed using approximations of belief propagation (BP) and/or free-energy minimization. Among these are the ADATAP [12], EP [13], EC [14]–[16], S-AMP [17], [18], and OAMP [19] approaches. Although numerical simulation suggests that some of these algorithms are more robust than AMP Algorithm 1 to the choice of $\mathbf{A}$, their performance and convergence has not been rigorously analyzed. In particular, there remains the question of whether there exists an AMP-like algorithm with a rigorous SE analysis that holds for a larger class of matrices than i.i.d. sub-Gaussian. In the sequel, we describe one such algorithm.

D. Contributions

In this paper, we propose a computationally efficient iterative algorithm for the estimation of the vector $\mathbf{x}^0$ from noisy linear measurements $\mathbf{y}$ of the form in (1). (See Algorithm 2.) We call the algorithm "*vector AMP*" (VAMP) because i) its behavior can be rigorously characterized by a scalar SE under large random $\mathbf{A}$, and ii) it can be derived using an approximation of BP on a factor graph with vector-valued variable nodes. We outline VAMP's derivation in Section III with the aid of some background material that is reviewed in Section II.

In Section IV, we establish the VAMP SE in the case of large *right-orthogonally invariant* random $\mathbf{A}$ and separable Lipschitz denoisers $\mathbf{g}_1(\cdot, \gamma_k)$, using techniques similar to those used by Bayati and Montanari in [4]. Importantly, these right-orthogonally invariant $\mathbf{A}$ allow arbitrary singular values and arbitrary left singular vectors, making VAMP much more robust than AMP in regards to the construction of $\mathbf{A}$. In Section V, we establish that the asymptotic MSE predicted by VAMP's SE agrees with the MMSE predicted by the replica method [20] when VAMP's priors are matched to the true data. Finally, in Section VI, we present numerical experiments demonstrating that VAMP's empirical behavior matches its SE at moderate dimensions, even when $\mathbf{A}$ is highly ill-conditioned or non-zero-mean.

E. Relation to Existing Work

The idea to construct algorithms from graphical models with vector-valued nodes is not new, and in fact underlies the EC- and EP-based algorithms described in [13]–[16]. The use of vector-valued nodes is also central to the derivation of S-AMP [17], [18]. In the sequel, we present a simple derivation of VAMP that uses an EP-like approach. But we note that VAMP can also be derived from the "diagonal restricted" [14, App. D] or "uniform diagonalization" [16] incarnations of EC, although those derivations are much more involved.

It was recently shown [15] that, for large right-orthogonally invariant $\mathbf{A}$, the fixed points of diagonal-restricted EC are "good" in the sense that they are consistent with a certain replica prediction of the MMSE that is derived in [15]. Since the fixed points of ADATAP and S-AMP are known [17] to coincide with those of diagonal-restricted EC (and thus VAMP), all of these algorithms can be understood to have good fixed points. The trouble is that these algorithms do not necessarily converge to their fixed points. For example, S-AMP diverges with even mildly ill-conditioned or non-zero-mean $\mathbf{A}$, as demonstrated in Section VI. *Our main contribution is establishing that VAMP's behavior can be exactly predicted by an SE analysis analogous to that for AMP. This SE analysis then provides precise convergence guarantees for large right-orthogonally invariant* $\mathbf{A}$. The numerical results presented in Section VI confirm that, in practice, VAMP's convergence is remarkably robust, even with very ill-conditioned or mean-perturbed matrices $\mathbf{A}$ of finite dimension.

The main insight that leads to both the VAMP algorithm and its SE analysis comes from a consideration of the singular value decomposition (SVD) of $\mathbf{A}$. Specifically, take the "economy" SVD,

$$\mathbf{A} = \overline{\mathbf{U}}\mathrm{Diag}(\overline{\mathbf{s}})\overline{\mathbf{V}}^{\mathsf{T}}, \tag{11}$$

where $\overline{\mathbf{s}} \in \mathbb{R}^R$ for $R := \mathrm{rank}(\mathbf{A}) \leq \min(M, N)$. The VAMP iterations can be performed by matrix-vector multiplications with $\overline{\mathbf{V}} \in \mathbb{R}^{N\times R}$ and $\overline{\mathbf{V}}^{\mathsf{T}}$, yielding a structure very similar to that of AMP. Computationally, the SVD form of VAMP (i.e., Algorithm 2) has the benefit that, once the SVD has been computed, VAMP's per-iteration cost will be dominated by $O(RN)$ floating-point operations (flops), as opposed to $O(N^3)$ for the EC methods from [14, App. D] or [16]. Furthermore, if these matrix-vector multiplications have fast implementations (e.g., $O(N)$ when $\overline{\mathbf{V}}$ is a discrete wavelet transform), then the complexity of VAMP reduces accordingly. We emphasize that VAMP uses a single SVD, not a per-iteration SVD. In many applications, this SVD can be computed off-line. In the case that SVD complexity may be an issue, we note that it costs $O(MNR)$ flops by classical methods or $O(MN\log K)$ by modern approaches [21].

The SVD offers more than just a fast algorithmic implementation. More importantly, it connects VAMP to AMP in such a way that the Bayati and Montanari's SE analysis of AMP [4] can be extended to obtain a rigorous SE for VAMP. In this way, the SVD can be viewed as a proof technique. Since it will be useful for derivation/interpretation in the sequel, we note that the VAMP iterations can also be written without an explicit SVD (see Algorithm 3), in which case they coincide with the uniform-diagonalization variant of the generalized EC method from [16]. In this latter implementation, the linear MMSE (LMMSE) estimate (24) must be computed at each iteration, as well as the trace of its covariance matrix (25), which both involve the inverse of an $N \times N$ matrix.

The OAMP-LMMSE algorithm from [19] is similar to VAMP and diagonal-restricted EC, but different in that it approximates certain variance terms. This difference can be seen by comparing equations (30)-(31) in [19] to lines 8 and 10 in Algorithm 2 (or lines 14 and 7 in Algorithm 3). Furthermore, OAMP-LMMSE differs from VAMP in its reliance on matrix inversion (see, e.g., the comments in the Conclusion of [19]).

F. Notation

We use capital boldface letters like $\mathbf{A}$ for matrices, small boldface letters like $\mathbf{a}$ for vectors, $(\cdot)^{\mathsf{T}}$ for transposition, and $a_n = [\mathbf{a}]_n$ to denote the nth element of $\mathbf{a}$. Also, we use $\|\mathbf{a}\|_p = (\sum_n |a_n|^p)^{1/p}$ for the ℓ_p norm of $\mathbf{a}$, $\|\mathbf{A}\|_2$ for the spectral norm of $\mathbf{A}$, $\mathrm{Diag}(\mathbf{a})$ for the diagonal matrix created from vector $\mathbf{a}$, and $\mathrm{diag}(\mathbf{A})$ for the vector extracted from the diagonal of matrix $\mathbf{A}$. Likewise, we use $\mathbf{I}_N$ for the $N \times N$ identity matrix, $\mathbf{0}$ for the matrix of all zeros, and $\mathbf{1}$ for the matrix of all ones. For a random vector $\mathbf{x}$, we denote its probability density function (pdf) by $p(\mathbf{x})$, its expectation by $\mathbb{E}[\mathbf{x}]$, and its covariance matrix by $\mathrm{Cov}[\mathbf{x}]$. Similarly, we use $p(\mathbf{x}|\mathbf{y})$, $\mathbb{E}[\mathbf{x}|\mathbf{y}]$, and $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$ for the *conditional* pdf, expectation, and covariance, respectively. Also, we use $\mathbb{E}[\mathbf{x}|b]$ and $\mathrm{Cov}[\mathbf{x}|b]$ to denote the expectation and covariance of $\mathbf{x} \sim b(\mathbf{x})$, i.e., $\mathbf{x}$ distributed according to the pdf $b(\mathbf{x})$. We refer to the Dirac delta pdf using $\delta(\mathbf{x})$ and to the pdf of a Gaussian random vector $\mathbf{x} \in \mathbb{R}^N$ with mean $\mathbf{a}$ and covariance $\mathbf{C}$ using $\mathcal{N}(\mathbf{x}; \mathbf{a}, \mathbf{C}) = \exp(-(\mathbf{x}-\mathbf{a})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{a})/2)/\sqrt{(2\pi)^N|\mathbf{C}|}$. Finally, $p(\mathbf{x}) \propto f(\mathbf{x})$ says that functions $p(\cdot)$ and $f(\cdot)$ are equal up to a scaling that is invariant to $\mathbf{x}$.

II. Background on the AMP Algorithm

In this section, we provide background on the AMP algorithm that will be useful in the sequel.

A. Applications to Bayesian Inference

We first detail the application of the AMP Algorithm 1 to the Bayesian inference problems from Section I-A. Suppose that the prior on $\mathbf{x}$ is i.i.d., so that it takes the form

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n). \tag{12}$$

Then AMP can be applied to MAP problem (4) by choosing the scalar denoiser as

$$g_1(r_{kn}, \gamma_k) = \underset{x_n\in\mathbb{R}}{\arg\min} \left[\frac{\gamma_k}{2}|x_n - r_{kn}|^2 - \ln p(x_n)\right]. \tag{13}$$

Likewise, AMP can be applied to the MMSE problem (5) by choosing

$$g_1(r_{kn}, \gamma_k) = \mathbb{E}[x_n|r_{kn}, \gamma_k], \tag{14}$$

where the expectation in (14) is with respect to the conditional density

$$p(x_n|r_{kn}, \gamma_k) \propto \exp\left[-\frac{\gamma_k}{2}|r_{kn} - x_n|^2 + \ln p(x_n)\right]. \tag{15}$$

In addition, $p(x_n|r_{kn}, \gamma_k)$ in (15) acts as AMP's iteration-k approximation of the marginal posterior $p(x_n|\mathbf{y})$. For later use, we note that the derivative of the MMSE scalar denoiser (14) w.r.t. its first argument can be expressed as

$$g_1'(r_{kn}, \gamma_k) = \gamma_k \mathrm{var}\left[x_n|r_{kn}, \gamma_k\right], \tag{16}$$

where the variance is computed with respect to the density (15) (see, e.g., [22]).

In (13)-(15), γ_k can be interpreted as an estimate of τ_k^{-1}, the iteration-k precision of $\mathbf{r}_k$ from (9). In the case that τ_k is known, the "matched" assignment

$$\gamma_k = \tau_k^{-1} \tag{17}$$

leads to the interpretation of (13) and (14) as the scalar MAP and MMSE denoisers of r_{kn}, respectively. Since, in practice, τ_k is usually not known, it has been suggested to use

$$\gamma_{k+1} = \frac{M}{\|\mathbf{v}_k\|^2}, \tag{18}$$

although other choices are possible [23].

B. Relation of AMP to IST

The AMP Algorithm 1 is closely related to the well-known *iterative soft thresholding* (IST) algorithm [24], [25] that can be used[2] to solve (2) with convex $f(\cdot)$. In particular, if the term

$$\frac{N}{M}\alpha_{k-1}\mathbf{v}_{k-1} \tag{19}$$

is removed from line 5 of Algorithm 1, then what remains is the IST algorithm.

The term (19) is known as the *Onsager* term in the statistical physics literature [26]. Under large i.i.d. sub-Gaussian $\mathbf{A}$, the Onsager correction ensures the behavior in (9). When (9) holds, the denoiser $g_1(\cdot,\gamma_k)$ can be optimized accordingly, in which case each iteration of AMP becomes very productive. As a result, AMP converges much faster than ISTA for i.i.d. Gaussian $\mathbf{A}$ (see, e.g., [23] for a comparison).

C. Derivations of AMP

The AMP algorithm can be derived in several ways. One way is through approximations of loopy belief propagation (BP) [27], [28] on a bipartite factor graph constructed from the factorization

$$p(\mathbf{y},\mathbf{x}) = \left[\prod_{m=1}^{M}\mathcal{N}(y_m;\mathbf{a}_m^{\mathsf{T}}\mathbf{x},\gamma_w^{-1})\right]\left[\prod_{n=1}^{N}p(x_n)\right], \tag{20}$$

where $\mathbf{a}_m^{\mathsf{T}}$ denotes the mth row of $\mathbf{A}$. We refer the reader to [3], [22] for details on the message-passing derivation of AMP, noting connections to the general framework of *expectation propagation* (EP) [29], [30]. AMP can also be derived through a "free-energy" approach, where one i) proposes a cost function, ii) derives conditions on its stationary points, and iii) constructs an algorithm whose fixed points coincide with those stationary points. We refer the reader to [17], [31], [32] for details, and note connections to the general framework of *expectation consistent approximation* (EC) [14], [16].

III. The Vector AMP Algorithm

The *Vector AMP* (VAMP) algorithm is stated in Algorithm 2. In line 9, "$\overline{\mathbf{s}}^2$" refers to the componentwise square of vector $\overline{\mathbf{s}}$. Also, $\mathrm{Diag}(\mathbf{a})$ denotes the diagonal matrix whose diagonal components are given by the vector $\mathbf{a}$.

A. Relation of VAMP to AMP

A visual examination of VAMP Algorithm 2 shows many similarities with AMP Algorithm 1. In particular, the denoising and divergence steps in lines 5-6 of Algorithm 2 are identical to those in lines 3-4 of Algorithm 1. Likewise, an Onsager term $\alpha_k\mathbf{r}_k$ is visible in line 7 of Algorithm 2, analogous to the one in line 5 of Algorithm 1. Finally, the per-iteration computational complexity of each algorithm is dominated by two matrix-vector multiplications: those involving $\mathbf{A}$ and $\mathbf{A}^{\mathsf{T}}$ in Algorithm 1 and those involving $\overline{\mathbf{V}}$ and $\overline{\mathbf{V}}^{\mathsf{T}}$ in Algorithm 2.

The most important similarity between the AMP and VAMP algorithms is not obvious from visual inspection and will be established rigorously in the sequel. It is the following: for certain large random $\mathbf{A}$, the VAMP quantity $\mathbf{r}_k$ behaves like a white-Gaussian-noise corrupted version of the true signal $\mathbf{x}^0$, i.e.,

$$\mathbf{r}_k = \mathbf{x}^0 + \mathcal{N}(\mathbf{0},\tau_k\mathbf{I}), \tag{21}$$

for some variance $\tau_k > 0$. Moreover, the noise variance τ_k can be tracked through a scalar SE formalism whose details will be provided in the sequel. Furthermore, the VAMP quantity γ_k can be interpreted as an estimate of τ_k^{-1} in (21), analogous to the AMP quantity γ_k discussed around (17).

It should be emphasized that the class of matrices $\mathbf{A}$ under which the VAMP SE holds is much bigger than the class under which the AMP SE holds. In particular, VAMP's SE holds for large random matrices $\mathbf{A}$ whose right singular[3] vector matrix $\mathbf{V}\in\mathbb{R}^{N\times N}$ is uniformly distributed on the group of orthogonal matrices. Notably, VAMP's SE holds for *arbitrary* (i.e., deterministic) left singular vector matrices $\mathbf{U}$ and singular values, apart from some mild regularity conditions that will be detailed in the sequel. In contrast, AMP's SE is known to hold [4], [5] only for large i.i.d. sub-Gaussian matrices $\mathbf{A}$, which implies i) random orthogonal $\mathbf{U}$ and $\mathbf{V}$ and ii) a particular distribution on the singular values of $\mathbf{A}$.

Algorithm 2 Vector AMP (SVD Form)

Require: Matrix $\mathbf{A}\in\mathbb{R}^{M\times N}$; measurements $\mathbf{y}\in\mathbb{R}^M$; denoiser $\mathbf{g}_1(\cdot,\gamma_k)$; assumed noise precision $\gamma_w\geq 0$; and number of iterations $K_{\rm it}$.

1: Compute economy SVD $\overline{\mathbf{U}}\mathrm{Diag}(\overline{\mathbf{s}})\overline{\mathbf{V}}^{\mathsf{T}} = \mathbf{A}\in\mathbb{R}^{M\times N}$ with $\overline{\mathbf{U}}^{\mathsf{T}}\overline{\mathbf{U}}=\mathbf{I}_R$, $\overline{\mathbf{V}}^{\mathsf{T}}\overline{\mathbf{V}}=\mathbf{I}_R$, $\overline{\mathbf{s}}\in\mathbb{R}_+^R$, $R=\mathrm{rank}(\mathbf{A})$.
2: Compute preconditioned $\widetilde{\mathbf{y}} := \mathrm{Diag}(\overline{\mathbf{s}})^{-1}\overline{\mathbf{U}}^{\mathsf{T}}\mathbf{y}$
3: Select initial $\mathbf{r}_0$ and $\gamma_0\geq 0$.
4: **for** $k=0,1,\ldots,K_{\rm it}$ **do**
5: $\quad\widehat{\mathbf{x}}_k = \mathbf{g}_1(\mathbf{r}_k,\gamma_k)$
6: $\quad\alpha_k = \langle\mathbf{g}_1'(\mathbf{r}_k,\gamma_k)\rangle$
7: $\quad\widetilde{\mathbf{r}}_k = (\widehat{\mathbf{x}}_k-\alpha_k\mathbf{r}_k)/(1-\alpha_k)$
8: $\quad\widetilde{\gamma}_k = \gamma_k(1-\alpha_k)/\alpha_k$
9: $\quad\mathbf{d}_k = \gamma_w\mathrm{Diag}\big(\gamma_w\overline{\mathbf{s}}^2+\widetilde{\gamma}_k\mathbf{1}\big)^{-1}\overline{\mathbf{s}}^2$
10: $\quad\gamma_{k+1} = \widetilde{\gamma}_k R\langle\mathbf{d}_k\rangle/(N-R\langle\mathbf{d}_k\rangle)$
11: $\quad\mathbf{r}_{k+1} = \widetilde{\mathbf{r}}_k + \frac{N}{R}\overline{\mathbf{V}}\mathrm{Diag}\big(\mathbf{d}_k/\langle\mathbf{d}_k\rangle\big)\big(\widetilde{\mathbf{y}}-\overline{\mathbf{V}}^{\mathsf{T}}\widetilde{\mathbf{r}}_k\big)$
12: **end for**
13: Return $\widehat{\mathbf{x}}_{K_{\rm it}}$.

B. EP Derivation of VAMP

As with AMP Algorithm 1, in the VAMP algorithm in Algorithm 2 can be derived in many ways. Here we present a very simple derivation based on an EP-like approximation of the sum-product (SP) belief-propagation algorithm. Unlike the AMP algorithm, whose message-passing derivation uses

[2]The IST algorithm is guaranteed to converge [25] when $\|\mathbf{A}\|_2<1$.

[3]We use several forms of SVD in this paper. Algorithm 2 uses the "economy" SVD $\mathbf{A}=\overline{\mathbf{U}}\mathrm{Diag}(\overline{\mathbf{s}})\overline{\mathbf{V}}^{\mathsf{T}}\in\mathbb{R}^{M\times N}$, where $\overline{\mathbf{s}}\in\mathbb{R}_+^R$ with $R=\mathrm{rank}(\mathbf{A})$, so that $\overline{\mathbf{U}}$ and/or $\overline{\mathbf{V}}$ may be tall. The discussion in Section III-A uses the "standard" SVD $\mathbf{A}=\mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$, where $\mathbf{S}\in\mathbb{R}^{M\times N}$ and both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. Finally, the state-evolution proof in Section IV uses the standard SVD on square $\mathbf{A}\in\mathbb{R}^{N\times N}$.

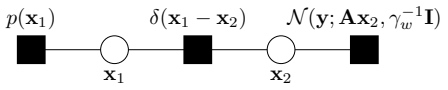

Fig. 1. The factor graph used for the derivation of VAMP. The circles represent variable nodes and the squares represent factor nodes from (23).

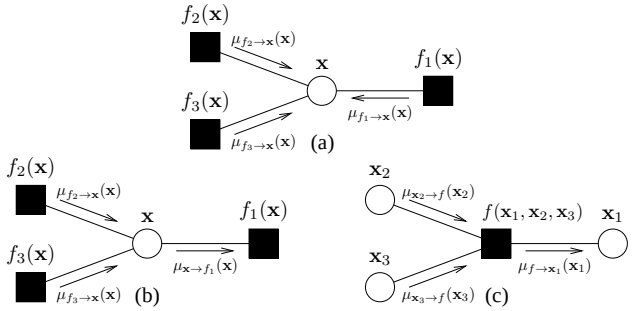

Fig. 2. Factor graphs to illustrate (a) messaging through a factor node and (b) messaging through a variable node.

a loopy factor graph with *scalar*-valued nodes, the VAMP algorithm uses a non-loopy graph with *vector*-valued nodes, hence the name "vector AMP." We note that VAMP can also be derived using the "diagonal restricted" or "uniform diagonalization" EC approach [14], [16], but that derivation is much more complicated.

To derive VAMP, we start with the factorization

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x})\mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{x}, \gamma_w^{-1}\mathbf{I}), \tag{22}$$

and split $\mathbf{x}$ into two identical variables $\mathbf{x}_1 = \mathbf{x}_2$, giving an equivalent factorization

$$p(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1)\delta(\mathbf{x}_1 - \mathbf{x}_2)\mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{x}_2, \gamma_w^{-1}\mathbf{I}), \tag{23}$$

where $\delta(\cdot)$ is the Dirac delta distribution. The factor graph corresponding to (23) is shown in Figure 1. We then pass messages on this factor graph according to the following rules.

1) *Approximate beliefs:* The approximate belief $b_{\mathsf{app}}(\mathbf{x})$ on variable node $\mathbf{x}$ is $\mathcal{N}(\mathbf{x}; \widehat{\mathbf{x}}, \eta^{-1}\mathbf{I})$, where $\widehat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|b_{\mathsf{sp}}]$ and $\eta^{-1} = \langle \mathrm{diag}(\mathrm{Cov}[\mathbf{x}|b_{\mathsf{sp}}])\rangle$ are the mean and average variance of the corresponding SP belief $b_{\mathsf{sp}}(\mathbf{x}) \propto \prod_i \mu_{f_i\to\mathbf{x}}(\mathbf{x})$, i.e., the normalized product of all messages impinging on the node. See Figure 2(a) for an illustration.
2) *Variable-to-factor messages:* The message from a variable node $\mathbf{x}$ to a connected factor node f_i is $\mu_{\mathbf{x}\to f_i}(\mathbf{x}) \propto b_{\mathsf{app}}(\mathbf{x})/\mu_{f_i\to\mathbf{x}}(\mathbf{x})$, i.e., the ratio of the most recent approximate belief $b_{\mathsf{app}}(\mathbf{x})$ to the most recent message from f_i to $\mathbf{x}$. See Figure 2(b) for an illustration.
3) *Factor-to-variable messages:* The message from a factor node f to a connected variable node $\mathbf{x}_i$ is $\mu_{f\to\mathbf{x}_i}(\mathbf{x}_i) \propto \int f(\mathbf{x}_i, \{\mathbf{x}_j\}_{j\neq i}) \prod_{j\neq i} \mu_{\mathbf{x}_j\to f}(\mathbf{x}_j)\, \mathrm{d}\mathbf{x}_j$. See Figure 2(c) for an illustration.

By applying the above message-passing rules to the factor graph in Figure 1, one obtains Algorithm 3. (See Appendix A for a detailed derivation.) Lines 11–12 of Algorithm 3 use

$$\mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k}) := \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_{2k}\mathbf{I}\right)^{-1} \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{y} + \gamma_{2k}\mathbf{r}_{2k}\right), \tag{24}$$

which can be recognized as the MMSE estimate of a random vector $\mathbf{x}_2$ under likelihood $\mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{x}_2, \gamma_w^{-1}\mathbf{I})$ and prior $\mathbf{x}_2 \sim \mathcal{N}(\mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$. Since this estimate is linear in $\mathbf{r}_2$, we will refer to it as the "LMMSE" estimator. From (6)-(7) and (24), it follows that line 12 of Algorithm 3 uses

$$\langle \mathbf{g}_2'(\mathbf{r}_{2k}, \gamma_{2k})\rangle = \frac{\gamma_{2k}}{N}\mathrm{Tr}\left[\left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_2\mathbf{I}\right)^{-1}\right]. \tag{25}$$

Algorithm 3 is merely a restatement of VAMP Algorithm 2. Their equivalence can then be seen by substituting the "economy" SVD $\mathbf{A} = \overline{\mathbf{U}}\mathrm{Diag}(\overline{\mathbf{s}})\overline{\mathbf{V}}^\mathsf{T}$ into Algorithm 3, simplifying, and equating $\widehat{\mathbf{x}}_k \equiv \widehat{\mathbf{x}}_{1k}$, $\mathbf{r}_k \equiv \mathbf{r}_{1k}$, $\gamma_k \equiv \gamma_{1k}$, $\widetilde{\gamma}_k \equiv \gamma_{2k}$, and $\alpha_k \equiv \alpha_{1k}$.

As presented in Algorithm 3, the steps of VAMP exhibit an elegant symmetry. The first half of the steps perform denoising on $\mathbf{r}_{1k}$ and then Onsager correction in $\mathbf{r}_{2k}$, while the second half of the steps perform LMMSE estimation $\mathbf{r}_{2k}$ and Onsager correction in $\mathbf{r}_{1,k+1}$.

Algorithm 3 Vector AMP (LMMSE form)

Require: LMMSE estimator $\mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k})$ from (24), denoiser $\mathbf{g}_1(\cdot, \gamma_{1k})$, and number of iterations K_{it}.

1: Select initial $\mathbf{r}_{10}$ and $\gamma_{10} \geq 0$.
2: **for** $k = 0, 1, \ldots, K_{\mathrm{it}}$ **do**
3: // Denoising
4: $\widehat{\mathbf{x}}_{1k} = \mathbf{g}_1(\mathbf{r}_{1k}, \gamma_{1k})$
5: $\alpha_{1k} = \langle \mathbf{g}_1'(\mathbf{r}_{1k}, \gamma_{1k})\rangle$
6: $\eta_{1k} = \gamma_{1k}/\alpha_{1k}$
7: $\gamma_{2k} = \eta_{1k} - \gamma_{1k}$
8: $\mathbf{r}_{2k} = (\eta_{1k}\widehat{\mathbf{x}}_{1k} - \gamma_{1k}\mathbf{r}_{1k})/\gamma_{2k}$
9:
10: // LMMSE estimation
11: $\widehat{\mathbf{x}}_{2k} = \mathbf{g}_2(\mathbf{r}_{2k}, \gamma_{2k})$
12: $\alpha_{2k} = \langle \mathbf{g}_2'(\mathbf{r}_{2k}, \gamma_{2k})\rangle$
13: $\eta_{2k} = \gamma_{2k}/\alpha_{2k}$
14: $\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$
15: $\mathbf{r}_{1,k+1} = (\eta_{2k}\widehat{\mathbf{x}}_{2k} - \gamma_{2k}\mathbf{r}_{2k})/\gamma_{1,k+1}$
16: **end for**
17: Return $\widehat{\mathbf{x}}_{1K_{\mathrm{it}}}$.

C. Implementation Details

For practical implementation with finite-dimensional $\mathbf{A}$, we find that it helps to make some small enhancements to VAMP. In the discussion below we will refer to Algorithm 2, but the same approaches apply to Algorithm 3.

First, we suggest to clip the precisions γ_k and $\widetilde{\gamma}_k$ to a positive interval $[\gamma_{\min}, \gamma_{\max}]$. It is possible, though uncommon, for line 6 of Algorithm 2 to return a negative α_k, which will lead to negative γ_k and $\widetilde{\gamma}_k$ if not accounted for. For the numerical results in Section VI, we used $\gamma_{\min} = 1 \times 10^{-11}$ and $\gamma_{\max} = 1 \times 10^{11}$.

Second, we find that a small amount of damping can be helpful when $\mathbf{A}$ is highly ill-conditioned. In particular, we suggest to replace lines 5 and 10 of Algorithm 2 with the damped versions

$$\widehat{\mathbf{x}}_k = \rho\mathbf{g}_1(\mathbf{r}_k, \gamma_k) + (1-\rho)\widehat{\mathbf{x}}_{k-1} \tag{26}$$

$$\gamma_{k+1} = \rho\widetilde{\gamma}_k\langle \mathbf{d}_k\rangle R/(N - \langle \mathbf{d}_k\rangle R) + (1-\rho)\gamma_k \tag{27}$$

for all iterations $k > 1$, where $\rho \in (0, 1]$ is a suitably chosen damping parameter. Note that, when $\rho = 1$, the damping has no effect. For the numerical results in Section VI, we used $\rho = 0.95$.

Third, rather than requiring VAMP to complete $K_{\rm it}$ iterations, we suggest that the iterations are stopped when the normalized difference $\|\mathbf{r}_{1k} - \mathbf{r}_{1,k-1}\|/\|\mathbf{r}_{1k}\|$ falls below a tolerance τ. For the numerical results in Section VI, we used $\tau = 1 \times 10^{-4}$.

Finally, we note that the VAMP algorithm requires the user to choose the measurement-noise precision γ_w and the denoiser $\mathbf{g}_1(\cdot, \gamma_k)$. Ideally, the true noise precision γ_{w0} is known and the signal $\mathbf{x}^0$ is i.i.d. with known prior $p(x_j)$, in which case the MMSE denoiser can be straightforwardly designed. In practice, however, γ_{w0} and $p(x_j)$ are usually unknown. Fortunately, there is a simple expectation-maximization (EM)-based method to estimate both quantities on-line, whose details are given in [33]. The numerical results in [33] show that the convergence and asymptotic performance of EM-VAMP is nearly identical to that of VAMP with known γ_{w0} and $p(x_j)$. For the numerical results in Section VI, however, we assume that γ_{w0} and $p(x_j)$ are known.

Matlab implementations of VAMP and EM-VAMP can be found in the public-domain GAMPmatlab toolbox [34].

IV. State Evolution

A. Large-System Analysis

Our primary goal is to understand the behavior of the VAMP algorithm for a certain class of matrices in the high-dimensional regime. We begin with an overview of our analysis framework and follow with more details in later sections.

1) Linear measurement model: Our analysis considers a sequence of problems indexed by the signal dimension N. For each N, we assume that there is a "true" vector $\mathbf{x}^0 \in \mathbb{R}^N$ which is observed through measurements of the form,

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w} \in \mathbb{R}^N, \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \gamma_{w0}^{-1}\mathbf{I}_N), \tag{28}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a known transform and $\mathbf{w}$ is Gaussian noise with precision γ_{w0}. Note that we use γ_{w0} to denote the "true" noise precision to distinguish it from γ_w, which is the noise precision postulated by the estimator.

For the transform $\mathbf{A}$, our key assumption is that it can be modeled as a large, *right-orthogonally invariant* random matrix. Specifically, we assume that it has an SVD of the form

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}, \quad \mathbf{S} = \mathrm{Diag}(\mathbf{s}), \tag{29}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N \times N$ orthogonal matrices such that $\mathbf{U}$ is deterministic and $\mathbf{V}$ is Haar distributed (i.e. uniformly distributed on the set of orthogonal matrices). We refer to $\mathbf{A}$ as "right-orthogonally invariant" because the distribution of $\mathbf{A}$ is identical to that of $\mathbf{A}\mathbf{V}_0$ for any fixed orthogonal matrix $\mathbf{V}_0$. We will discuss the distribution of the singular values $\mathbf{s} \in \mathbb{R}^N$ below.

Although we have assumed that $\mathbf{A}$ is square to streamline the analysis, we make this assumption without loss of generality. For example, by setting

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} \mathbf{s}_0 \\ \mathbf{0} \end{bmatrix},$$

our formulation can model a wide rectangular matrix whose SVD is $\mathbf{U}_0\mathbf{S}_0\mathbf{V}^{\mathsf{T}}$ with $\mathrm{diag}(\mathbf{S}_0) = \mathbf{s}_0$. A similar manipulation allows us to model a tall rectangular matrix.

2) Denoiser: Our analysis applies to a fairly general class of denoising functions $\mathbf{g}_1(\cdot, \gamma_{1k})$ indexed by the parameter $\gamma_{1k} \geq 0$. Our main assumption is that the denoiser is separable, meaning that it is of the form (8) for some scalar denoiser $g_1(\cdot, \gamma_{1k})$. As discussed above, this separability assumption will occur for the MAP and MMSE denoisers under the assumption of an i.i.d. prior. However, we do not require the denoiser to be MAP or MMSE for any particular prior. We will impose certain Lipschitz continuity conditions on $g_1(\cdot, \gamma_1 k)$ in the sequel.

3) Asymptotic distributions: It remains to describe the distributions of the true vector $\mathbf{x}^0$ and the singular-value vector $\mathbf{s}$. A simple model would be to assume that they are random i.i.d. sequences that grow with N. However, following the Bayati-Montanari analysis [4], we will consider a more general framework where each of these vectors is modeled as deterministic sequence for which the empirical distribution of the components converges in distribution. When the vectors $\mathbf{x}^0$ and $\mathbf{s}$ are i.i.d. random sequences, they will satisfy this condition almost surely. Details of this analysis framework are reviewed in Appendix B.

Using the definitions in Appendix B, we assume that the components of the singular-value vector $\mathbf{s} \in \mathbb{R}^N$ in (29) converge empirically with second-order moments as

$$\lim_{N\to\infty} \{s_n\} \overset{PL(2)}{=} S, \tag{30}$$

for some positive random variable S. We assume that $\mathbb{E}[S] > 0$ and $S \in [0, S_{max}]$ for some finite maximum value S_{max}. Additionally, we assume that the components of the true vector, $\mathbf{x}^0$, and the initial input to the denoiser, $\mathbf{r}_{10}$, converge empirically as

$$\lim_{N\to\infty} \{(r_{10,n}, x_n^0)\} \overset{PL(2)}{=} (R_{10}, X^0), \tag{31}$$

for some random variables (R_{10}, X^0). Note that the convergence with second-order moments requires that $\mathbb{E}[(X^0)^2] < \infty$ and $\mathbb{E}[R_{10}^2] < \infty$, so they have bounded second moments. We also assume that the initial second-order term converges as

$$\lim_{N\to\infty} \gamma_{10} = \overline{\gamma}_{10}, \tag{32}$$

for some $\overline{\gamma}_{10} > 0$.

As stated above, most of our analysis will apply to general separable denoisers $g_1(\cdot, \gamma_{1k})$. However, some results will apply specifically to MMSE denoisers. Under the assumption that the components of the true vector $\mathbf{x}^0$ are asymptotically distributed like the random variable X^0, the MMSE denoiser (14) and its derivative (16) reduce to

$$\begin{aligned} g_1(r_1, \gamma_1) &= \mathbb{E}\left[X^0 | R_1 = r_1\right], \\ g_1'(r_1, \gamma_1) &= \gamma_1 \mathrm{var}\left[X^0 | R_1 = r_1\right], \end{aligned} \tag{33}$$

where R_1 is the random variable representing X^0 corrupted by AWGN noise, i.e.,

$$R_1 = X^0 + P, \quad P \sim \mathcal{N}(0, \gamma_1^{-1}),$$

with P being independent of X^0. Thus, the MMSE denoiser and its derivative can be computed from the posterior mean and variance of X^0 under an AWGN measurement.

B. Error Functions

Before describing the state evolution (SE) equations and the analysis in the LSL, we need to introduce two key functions: *error functions* and *sensitivity functions*. We begin by describing the error functions.

The error functions, in essence, describe the mean squared error (MSE) of the denoiser and LMMSE estimators under AWGN measurements. Specifically, for the denoiser $g_1(\cdot, \gamma_1)$, we define the error function as

$$\begin{aligned} \mathcal{E}_1(\gamma_1, \tau_1) &:= \mathbb{E}\left[(g_1(R_1, \gamma_1) - X^0)^2\right], \\ R_1 &= X^0 + P, \quad P \sim \mathcal{N}(0, \tau_1). \end{aligned} \tag{34}$$

The function $\mathcal{E}_1(\gamma_1, \tau_1)$ thus represents the MSE of the estimate $\widehat{X} = g_1(R_1, \gamma_1)$ from a measurement R_1 corrupted by Gaussian noise of variance τ_1. For the LMMSE estimator, we define the error function as

$$\begin{aligned} \mathcal{E}_2(\gamma_2, \tau_2) &:= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\|\mathbf{g}_2(\mathbf{r}_2, \gamma_2) - \mathbf{x}^0\|^2, \\ \mathbf{r}_2 &= \mathbf{x}^0 + \mathbf{q}, \quad \mathbf{q} \sim \mathcal{N}(0, \tau_2\mathbf{I}), \\ \mathbf{y} &= \mathbf{A}\mathbf{x}^0 + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0, \gamma_{w0}^{-1}\mathbf{I}), \end{aligned} \tag{35}$$

which is the average per component error of the vector estimate under Gaussian noise. Note that $\mathcal{E}_2(\gamma_2, \tau_2)$ is implicitly a function of the noise precision levels γ_{w0} and γ_w, but this dependence is omitted to simplify the notation.

We will say that both estimators are "matched" when

$$\tau_1 = \gamma_1^{-1}, \quad \tau_2 = \gamma_2^{-1}, \quad \gamma_w = \gamma_{w0},$$

so that the noise levels used by the estimators both match the true noise levels. Under the matched condition, we will use the simplified notation

$$\mathcal{E}_1(\gamma_1) := \mathcal{E}_1(\gamma_1, \gamma_1^{-1}), \quad \mathcal{E}_2(\gamma_2) := \mathcal{E}_2(\gamma_2, \gamma_2^{-1}).$$

The following lemma establishes some basic properties of the error functions.

Lemma 1. *Recall the error functions $\mathcal{E}_1, \mathcal{E}_2$ defined above.*

(a) For the MMSE denoiser (33) *under the matched condition $\tau_1 = \gamma_1^{-1}$, the error function is the conditional variance*

$$\mathcal{E}_1(\gamma_1) = \text{var}\left[X^0 | R_1 = X^0 + \mathcal{N}(0, \gamma_1^{-1})\right]. \tag{36}$$

(b) The LMMSE error function is given by

$$\mathcal{E}_2(\gamma_2, \tau_2) = \lim_{N\to\infty} \frac{1}{N}\text{Tr}\left[\mathbf{Q}^{-2}\widetilde{\mathbf{Q}}\right], \tag{37}$$

where $\mathbf{Q}$ and $\widetilde{\mathbf{Q}}$ are the matrices

$$\mathbf{Q} := \gamma_w\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_2\mathbf{I}, \quad \widetilde{\mathbf{Q}} := \frac{\gamma_w^2}{\gamma_{w0}}\mathbf{A}^\mathsf{T}\mathbf{A} + \tau_2\gamma_2^2\mathbf{I}. \tag{38}$$

Under the matched condition $\tau_2 = \gamma_2^{-1}$ and $\gamma_w = \gamma_{w0}$,

$$\mathcal{E}_2(\gamma_2) = \lim_{N\to\infty} \frac{1}{N}\text{Tr}\left[\mathbf{Q}^{-1}\right]. \tag{39}$$

(c) The LMMSE error function is also given by

$$\mathcal{E}_2(\gamma_2, \tau_2) = \mathbb{E}\left[\frac{\gamma_w^2 S^2/\gamma_{w0} + \tau_2\gamma_2^2}{(\gamma_w S^2 + \gamma_2)^2}\right], \tag{40}$$

where S is the random variable (30) *representing the distribution of the singular values of $\mathbf{A}$. For the matched condition $\tau_2 = \gamma_2^{-1}$ and $\gamma_w = \gamma_{w0}$,*

$$\mathcal{E}_2(\gamma_2) = \mathbb{E}\left[\frac{1}{\gamma_w S^2 + \gamma_2}\right]. \tag{41}$$

Proof: See Appendix C. ∎

C. Sensitivity Functions

The sensitivity functions describe the expected divergence of the estimator. For the denoiser, the sensitivity function is defined as

$$\begin{aligned} A_1(\gamma_1, \tau_1) &:= \mathbb{E}\left[g_1'(R_1, \gamma_1)\right], \\ R_1 &= X^0 + P, \quad P \sim \mathcal{N}(0, \tau_1), \end{aligned} \tag{42}$$

which is the average derivative under a Gaussian noise input. For the LMMSE estimator, the sensitivity is defined as

$$A_2(\gamma_2) := \lim_{N\to\infty} \frac{1}{N}\text{Tr}\left[\frac{\partial \mathbf{g}_2(\mathbf{r}_2, \gamma_2)}{\partial \mathbf{r}_2}\right]. \tag{43}$$

Lemma 2. *For the sensitivity functions above:*

(a) For the MMSE denoiser (33) *under the matched condition $\tau_1 = \gamma_1^{-1}$, the sensitivity function is given by*

$$A_1(\gamma_1, \gamma_1^{-1}) = \gamma_1 \text{var}\left[X^0 | R_1 = X^0 + \mathcal{N}(0, \gamma_1^{-1})\right], \tag{44}$$

which is the ratio of the conditional variance to the measurement variance γ_1^{-1}.

(b) The LMMSE estimator's sensitivity function is given by

$$A_2(\gamma_2) = \lim_{N\to\infty} \frac{1}{N}\gamma_2\text{Tr}\left[(\gamma_w\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_2\mathbf{I})^{-1}\right].$$

(c) The LMMSE estimator's sensitivity function can also be written as

$$A_2(\gamma_2) = \mathbb{E}\left[\frac{\gamma_2}{\gamma_w S^2 + \gamma_2}\right].$$

Proof: See Appendix C. ∎

D. State Evolution Equations

We can now describe our main result, which is the SE equations for VAMP. For a given iteration $k \geq 1$, consider the set of components,

$$\{(\widehat{x}_{1k,n}, r_{1k,n}, x_n^0),\ n = 1, \ldots, N\}.$$

This set represents the components of the true vector $\mathbf{x}^0$, its corresponding estimate $\widehat{\mathbf{x}}_{1k}$ and the input to the denoiser estimator $\mathbf{r}_{1k}$. We will show that, under certain assumptions, these components converge empirically as

$$\lim_{N\to\infty} \{(\widehat{x}_{1k,n}, r_{1k,n}, x_n^0)\} \overset{PL(2)}{=} (\widehat{X}_{1k}, R_{1k}, X^0), \tag{45}$$

where the random variables $(\widehat{X}_{1k}, R_{1k}, X^0)$ are given by

$$R_{1k} = X^0 + P_k, \quad P_k \sim \mathcal{N}(0, \tau_{1k}), \tag{46a}$$

$$\widehat{X}_{1k} = g_1(R_{1k}, \overline{\gamma}_{1k}), \tag{46b}$$

for constants $\overline{\gamma}_{1k}$ and τ_{1k} that will be defined below. Thus, each component $r_{1k,n}$ appears as the true component x_n^0 plus Gaussian noise. The corresponding estimate $\widehat{x}_{1k,n}$ then appears as the denoiser output with $r_{1k,n}$ as the input. Hence, the asymptotic behavior of any component x_n^0 and its corresponding $\widehat{x}_{1k,n}$ is identical to a simple scalar system. We will refer to (45)-(46) as the denoiser's *scalar equivalent model*.

For the LMMSE estimation function, we define the transformed error and transformed noise,

$$\mathbf{q}_k := \mathbf{V}^{\mathsf{T}}(\mathbf{r}_{2k} - \mathbf{x}^0), \quad \boldsymbol{\xi} := \mathbf{U}^{\mathsf{T}}\mathbf{w}, \tag{47}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the matrices in the SVD decomposition (29). We will show that these transformed errors and singular values s_n converge as

$$\lim_{N\to\infty} \{(q_{k,n}, \xi_n, s_n)\} \overset{PL(2)}{=} (Q_k, \Xi, S), \tag{48}$$

to a set of random variables (Q_k, Ξ, S). These random variables are independent, with S defined in the limit (30) and

$$Q_k \sim \mathcal{N}(0, \tau_{2k}), \quad \Xi \sim \mathcal{N}(0, \gamma_{w0}^{-1}), \tag{49}$$

where τ_{2k} is a variance that will be defined below and γ_{w0} is the noise precision in the measurement model (28). Thus (48)-(49) is a scalar equivalent model for the LMMSE estimator.

The variance terms are defined recursively through what are called *state evolution* equations,

$$\overline{\alpha}_{1k} = A_1(\overline{\gamma}_{1k}, \tau_{1k}) \tag{50a}$$

$$\overline{\eta}_{1k} = \frac{\overline{\gamma}_{1k}}{\overline{\alpha}_{1k}}, \quad \overline{\gamma}_{2k} = \overline{\eta}_{1k} - \overline{\gamma}_{1k} \tag{50b}$$

$$\tau_{2k} = \frac{1}{(1-\overline{\alpha}_{1k})^2}\left[\mathcal{E}_1(\overline{\gamma}_{1k}, \tau_{1k}) - \overline{\alpha}_{1k}^2\tau_{1k}\right], \tag{50c}$$

$$\overline{\alpha}_{2k} = A_2(\overline{\gamma}_{2k}, \tau_{2k}) \tag{50d}$$

$$\overline{\eta}_{2k} = \frac{\overline{\gamma}_{2k}}{\overline{\alpha}_{2k}}, \quad \overline{\gamma}_{1,k+1} = \overline{\eta}_{2k} - \overline{\gamma}_{2k} \tag{50e}$$

$$\tau_{1,k+1} = \frac{1}{(1-\overline{\alpha}_{2k})^2}\left[\mathcal{E}_2(\overline{\gamma}_{2k}, \tau_{2k}) - \overline{\alpha}_{2k}^2\tau_{2k}\right], \tag{50f}$$

which are initialized with

$$\tau_{10} = \mathbb{E}[(R_{10} - X^0)^2], \tag{51}$$

and $\overline{\gamma}_{10}$ defined from the limit (32).

Theorem 1. *Under the above assumptions and definitions, assume additionally that for all iterations k:*

(i) The solution $\overline{\alpha}_{1k}$ from the SE equations (50) *satisfies*

$$\overline{\alpha}_{1k} \in (0, 1). \tag{52}$$

(ii) The functions $A_i(\gamma_i, \tau_i)$ and $\mathcal{E}_i(\gamma_i, \tau_i)$ are continuous at $(\gamma_i, \tau_i) = (\overline{\gamma}_{ik}, \tau_{ik})$.

(iii) The denoiser function $g_1(r_1, \gamma_1)$ and its derivative $g_1'(r_1, \gamma_1)$ are uniformly Lipschitz in r_1 at $\gamma_1 = \overline{\gamma}_{1k}$. (See Appendix B for a precise definition of uniform Lipschitz continuity.)

Then, for any fixed iteration $k \geq 0$,

$$\lim_{N\to\infty} (\alpha_{ik}, \eta_{ik}, \gamma_{ik}) = (\overline{\alpha}_{ik}, \overline{\eta}_{ik}, \overline{\gamma}_{ik}) \tag{53}$$

almost surely. In addition, the empirical limit (45) *holds almost surely for all $k > 0$, and* (48) *holds almost surely for all $k \geq 0$.*

E. Mean Squared Error

One important use of the scalar equivalent model is to predict the asymptotic performance of the VAMP algorithm in the LSL. For example, define the asymptotic mean squared error (MSE) of the iteration-k estimate $\widehat{\mathbf{x}}_{ik}$ as

$$\mathsf{MSE}_{ik} := \lim_{N\to\infty} \frac{1}{N}\|\widehat{\mathbf{x}}_{ik} - \mathbf{x}^0\|^2. \tag{54}$$

For this MSE, we claim that

$$\mathsf{MSE}_{ik} = \mathcal{E}_i(\overline{\gamma}_{ik}, \tau_{ik}). \tag{55}$$

To prove (55) for $i = 1$, we write

$$\begin{aligned}\mathsf{MSE}_{1k} &= \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N}(\widehat{x}_{1k,n} - x_n^0)^2 \\ &\overset{(a)}{=} \mathbb{E}[(\widehat{X}_{1k} - X^0)^2] \\ &\overset{(b)}{=} \mathbb{E}[(g_1(R_1, \overline{\gamma}_{1k}) - X^0)^2] \overset{(c)}{=} \mathcal{E}_1(\overline{\gamma}_{1k})\end{aligned}$$

where (a) and (b) follow from the convergence in (45) and the scalar equivalent model (45), and where (c) follows from (34). Using the scalar equivalent model (48), the definition of $\mathcal{E}_2(\cdot)$ in (35), and calculations similar to the proof of Lemma 1, one can also show that (55) holds for $i = 2$.

Interestingly, this type of calculation can be used to compute any other componentwise distortion metric. Specifically, given any distortion function $d(x, \widehat{x})$ that is pseudo-Lipschitz of order two, its average value is given by

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} d(x_n^0, \widehat{x}_{1k,n}) = \mathbb{E}\left[d(X^0, \widehat{X}_{1k})\right],$$

where the expectation is from the scalar equivalent model (45).

F. Contractiveness of the Denoiser

An essential requirement of Theorem 1 is the condition (52) that $\overline{\alpha}_{1k} \in (0, 1)$. This assumption requires that, in a certain average, the denoiser function $g_1(\cdot, \gamma_1)$ is increasing (i.e., $g_1'(r_{1n}, \gamma_1) > 0$) and is a contraction (i.e., $g_1'(r_{1n}, \gamma_1) < 1$). If these conditions are not met, then $\overline{\alpha}_{1k} \leq 0$ or $\overline{\alpha}_{1k} \geq 1$, and either the estimated precision $\overline{\eta}_{1k}$ or $\overline{\gamma}_{2k}$ in (50b) may be negative, causing subsequent updates to be invalid. Thus, $\overline{\alpha}_{1k}$ must be in the range $(0, 1)$. There are two important conditions under which this increasing contraction property are provably guaranteed:

Strongly convex penalties: Suppose that $g_1(r_{1n},\gamma_1)$ is the either the MAP denoiser (13) or the MMSE denoiser (14) for a density $p(x_n)$ that is strongly log-concave. That is, there exists constants $c_1, c_2 > 0$ such that

$$c_1 \leq -\frac{\partial^2}{\partial x_n^2}\ln p(x_n) \leq c_2.$$

Then, using results from log-concave functions [35], it is shown in [11] that

$$g_1'(r_{1n},\gamma_1) \in \left[\frac{\gamma_1}{c_2+\gamma_1}, \frac{\gamma_1}{c_1+\gamma_1}\right] \subset (0,1),$$

for all r_{1n} and $\gamma_1 > 0$. Hence, from the definition of the sensitivity function (42), the sensitivity $\overline{\alpha}_{1k}$ in (50a) will be in the range $(0,1)$.

Matched MMSE denoising: Suppose that $g_1(r_{1n},\gamma_1)$ is the MMSE denoiser in the matched condition where $\overline{\gamma}_{1k} = \tau_{1k}^{-1}$ for some iteration k. From (44),

$$A_1(\gamma_1,\gamma_1^{-1}) = \gamma_1 \text{var}\left[X^0 | R_1 = X^0 + \mathcal{N}(0,\gamma_1^{-1})\right].$$

Since the conditional variance is positive, $A_1(\gamma_1,\gamma_1^{-1}) > 0$. Also, since the variance is bounded above by the MSE of a linear estimator,

$$\begin{aligned}\gamma_1 \text{var}\left[X^0 | R_1 = X^0 + \mathcal{N}(0,\gamma_1^{-1})\right] \\ \leq \gamma_1 \frac{\gamma_1^{-1}\tau_{x_0}}{\tau_{x_0}+\gamma_1^{-1}} = \frac{\gamma_1\tau_{x_0}}{1+\gamma_1\tau_{x_0}} < 1,\end{aligned}$$

where $\tau_{x0} = \text{var}(X^0)$. Thus, we have $A_1(\gamma_1,\gamma_1^{-1}) \in (0,1)$ and $\overline{\alpha}_{1k} \in (0,1)$.

In the case when the prior is not log-concave and the estimator uses an denoiser that is not perfectly matched, $\overline{\alpha}_{1k}$ may not be in the valid range $(0,1)$. In these cases, VAMP may obtain invalid (i.e. negative) variance estimates.

V. MMSE Denoising, Optimality, and Connections to the Replica Method

An important special case of the VAMP algorithm is when we apply the MMSE optimal denoiser. In this case, the SE equations simplify considerably.

Theorem 2. *Consider the SE equations* (50) *with the MMSE optimal denoiser* (33) *and matched initial condition* $\overline{\gamma}_{10} = \tau_{10}^{-1}$. *Then, for all iterations* $k \geq 0$,

$$\overline{\eta}_{1k} = \frac{1}{\mathcal{E}_1(\overline{\gamma}_{1k})}, \quad \overline{\gamma}_{2k} = \tau_{2k}^{-1} = \overline{\eta}_{1k} - \overline{\gamma}_{1k}, \tag{56a}$$

$$\overline{\eta}_{2k} = \frac{1}{\mathcal{E}_2(\overline{\gamma}_{2k})}, \quad \overline{\gamma}_{1,k+1} = \tau_{1,k+1}^{-1} = \overline{\eta}_{2k} - \overline{\gamma}_{2k}. \tag{56b}$$

In addition, for estimators $i = 1,2$, $\overline{\eta}_{ik}$ *is the inverse MSE:*

$$\overline{\eta}_{ik}^{-1} = \lim_{N\to\infty}\frac{1}{N}\|\widehat{\mathbf{x}}_{ik} - \mathbf{x}^0\|^2. \tag{56c}$$

Proof: See Appendix H. ∎

It is useful to compare this result with the work [20], which uses the *replica method* from statistical physics to predict the asymptotic MMSE error in the LSL. To state the result, given a positive semidefinite matrix $\mathbf{C}$, we define its Stieltjes transform as

$$S_{\mathbf{C}}(\omega) = \frac{1}{N}\text{Tr}\left[(\mathbf{C}-\omega\mathbf{I}_N)^{-1}\right] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\lambda_n - \omega}, \tag{57}$$

where λ_n are the eigenvalues of $\mathbf{C}$. Also, let $R_{\mathbf{C}}(\omega)$ denote the so-called R-transform of $\mathbf{C}$, given by

$$R_{\mathbf{C}}(\omega) = S_{\mathbf{C}}^{-1}(-\omega) - \frac{1}{\omega}, \tag{58}$$

where the inverse $S_{\mathbf{C}}^{-1}(\cdot)$ is in terms of composition of functions. The Stieltjes and R-transforms are discussed in detail in [36]. The Stieltjes and R-transforms can be extended to random matrix sequences by taking limits as $N\to\infty$ (for matrix sequences where these limits converge almost surely).

Now suppose that $\widehat{\mathbf{x}} = \mathbb{E}(\mathbf{x}^0|\mathbf{y})$ is the Bayes optimal estimator (i.e. posterior mean or MMSE estimate of $\mathbf{x}^0$ given $\mathbf{y}$). Let $\overline{\eta}^{-1}$ be the asymptotic inverse MSE

$$\overline{\eta}^{-1} := \lim_{N\to\infty}\frac{1}{N}\|\widehat{\mathbf{x}} - \mathbf{x}^0\|^2.$$

Using a so-called replica symmetric analysis, it is argued in [20] that this MSE should satisfy the fixed point equations

$$\overline{\gamma}_1 = R_{\mathbf{C}}(-\overline{\eta}^{-1}), \quad \overline{\eta}^{-1} = \mathcal{E}_1(\overline{\gamma}_1), \tag{59}$$

where $\mathbf{C} = \gamma_{w0}\mathbf{A}^{\mathsf{T}}\mathbf{A}$. A similar result is given in [15].

Theorem 3. *Let* $\overline{\gamma}_i, \overline{\eta}_i$ *be any fixed point solutions to the SE equations* (56) *of VAMP under MMSE denoising. Then* $\overline{\eta}_1 = \overline{\eta}_2$. *If we define* $\overline{\eta} := \overline{\eta}_i$ *as the common value, then* $\overline{\gamma}_1$ *and* $\overline{\eta}$ *satisfy the replica fixed point equation* (59).

Proof: Note that we have dropped the iteration index k since we are discussing a fixed point. First, (56) shows that, at any fixed point,

$$\overline{\gamma}_1 + \overline{\gamma}_2 = \overline{\eta}_1 = \overline{\eta}_2,$$

so that $\overline{\eta}_1 = \overline{\eta}_2$. Also, in the matched case, (41) shows that

$$\mathcal{E}_2(\overline{\gamma}_2) = S_{\mathbf{C}}(-\overline{\gamma}_2).$$

Since $\overline{\eta}^{-1} = \mathcal{E}_2(\overline{\gamma}_2)$, we have that

$$\overline{\gamma}_1 = \overline{\eta} - \overline{\gamma}_2 = \overline{\eta} + S_{\mathbf{C}}^{-1}(\overline{\eta}^{-1}) = R_{\mathbf{C}}(-\overline{\eta}^{-1}).$$

Also, $\overline{\eta}^{-1} = \overline{\eta}_1^{-1} = \mathcal{E}(\overline{\gamma}_1)$. ∎

The consequence of Theorem 3 is that, if the replica equations (59) have a unique fixed point, then the MSE achieved by the VAMP algorithm exactly matches the Bayes optimal MSE as predicted by the replica method. Hence, if this replica prediction is correct, then the VAMP method provides a computationally efficient method for finding MSE optimal estimates under very general priors—including priors for which the associated penalty functions are not convex.

The replica method, however, is generally heuristic. But in the case of i.i.d. Gaussian matrices, it has recently been shown that the replica prediction can be rigorously proven [37], [38].

VI. Numerical Experiments

In this section, we present numerical experiments that compare the VAMP[4] Algorithm 2 to the VAMP state evolution from Section IV, the replica prediction from [20], the AMP Algorithm 1 from [3], the S-AMP algorithm from [17, Sec. IV], the adaptively damped (AD) GAMP algorithm from [9], and the support-oracle MMSE estimator, whose MSE lower bounds that achievable by any practical method. In all cases, we consider the recovery of vectors $\mathbf{x}^0 \in \mathbb{R}^N$ from AWGN-corrupted measurements $\mathbf{y} \in \mathbb{R}^M$ constructed from (1), where $\mathbf{x}^0$ was drawn i.i.d. zero-mean Bernoulli-Gaussian with $\Pr\{x_j^0 \neq 0\} = 0.1$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\gamma_{w0})$, and where $M = 512$ and $N = 1024$. All methods under test were matched to the true signal and noise statistics. When computing the support-oracle MMSE estimate, the support of $\mathbf{x}^0$ is assumed to be known, in which case the problem reduces to estimating the non-zero coefficients of $\mathbf{x}^0$. Since these non-zero coefficients are Gaussian, their MMSE estimate can be computed in closed form. For VAMP we used the implementation enhancements described in Section III-C. For line 7 of AMP Algorithm 1, we used $1/\gamma_{k+1} = 1/\gamma_{w0} + \frac{N}{M}\alpha_k/\gamma_k$, as specified in [3, Eq. (25)]. For the AMP, S-AMP, and AD-GAMP algorithms, we allowed a maximum of 1000 iterations, and for the VAMP algorithm we allowed a maximum of 100 iterations.

A. Ill-conditioned $\mathbf{A}$

First we investigate algorithm robustness to the condition number of $\mathbf{A}$. For this study, realizations of $\mathbf{A}$ were constructed from the SVD $\mathbf{A} = \overline{\mathbf{U}}\text{Diag}(\overline{\mathbf{s}})\overline{\mathbf{V}}^{\mathsf{T}} \in \mathbb{R}^{M\times N}$ with geometric singular values $\overline{\mathbf{s}} \in \mathbb{R}^M$. That is, $\bar{s}_i/\bar{s}_{i-1} = \rho\ \forall i$, with ρ chosen to achieve a desired condition number $\kappa(\mathbf{A}) := \bar{s}_1/\bar{s}_M$ and with $\bar{s}_1$ chosen so that $\|\mathbf{A}\|_F^2 = N$. The singular vector matrices $\overline{\mathbf{U}}, \overline{\mathbf{V}}$ were drawn uniformly at random from the group of orthogonal matrices, i.e., from the Haar distribution. Finally, the signal and noise variances were set to achieve a signal-to-noise ratio (SNR) $\mathbb{E}[\|\mathbf{Ax}\|^2]/\mathbb{E}[\|\mathbf{w}\|^2]$ of 40 dB.

Figure 3 plots the median normalized MSE (NMSE) achieved by each algorithm over 500 independent realizations of $\{\mathbf{A}, \mathbf{x}, \mathbf{w}\}$, where $\text{NMSE}(\widehat{\mathbf{x}}) := \|\widehat{\mathbf{x}} - \mathbf{x}^0\|^2/\|\mathbf{x}^0\|^2$. To enhance visual clarity, NMSEs were clipped to a maximum value of 1. Also, error bars are shown that (separately) quantify the positive and negative standard deviations of VAMP's NMSE from the median value. The NMSE was evaluated for condition numbers $\kappa(\mathbf{A})$ ranging from 1 (i.e., row-orthogonal $\mathbf{A}$) to 1×10^6 (i.e., highly ill-conditioned $\mathbf{A}$).

In Figure 3, we see that AMP and S-AMP diverged for even mildly ill-conditioned $\mathbf{A}$. We also see that, while adaptive damping helped to extend the operating range of AMP, it had a limited effect. In contrast, Figure 3 shows that VAMP's NMSE stayed relatively close to the replica prediction for all condition numbers $\kappa(\mathbf{A})$. The small gap between VAMP and the replica prediction is due to finite-dimensional effects; the SE analysis from Section IV establishes that this gap closes in the large-system limit. Finally, Figure 3 shows that the oracle bound is close to the replica prediction at small $\kappa(\mathbf{A})$ but not at large $\kappa(\mathbf{A})$.

[4]A Matlab implementation of VAMP can be found in the public-domain GAMPmatlab toolbox [34].

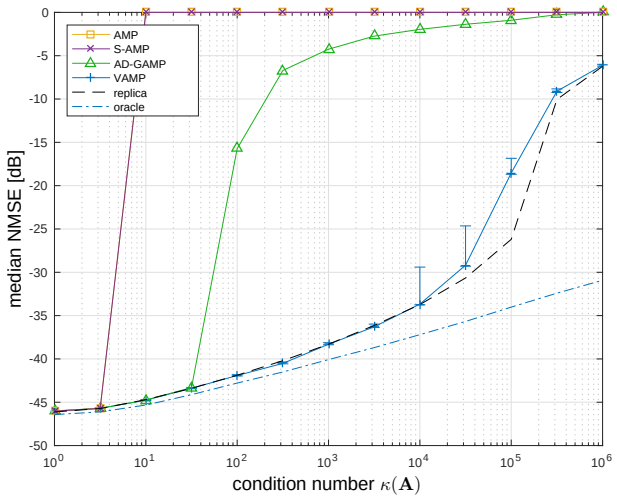

Fig. 3. NMSE versus condition number $\kappa(\mathbf{A})$ at final algorithm iteration. The reported NMSE is the median over 500 realizations, with error bars shown on the VAMP trace.

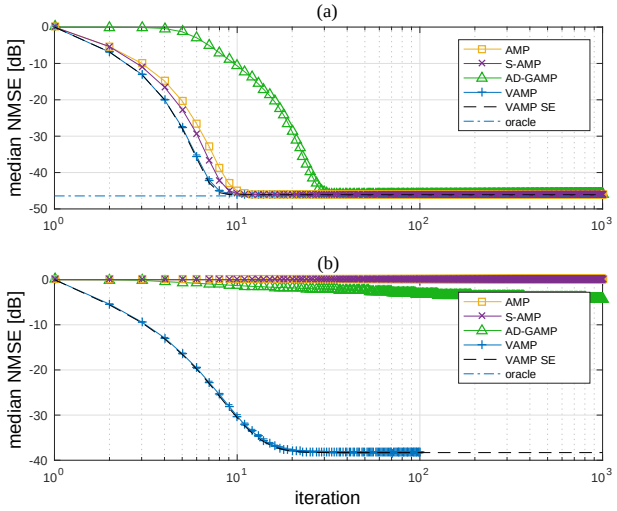

Fig. 4. NMSE versus algorithm iteration for condition number $\kappa(\mathbf{A}) = 1$ in (a) and $\kappa(\mathbf{A}) = 1000$ in (b). The reported NMSE is the median over 500 realizations, with error bars shown on the VAMP traces.

Figure 4(a) plots NMSE versus algorithm iteration for condition number $\kappa(\mathbf{A}) = 1$ and Figure 4(b) plots the same for $\kappa(\mathbf{A}) = 1000$, again with error bars on the VAMP traces. Both figures show that the VAMP trajectory stayed very close to the VAMP-SE trajectory at every iteration. The figures also show that VAMP converges a bit quicker than AMP, S-AMP, and AD-GAMP when $\kappa(\mathbf{A}) = 1$, and that VAMP's convergence rate is relatively insensitive to the condition number $\kappa(\mathbf{A})$.

B. Non-zero-mean $\mathbf{A}$

In this section, we investigate algorithm robustness to the componentwise mean of $\mathbf{A}$. For this study, realizations of $\mathbf{A}$ were constructed by first drawing an i.i.d. $\mathcal{N}(\mu, 1/M)$ matrix and then scaling it so that $\|\mathbf{A}\|_F^2 = N$ (noting that essentially no scaling is needed when $\mu \approx 0$). As before, the signal and

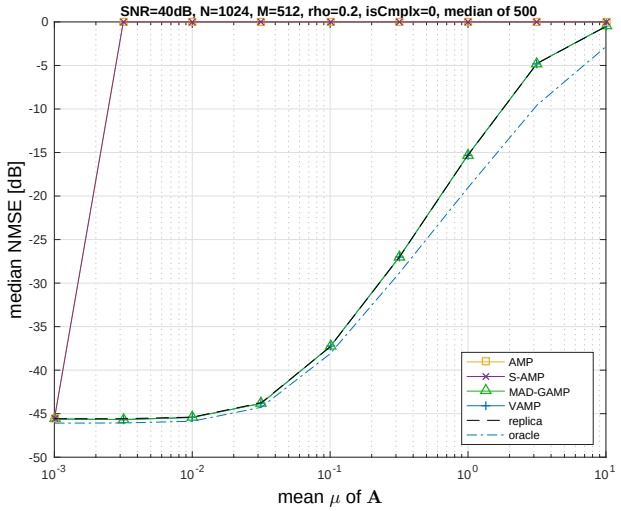

Fig. 5. NMSE versus mean μ at final algorithm iteration. The reported NMSE is the median over 200 realizations, with error bars shown on the VAMP trace.

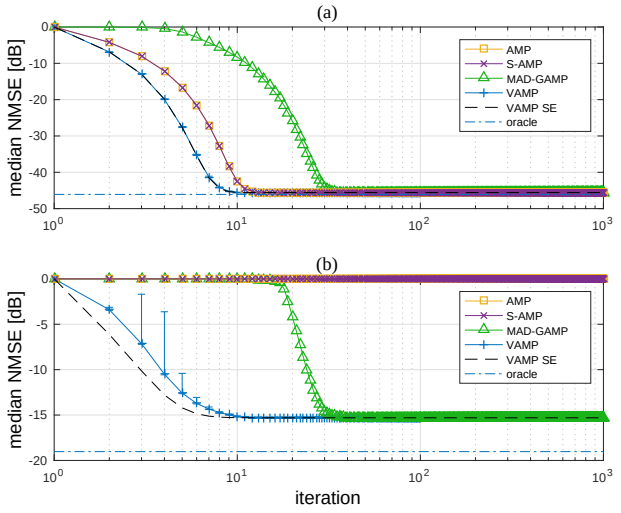

Fig. 6. NMSE versus algorithm iteration when $\mathbf{A}$ has mean $\mu = 0.001$ in (a) and $\mu = 1$ in (b). The reported NMSE is the median over 200 realizations, with error bars shown on the VAMP traces.

noise variances were set to achieve an SNR of 40 dB. For AD-GAMP, we used the mean-removal trick proposed in [9].

Figure 5 plots the NMSE achieved by each algorithm over 200 independent realizations of $\{\mathbf{A}, \mathbf{x}, \mathbf{w}\}$. The NMSE was evaluated for mean parameters μ between 0.001 and 10. Note that, when $\mu > 0.044$, the mean is larger than the standard deviation. Thus, the values of μ that we consider are quite extreme relative to past studies like [8].

Figure 5 shows that AMP and S-AMP diverged for even mildly mean-perturbed $\mathbf{A}$. In contrast, the figure shows that VAMP and mean-removed AD-GAMP (MAD-GAMP) closely matched the replica prediction for all mean parameters μ. It also shows a relatively small gap between the replica prediction and the oracle bound, especially for small μ.

Figure 6(a) plots NMSE versus algorithm iteration for matrix mean $\mu = 0.001$ and Figure 6(b) plots the same for $\mu = 1$. When $\mu = 0.001$, VAMP closely matched its SE at all iterations and converged noticeably quicker than AMP, S-AMP, and MAD-VAMP. When $\mu = 1$, there was a small but noticeable gap between VAMP and its SE for the first few iterations, although the gap closed after about 10 iterations. This gap may be due to the fact that the random matrix $\mathbf{A}$ used for this experiment was not right-rotationally invariant, since the dominant singular vectors are close to (scaled versions of) the $\mathbf{1}$s vector for sufficiently large μ.

SNR	replica	VAMP(stderr)	S-AMP(stderr)	AMP(stderr)
10 dB	5.09e-02	5.27e-02(4.3e-04)	5.27e-02(4.3e-04)	5.42e-02(4.2e-04)
20 dB	3.50e-03	3.57e-03(2.7e-05)	3.58e-03(2.7e-05)	3.62e-03(2.6e-05)
30 dB	2.75e-04	2.84e-04(2.2e-06)	2.85e-04(2.2e-06)	2.85e-04(2.1e-06)

TABLE I
AVERAGE NMSE VERSUS SNR FOR ROW-ORTHOGONAL $\mathbf{A}$, WHERE THE AVERAGE WAS COMPUTED FROM 1000 REALIZATIONS. STANDARD ERROR DEVIATIONS ARE ALSO REPORTED.

C. Row-orthogonal $\mathbf{A}$

In this section we investigate algorithm NMSE versus SNR for row-orthogonal $\mathbf{A}$, i.e., $\mathbf{A}$ constructed as in Section VI-A but with $\kappa(\mathbf{A}) = 1$. Previous studies [15], [18] have demonstrated that, when $\mathbf{A}$ is rotationally invariant but not i.i.d. Gaussian (e.g., row-orthogonal), the fixed points of S-AMP and diagonal-restricted EC are better than those of AMP because the former approaches exploit the singular-value spectrum of $\mathbf{A}$, whereas AMP does not.

Table I reports the NMSE achieved by VAMP, S-AMP, and AMP at three levels of SNR: 10 dB, 20 dB, and 30 dB. The NMSEs reported in the table were computed from an average of 1000 independent realizations of $\{\mathbf{A}, \mathbf{x}, \mathbf{w}\}$. Since the NMSE differences between the algorithms are quite small, the table also reports the standard error on each NMSE estimate to confirm its accuracy.

Table I shows that VAMP and S-AMP gave nearly identical NMSE at all tested SNRs, which is expected because these two algorithms share the same fixed points. The table also shows that VAMP's NMSE was strictly better than AMP's NMSE at low SNR (as expected), but that the NMSE difference narrows as the SNR increases. Finally, the table reports the replica prediction of the NMSE, which is about 3% lower (i.e., -0.15 dB) than VAMP's empirical NMSE at each SNR. We attribute this difference to finite-dimensional effects.

D. Discussion

Our numerical results confirm what is already known about the *fixed points* of diagonally restricted EC (via VAMP) and S-AMP. That is, when $\mathbf{A}$ is large and right-rotationally invariant, they agree with each other and with the replica prediction; and when $\mathbf{A}$ is large i.i.d. Gaussian (which is a special case of right-rotationally invariant [36]), they furthermore agree with the fixed points of AMP [15], [18].

But our numerical results also clarify that it is not enough for an algorithm to have good fixed points, because it may not converge to its fixed points. For example, although the fixed points of S-AMP are good (i.e., replica matching) for *any* large right-rotationally invariant $\mathbf{A}$, our numerical results

indicate that S-AMP converges only for a small subset of large right-rotationally invariant $\mathbf{A}$: those with singular-value spectra similar (or flatter than) i.i.d. Gaussian $\mathbf{A}$.

The SE analysis from Section IV establishes that, in the large-system limit and under matched priors, VAMP is guaranteed to converge to a fixed point that is also a fixed point of the replica equation (59). Our numerical results suggest that, even with large but finite-dimensional right rotationally invariant $\mathbf{A}$ (i.e., 512×1024 in our simulations), VAMP attains NMSEs that are very close to the replica prediction.

VII. Conclusions

In this paper, we considered the standard linear regression (SLR) problem (1), where the goal is to recover the vector $\mathbf{x}^0$ from noisy linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w}$. Our work is inspired by Donoho, Maleki, and Montanari's AMP algorithm [2], which offers a computationally efficient approach to SLR. AMP has the desirable property that its behavior is rigorously characterized under large i.i.d. sub-Gaussian $\mathbf{A}$ by a scalar state evolution whose fixed points, when unique, are Bayes optimal [4]. A major shortcoming of AMP, however, is its fragility with respect to the i.i.d. sub-Gaussian model on $\mathbf{A}$: even small perturbations from this model can cause AMP to diverge.

In response, we proposed a vector AMP (VAMP) algorithm that (after performing an initial SVD) has similar complexity to AMP but is much more robust with respect to the matrix $\mathbf{A}$. Our main contribution is establishing that VAMP's behavior can be rigorously characterized by a scalar state-evolution that holds when $\mathbf{A}$ is large and right-rotationally invariant. The fixed points of VAMP's state evolution are, in fact, consistent with the replica prediction of the minimum mean-squared error recently derived in [20]. We also showed how VAMP can be derived as an approximation of belief propagation on a factor graph with vector-valued nodes, hence the name "vector AMP." Finally, we presented numerical experiments to demonstrate VAMP's robust convergence for ill-conditioned and mean-perturbed matrices $\mathbf{A}$ that cause earlier AMP algorithms to diverge.

As future work, it would be interesting to extend VAMP to the generalized linear model, where the outputs $\mathbf{A}\mathbf{x}^0$ are non-linearly mapped to $\mathbf{y}$. Also, it would be interesting to design and analyze extensions of VAMP that are robust to more general models of $\mathbf{A}$, such as $\mathbf{A}$ that are statistically coupled to $\mathbf{x}^0$.

Appendix A
Message-Passing Derivation of VAMP

In this appendix, we detail the message-passing derivation of Algorithm 3. Below, we will use k to denote the VAMP iteration and n to index the elements of N-dimensional vectors like $\mathbf{x}_1, \mathbf{r}_{1k}$ and $\widehat{\mathbf{x}}_{1k}$. We start by initializing the message-passing with $\mu_{\delta\to\mathbf{x}_1}(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1;\mathbf{r}_{10},\gamma_{10}^{-1}\mathbf{I})$. The following steps are then repeated for $k = 0, 1, 2, \ldots$.

From Rule 1, we first set the approximate belief on $\mathbf{x}_1$ as $\mathcal{N}(\mathbf{x}_1;\widehat{\mathbf{x}}_{1k},\eta_{1k}^{-1}\mathbf{I})$, where $\widehat{\mathbf{x}}_{1k} = \mathbb{E}[\mathbf{x}_1|b_{\mathsf{sp}}(\mathbf{x}_1)]$ and $\eta_{1k}^{-1} = \langle\mathrm{diag}(\mathrm{Cov}[\mathbf{x}_1|b_{\mathsf{sp}}(\mathbf{x}_1)])\rangle$ for the SP belief $b_{\mathsf{sp}}(\mathbf{x}_1) \propto p(\mathbf{x}_1)\mathcal{N}(\mathbf{x}_1;\mathbf{r}_{1k},\gamma_{1k}^{-1}\mathbf{I})$. With an i.i.d. prior $p(\mathbf{x}_1)$ as in (12), we have that $[\widehat{\mathbf{x}}_{1k}]_n = g_1(r_{1k,n},\gamma_{1k})$ for the conditional-mean estimator $g_1(\cdot,\gamma_{1k})$ given in (14), yielding line 4 of Algorithm 3. Furthermore, from (16) we see that the corresponding conditional covariance is $\gamma_{1k}^{-1}g_1'(r_{1k,n},\gamma_{1k})$, yielding lines 5-6 of Algorithm 3.

Next, Rule 2 says to set the message $\mu_{\mathbf{x}_1\to\delta}(\mathbf{x}_1)$ proportional to $\mathcal{N}(\mathbf{x}_1;\widehat{\mathbf{x}}_{1k},\eta_{1k}^{-1}\mathbf{I})/\mathcal{N}(\mathbf{x}_1;\mathbf{r}_{1k},\gamma_{1k}^{-1}\mathbf{I})$. Since

$$\begin{aligned}&\mathcal{N}(\mathbf{x};\widehat{\mathbf{x}},\eta^{-1}\mathbf{I})/\mathcal{N}(\mathbf{x};\mathbf{r},\gamma^{-1}\mathbf{I})\\&\quad\propto\mathcal{N}\big(\mathbf{x};(\widehat{\mathbf{x}}\eta-\mathbf{r}\gamma)/(\eta-\gamma),(\eta-\gamma)^{-1}\mathbf{I}\big),\end{aligned}\tag{60}$$

we have $\mu_{\mathbf{x}_1\to\delta}(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1;\mathbf{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})$ for $\mathbf{r}_{2k} = (\widehat{\mathbf{x}}_{1k}\eta_{1k} - \mathbf{r}_{1k}\gamma_{1k})/(\eta_{1k}-\gamma_{1k})$ and $\gamma_{2k} = \eta_{1k}-\gamma_{1k}$, yielding lines 7-8 of Algorithm 3. Rule 3 then implies that the message $\mu_{\mathbf{x}_1\to\delta}(\mathbf{x}_1)$ will flow rightward through the δ node unchanged, manifesting as $\mu_{\delta\to\mathbf{x}_2}(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2;\mathbf{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})$ on the other side.

Rule 1 then says to set the approximate belief on $\mathbf{x}_2$ at $\mathcal{N}(\mathbf{x}_2;\widehat{\mathbf{x}}_{2k},\eta_{2k}^{-1}\mathbf{I})$, where $\widehat{\mathbf{x}}_{2k} = \mathbb{E}[\mathbf{x}_2|b_{\mathsf{sp}}(\mathbf{x}_2)]$ and $\eta_{2k}^{-1} = \langle\mathrm{diag}(\mathrm{Cov}[\mathbf{x}_2|b_{\mathsf{sp}}(\mathbf{x}_2)])\rangle$ for the SP belief $b_{\mathsf{sp}}(\mathbf{x}_2) \propto \mathcal{N}(\mathbf{x}_2;\mathbf{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})\mathcal{N}(\mathbf{y};\mathbf{A}\mathbf{x}_2,\gamma_w^{-1}\mathbf{I})$. Using standard manipulations, it can be shown that this belief is Gaussian with mean

$$\widehat{\mathbf{x}}_{2k} = \Big(\gamma_w\mathbf{A}^{\mathsf{T}}\mathbf{A}+\gamma_{2k}\mathbf{I}\Big)^{-1}\Big(\gamma_w\mathbf{A}^{\mathsf{T}}\mathbf{y}+\gamma_{2k}\mathbf{r}_{2k}\Big)\tag{61}$$

and covariance $(\gamma_w\mathbf{A}^{\mathsf{T}}\mathbf{A}+\gamma_{2k}\mathbf{I})^{-1}$. The equivalence between (61) and (24) explains line 11 of Algorithm 3. Furthermore, it can be seen by inspection that the average of the diagonal of this covariance matrix coincides with $\gamma_{2k}^{-1}\langle\mathbf{g}_2'(\mathbf{r}_{2k},\gamma_{2k})\rangle$ for $\langle\mathbf{g}_2'(\mathbf{r}_{2k},\gamma_{2k})\rangle$ from (25), thus explaining lines 12-13 of Algorithm 3.

Rule 2 then says to set the message $\mu_{\mathbf{x}_2\to\delta}(\mathbf{x}_2)$ at $\mathcal{N}(\mathbf{x}_2;\widehat{\mathbf{x}}_{2k},\eta_{2k}^{-1}\mathbf{I})/\mathcal{N}(\mathbf{x}_2;\mathbf{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})$, which (60) simplifies to $\mathcal{N}(\mathbf{x}_2;\mathbf{r}_{1,k+1},\gamma_{1,k+1}^{-1}\mathbf{I})$ for $\mathbf{r}_{1,k+1} = (\widehat{\mathbf{x}}_{2k}\eta_{2k} - \mathbf{r}_{2k}\gamma_{2k})/(\eta_{2k} - \gamma_{2k})$ and $\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$, yielding lines 14-15 of Algorithm 3. Finally, Rule 3 implies that the message $\mu_{\mathbf{x}_2\to\delta}(\mathbf{x}_2)$ flows left through the δ node unchanged, manifesting as $\mu_{\delta\to\mathbf{x}_1}(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1;\mathbf{r}_{1k+1},\gamma_{1,k+1}^{-1}\mathbf{I})$ on the other side. The above messaging sequence is then repeated with $k \leftarrow k+1$.

Appendix B
Convergence of Vector Sequences

We review some definitions from the Bayati-Montanari paper [4], since we will use the same analysis framework in this paper. Fix a dimension $r > 0$, and suppose that, for each N, $\mathbf{x}(N)$ is a block vector of the form

$$\mathbf{x}(N) = (\mathbf{x}_1(N),\ldots,\mathbf{x}_N(N)),$$

where each component $\mathbf{x}_n(N) \in \mathbb{R}^r$. Thus, the total dimension of $\mathbf{x}(N)$ is rN. In this case, we will say that $\mathbf{x}(N)$ is a *block vector sequence that scales with N under blocks* $\mathbf{x}_n(N) \in \mathbb{R}^r$. When $r = 1$, so that the blocks are scalar, we will simply say that $\mathbf{x}(N)$ is a *vector sequence that scales with* N. Such vector sequences can be deterministic or random. In most cases, we will omit the notational dependence on N and simply write $\mathbf{x}$.

Now, given $p \geq 1$, a function $\mathbf{f} : \mathbb{R}^s \to \mathbb{R}^r$ is called *pseudo-Lipschitz of order* p, if there exists a constant $C > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^s$,

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leq C\|\mathbf{x}_1 - \mathbf{x}_2\| \left[1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}\right].$$

Observe that in the case $p = 1$, pseudo-Lipschitz continuity reduces to the standard Lipschitz continuity.

Given $p \geq 1$, we will say that the block vector sequence $\mathbf{x} = \mathbf{x}(N)$ converges *empirically with p-th order moments* if there exists a random variable $X \in \mathbb{R}^r$ such that

(i) $\mathbb{E}|X|^p < \infty$; and

(ii) for any scalar-valued pseudo-Lipschitz continuous function f of order p,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n(N)) = \mathbb{E}\left[f(X)\right].$$

Thus, the empirical mean of the components $f(x_n(N))$ converges to the expectation $\mathbb{E}(f(X))$. In this case, with some abuse of notation, we will write

$$\lim_{N\to\infty} \{x_n\} \overset{PL(p)}{=} X, \tag{62}$$

where, as usual, we have omitted the dependence $x_n = x_n(N)$. Importantly, empirical convergence can de defined on deterministic vector sequences, with no need for a probability space. If $\mathbf{x} = \mathbf{x}(N)$ is a random vector sequence, we will often require that the limit (62) holds almost surely.

We conclude with one final definition. Let $\boldsymbol{\phi}(\mathbf{r}, \gamma)$ be a function on $\mathbf{r} \in \mathbb{R}^s$ and $\gamma \in \mathbb{R}$. We say that $\boldsymbol{\phi}(\mathbf{r}, \gamma)$ is *uniformly Lipschitz continuous* in $\mathbf{r}$ at $\gamma = \overline{\gamma}$ if there exists constants L_1 and $L_2 \geq 0$ and an open neighborhood U of $\overline{\gamma}$, such that

$$\|\boldsymbol{\phi}(\mathbf{r}_1, \gamma) - \boldsymbol{\phi}(\mathbf{r}_2, \gamma)\| \leq L_1\|\mathbf{r}_1 - \mathbf{r}_2\|, \tag{63}$$

for all $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^s$ and $\gamma \in U$; and

$$\|\boldsymbol{\phi}(\mathbf{r}, \gamma_1) - \boldsymbol{\phi}(\mathbf{r}, \gamma_2)\| \leq L_2\left(1 + \|\mathbf{r}\|\right)|\gamma_1 - \gamma_2|, \tag{64}$$

for all $\mathbf{r} \in \mathbb{R}^s$ and $\gamma_1, \gamma_2 \in U$.

Appendix C
Proof of Lemmas 1 and 2

For Lemma 1, part (a) follows immediately from (33) and (34). To prove part (b), suppose

$$\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w}, \quad \mathbf{r}_2 = \mathbf{x}^0 + \mathbf{q}.$$

Then, the error is given by

$$\begin{aligned}
\mathbf{g}_2(\mathbf{r}_2, \gamma_2) - \mathbf{x}^0 &\overset{(a)}{=} \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_2 \mathbf{I}\right)^{-1} \\
&\quad \times \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}^0 + \gamma_w \mathbf{A}^\mathsf{T}\mathbf{w} + \gamma_2 \mathbf{r}_2\right) - \mathbf{x}^0 \\
&\overset{(b)}{=} \left(\gamma_w \mathbf{A}^\mathsf{T}\mathbf{A} + \gamma_2 \mathbf{I}\right)^{-1} \left(\gamma_2 \mathbf{q} + \gamma_w \mathbf{A}^\mathsf{T}\mathbf{w}\right), \\
&\overset{(c)}{=} \mathbf{Q}^{-1}\left(\gamma_2 \mathbf{q} + \gamma_w \mathbf{A}^\mathsf{T}\mathbf{w}\right),
\end{aligned}$$

where (a) follows by substituting $\mathbf{y} = \mathbf{A}\mathbf{x}^0 + \mathbf{w}$ into (24); part (b) follows from the substitution $\mathbf{r} = \mathbf{x}^0 + \mathbf{q}$ and collecting the terms with $\mathbf{x}^0$; and (c) follows from the definition of $\mathbf{Q}$ in (38). Hence, the error covariance matrix is given

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{g}_2(\mathbf{r}_2, \gamma_2) - \mathbf{x}^0)(\mathbf{g}_2(\mathbf{r}_2, \gamma_2) - \mathbf{x}^0)^\mathsf{T}\right] \\
&\quad = \mathbf{Q}^{-1}\left[\gamma_2^2 \mathbb{E}(\mathbf{q}\mathbf{q}^\mathsf{T}) + \gamma_w^2 \mathbf{A}\mathbb{E}(\mathbf{w}\mathbf{w}^\mathsf{T})\mathbf{A}^\mathsf{T}\right]\mathbf{Q}^{-1} \\
&\quad = \mathbf{Q}^{-1}\widetilde{\mathbf{Q}}\mathbf{Q}^{-1},
\end{aligned}$$

where we have used the the fact that $\mathbf{q}$ and $\mathbf{w}$ are independent Gaussians with variances τ_2 and γ_{w0}^{-1}. This proves (37). Under the matched condition, $\mathbf{Q} = \widetilde{\mathbf{Q}}$, which proves (39). Part (c) of Lemma 1 follows from part (b) by using the SVD (29).

For Lemma 2, part (a) follows from averaging (33) over r_1. Part (b) follows by taking the derivative in (24) and part (c) follows from using the SVD (29).

Appendix D
Orthogonal Matrices Under Linear Constraints

In preparation for proving Theorem 1, we derive various results on orthogonal matrices subject to linear constraints. To this end, suppose $\mathbf{V} \in \mathbb{R}^{N\times N}$ is an orthogonal matrix satisfying linear constraints

$$\mathbf{A} = \mathbf{V}\mathbf{B}, \tag{65}$$

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 $s \leq N$). Let

$$\mathbf{U}_\mathbf{A} = \mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1/2}, \quad \mathbf{U}_\mathbf{B} = \mathbf{B}(\mathbf{B}^\mathsf{T}\mathbf{B})^{-1/2}. \tag{66}$$

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. Define

$$\widetilde{\mathbf{V}} := \mathbf{U}_{\mathbf{A}^\perp}^\mathsf{T}\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp}, \tag{67}$$

which has dimension $(N - s) \times (N - s)$.

Lemma 3. *Under the above definitions $\widetilde{\mathbf{V}}$ satisfies*

$$\mathbf{V} = \mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{B}^\mathsf{T} + \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp}^\mathsf{T}. \tag{68}$$

Proof: Let $\widehat{\mathbf{V}}$ be the right hand side of (68). That is,

$$\widehat{\mathbf{V}} = \mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{B}^\mathsf{T} + \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp}^\mathsf{T}. \tag{69}$$

So, we need to show $\widehat{\mathbf{V}} = \mathbf{V}$. First observe that $\mathbf{U}_\mathbf{A}$ and $\mathbf{U}_\mathbf{B}$ in (66) are matrices whose columns are an orthonormal basis of $\mathrm{Range}(\mathbf{A})^\perp$ and $\mathrm{Range}(\mathbf{B})^\perp$, respectively. Hence,

$$[\mathbf{U}_\mathbf{A}\ \mathbf{U}_{\mathbf{A}^\perp}], \quad [\mathbf{U}_\mathbf{B}\ \mathbf{U}_{\mathbf{B}^\perp}],$$

are both orthogonal. Hence to show $\widehat{\mathbf{V}} = \mathbf{V}$, it suffices to prove that

$$\widehat{\mathbf{V}}\mathbf{U}_\mathbf{B} = \mathbf{V}\mathbf{U}_\mathbf{B} \tag{70}$$

$$\widehat{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{V}\mathbf{U}_{\mathbf{B}^\perp}. \tag{71}$$

To prove (70), note that since $\mathbf{V}$ is orthogonal

$$\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{B}^\mathsf{T}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{B} = \mathbf{B}\mathbf{B}.$$

Using this property and the definitions in (66), we see that

$$\mathbf{U}_\mathbf{A} = \mathbf{V}\mathbf{U}_\mathbf{B}.$$

Also,

$$\widehat{\mathbf{V}}\mathbf{U}_{\mathbf{B}} \stackrel{(a)}{=} \mathbf{A}(\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{U}_{\mathbf{B}} \stackrel{(b)}{=} \mathbf{U}_{\mathbf{A}},$$

where (a) follows from (69) and the fact that $\mathbf{U}_{\mathbf{B}^\perp}\mathbf{U}_{\mathbf{B}} = \mathbf{0}$ and (b) follows from the definitions in (66). This proves (70).

To prove (71), first observe that

$$\mathbf{A}^{\mathsf{T}}\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{B}^{\mathsf{T}}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{0},$$

and hence $\mathrm{Range}(\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp})$ is contained in the $\mathrm{Range}(\mathbf{A})^\perp$. Since $\mathbf{U}_{\mathbf{A}^\perp}\mathbf{U}_{\mathbf{A}^\perp}^{\mathsf{T}}$ is an orthogonal projection onto $\mathrm{Range}(\mathbf{A})^\perp$, we have

$$\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{U}_{\mathbf{A}^\perp}\mathbf{U}_{\mathbf{A}^\perp}^{\mathsf{T}}\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}}. \tag{72}$$

Therefore,

$$\begin{aligned}\widehat{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp} &\stackrel{(a)}{=} \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}^\perp}^{\mathsf{T}}\mathbf{U}_{\mathbf{B}^\perp}\\ &\stackrel{(b)}{=} \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{V}} \stackrel{(c)}{=} \mathbf{V}\mathbf{U}_{\mathbf{B}^\perp},\end{aligned}$$

where (a) follows from (69) and the fact that $\mathbf{B}^{\mathsf{T}}\mathbf{U}_{\mathbf{B}^\perp} = \mathbf{0}$; (b) follows from the fact that the columns of $\mathbf{U}_{\mathbf{B}^\perp}$ are orthonormal; and (c) from (72) This shows (71) and we have proven (68).

To prove that $\widetilde{\mathbf{V}}$ is orthogonal,

$$\begin{aligned}\widetilde{\mathbf{V}}^{\mathsf{T}}\widetilde{\mathbf{V}} &\stackrel{(a)}{=} \mathbf{U}_{\mathbf{B}^\perp}^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\mathbf{U}_{\mathbf{A}^\perp}\mathbf{U}_{\mathbf{A}^\perp}^{\mathsf{T}}\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp}\\ &\stackrel{(b)}{=} \mathbf{U}_{\mathbf{B}^\perp}^{\mathsf{T}}\mathbf{V}^{\mathsf{T}}\mathbf{V}\mathbf{U}_{\mathbf{B}^\perp} \stackrel{(c)}{=} \mathbf{I},\end{aligned}$$

where (a) uses (67); (b) follows from (72) and (c) follows from the fact that $\mathbf{V}$ and $\mathbf{U}_{\mathbf{B}^\perp}$ have orthonormal columns. ■

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* (65)*. Then, the conditional distribution given G, $\widetilde{\mathbf{V}}$ is Haar distributed matrix independent of G. Thus,*

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

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

Proof: Let O_N be the set of $N \times N$ orthogonal matrices and let $\mathcal{L}$ be the set of matrices $\mathbf{V} \in O_N$ that satisfy the linear constraints (65). If $p_{\mathbf{V}}(\mathbf{V})$ is the uniform density on O_N (i.e. the Haar measure), the conditional density on $\mathbf{V}$ given the event G,

$$p_{\mathbf{V}|G}(\mathbf{V}|G) = \frac{1}{Z}p_{\mathbf{V}}(\mathbf{V})\mathbb{1}_{\{\mathbf{V}\in\mathcal{L}\}},$$

where Z is the normalization constant. Now let $\phi : \widetilde{\mathbf{V}} \mapsto \mathbf{V}$ be the mapping described by (68) which maps O_{N-s} to $\mathcal{L}$. This mapping is invertible. Since ϕ is affine, the conditional density on $\widetilde{\mathbf{V}}$ is given by

$$\begin{aligned}p_{\widetilde{\mathbf{V}}|G}(\widetilde{\mathbf{V}}|G) &\propto p_{\mathbf{V}|G}(\phi(\widetilde{\mathbf{V}})|G)\\ &\propto p_{\mathbf{V}}(\phi(\widetilde{\mathbf{V}}))\mathbb{1}_{\{\phi(\widetilde{\mathbf{V}}\in\mathcal{L})\}} = p_{\mathbf{V}}(\phi(\widetilde{\mathbf{V}})),\end{aligned} \tag{73}$$

where in the last step we used the fact that, for any matrix $\widetilde{\mathbf{V}}$, $\phi(\widetilde{\mathbf{V}}) \in \mathcal{L}$ (i.e. satisfies the linear constraints (65)). Now to show that $\widetilde{\mathbf{V}}$ is conditionally Haar distributed, we need to show that for any orthogonal matrix $\mathbf{W}_0 \in O_{N-s}$,

$$p_{\widetilde{\mathbf{V}}|G}(\mathbf{W}_0\widetilde{\mathbf{V}}|G) = p_{\widetilde{\mathbf{V}}|G}(\widetilde{\mathbf{V}}|G). \tag{74}$$

To prove this, given $\mathbf{W}_0 \in O_{N-s}$, define the matrix,

$$\mathbf{W} = \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^{\mathsf{T}} + \mathbf{U}_{\mathbf{A}^\perp}\mathbf{W}_0\mathbf{U}_{\mathbf{A}^\perp}^{\mathsf{T}}.$$

One can verify that $\mathbf{W} \in O_N$ (i.e. it is orthogonal) and

$$\phi(\mathbf{W}_0\widetilde{\mathbf{V}}) = \mathbf{W}\phi(\widetilde{\mathbf{V}}). \tag{75}$$

Hence,

$$\begin{aligned}p_{\widetilde{\mathbf{V}}|G}(\mathbf{W}_0\widetilde{\mathbf{V}}|G) &\stackrel{(a)}{\propto} p_{\mathbf{V}}(\phi(\mathbf{W}_0\widetilde{\mathbf{V}}))\\ &\stackrel{(b)}{\propto} p_{\mathbf{V}}(\mathbf{W}\phi(\widetilde{\mathbf{V}})) \stackrel{(c)}{\propto} p_{\mathbf{V}}(\phi(\widetilde{\mathbf{V}})),\end{aligned}$$

where (a) follows from (73); (b) follows from (75); and (c) follows from the orthogonal invariance of $\mathbf{V}$. Hence, the conditional density of $\widetilde{\mathbf{V}}$ is invariant under orthogonal transforms and is thus Haar distributed. ■

We will use Lemma 4 in conjunction with the following simple result.

Lemma 5. *Fix a dimension $s \geq 0$, and suppose that $\mathbf{x}(N)$ and $\mathbf{U}(N)$ are sequences such that for each N,*

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

(ii) $\mathbf{x} = \mathbf{x}(N) \in \mathbb{R}^{N-s}$ *a random vector that is is isotropically distributed in that* $\mathbf{V}\mathbf{x} \stackrel{d}{=} \mathbf{x}$ *for any orthogonal* $(N-s)\times(N-s)$ *matrix* $\mathbf{V}$*.*

(iii) The mean squared component magnitude converges almost surely as

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

for some $\tau > 0$.

Then, if we define $\mathbf{y} = \mathbf{U}\mathbf{x}$, we have that the components of $\mathbf{y}$ converge empirically to a Gaussian random variable

$$\lim_{N\to\infty}\{y_n\} \stackrel{PL(2)}{=} Y \sim \mathcal{N}(0,\tau). \tag{76}$$

Proof: Since $\mathbf{x}$ is isotropically distributed, it can be generated as a normalized Gaussian, i.e.

$$\mathbf{x} \stackrel{d}{=} \frac{\|\mathbf{x}\|}{\|\mathbf{w}_0\|}\mathbf{w}_0, \quad \mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N-s}).$$

For each N, let $\mathbf{U}_\perp$ be an $N\times s$ matrix such that $\mathbf{S} := [\mathbf{U}\ \mathbf{U}_\perp]$ is orthogonal. That is, the s columns of $\mathbf{U}_\perp$ are an orthonormal basis of the orthogonal complement of the $\mathrm{Range}(\mathbf{U})$. If we let $\mathbf{w}_1 \sim \mathcal{N}(0, \mathbf{I}_s)$, then if we define

$$\mathbf{w} = \begin{bmatrix}\mathbf{w}_0\\ \mathbf{w}_1\end{bmatrix},$$

so that $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_N)$. With this definition, we can write $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{U}\mathbf{x} \stackrel{d}{=} \frac{\|\mathbf{x}\|}{\|\mathbf{w}_0\|}\left[\mathbf{S}\mathbf{w} - \mathbf{U}_\perp\mathbf{w}_1\right]. \tag{77}$$

Now,

$$\lim_{N\to\infty}\frac{\|\mathbf{x}\|}{\|\mathbf{w}_0\|} = \sqrt{\tau},$$

almost surely. Also, since $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{S}$ is orthogonal, $\mathbf{S}\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$. Finally, since $\mathbf{w}_1$ is s-dimensional,

$$\lim_{N\to\infty}\frac{1}{N}\|\mathbf{U}_\perp\mathbf{w}_1\|^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{w}_1\|^2 = 0,$$

almost surely. Substituting these properties into (77), we obtain (76). ■

APPENDIX E
A GENERAL CONVERGENCE RESULT

To analyze the VAMP method, we a consider the following more general recursion. We are given a dimension N, an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$, an initial vector $\mathbf{u}_0 \in \mathbb{R}^N$, disturbance vectors $\mathbf{w}^p, \mathbf{w}^q \in \mathbb{R}^N$. Then, we generate a sequence of iterates by the following recursion:

$$\begin{aligned}
\mathbf{p}_k &= \mathbf{V}\mathbf{u}_k && \text{(78a)}\\
\alpha_{1k} &= \langle \mathbf{f}_p'(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k})\rangle, \quad \gamma_{2k} = \Gamma_1(\gamma_{1k}, \alpha_{1k}) && \text{(78b)}\\
\mathbf{v}_k &= C_1(\alpha_{1k})\left[\mathbf{f}_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k}) - \alpha_{1k}\mathbf{p}_k\right] && \text{(78c)}\\
\mathbf{q}_k &= \mathbf{V}^{\mathsf{T}}\mathbf{v}_k && \text{(78d)}\\
\alpha_{2k} &= \langle \mathbf{f}_q'(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k})\rangle, \quad \gamma_{1,k+1} = \Gamma_2(\gamma_{2k}, \alpha_{2k}) && \text{(78e)}\\
\mathbf{u}_{k+1} &= C_2(\alpha_{2k})\left[\mathbf{f}_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) - \alpha_{2k}\mathbf{q}_k\right], && \text{(78f)}
\end{aligned}$$

which is initialized with some vector $\mathbf{u}_0$ and scalar γ_{10}. Here, $\mathbf{f}_p(\cdot)$ and $\mathbf{f}_q(\cdot)$ are separable functions, meaning

$$\begin{aligned}
\left[\mathbf{f}_p(\mathbf{p}, \mathbf{w}^p, \gamma_1)\right]_n &= f_p(p_n, w_n^p, \gamma_1)\ \forall n,\\
\left[\mathbf{f}_q(\mathbf{q}, \mathbf{w}^q, \gamma_2)\right]_n &= f_q(q_n, w_n^q, \gamma_2)\ \forall n,
\end{aligned} \tag{79}$$

for scalar-valued functions $f_p(\cdot)$ and $f_q(\cdot)$. The functions $\Gamma_i(\cdot)$ and $C_i(\cdot)$ are also scalar-valued.

Similar to our analysis of the VAMP, we consider the following large-system limit (LSL) analysis. We consider a sequence of runs of the recursions indexed by N. We model the initial condition $\mathbf{u}_0$ and disturbance vectors $\mathbf{w}^p$ and $\mathbf{w}^q$ as deterministic sequences that scale with N and assume that their components converge empirically as

$$\lim_{N\to\infty}\{u_{0n}\} \overset{PL(2)}{=} U_0, \tag{80}$$

and

$$\lim_{N\to\infty}\{w_n^p\} \overset{PL(2)}{=} W^p, \quad \lim_{N\to\infty}\{w_n^q\} \overset{PL(2)}{=} W^q, \tag{81}$$

to random variables U_0, W^p and W^q. We assume that the initial constant converges as

$$\lim_{N\to\infty}\gamma_{10} = \overline{\gamma}_{10}, \tag{82}$$

for some $\overline{\gamma}_{10}$. The matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$ is assumed to be uniformly distributed on the set of orthogonal matrices independent of $\mathbf{r}_0$, $\mathbf{w}^p$ and $\mathbf{w}^q$. Since $\mathbf{r}_0$, $\mathbf{w}^p$ and $\mathbf{w}^q$ are deterministic, the only randomness is in the matrix $\mathbf{V}$.

Under the above assumptions, define the SE equations

$$\begin{aligned}
\overline{\alpha}_{1k} &= \mathbb{E}\left[f_p'(P_k, W^p, \overline{\gamma}_{1k})\right], && \text{(83a)}\\
\tau_{2k} &= C_1^2(\overline{\alpha}_{1k})\left\{\mathbb{E}\left[f_p^2(P_k, W^p, \overline{\gamma}_{1k})\right] - \overline{\alpha}_{1k}^2\tau_{1k}\right\} && \text{(83b)}\\
\overline{\gamma}_{2k} &= \Gamma_1(\overline{\gamma}_{1k}, \overline{\alpha}_{1k}) && \text{(83c)}\\
\overline{\alpha}_{2k} &= \mathbb{E}\left[f_q'(Q_k, W^q, \overline{\gamma}_{2k}\right], && \text{(83d)}\\
\tau_{1,k+1} &= C_2^2(\overline{\alpha}_{2k})\left\{\mathbb{E}\left[f_q^2(Q_k, W^q, \overline{\gamma}_{2k})\right] - \overline{\alpha}_{2k}^2\tau_{2k}\right\} && \text{(83e)}\\
\gamma_{1,k+1} &= \Gamma_2(\overline{\gamma}_{2k}, \overline{\alpha}_{2k}), && \text{(83f)}
\end{aligned}$$

which are initialized with $\overline{\gamma}_{10}$ in (82) and

$$\tau_{10} = \mathbb{E}U_0^2, \tag{84}$$

where U_0 is the random variable in (80). In the SE equations (83), the expectations are taken with respect to random variables

$$P_k \sim \mathcal{N}(0, \tau_{1k}), \quad Q_k \sim \mathcal{N}(0, \tau_{2k}),$$

where P_k is independent of W^p and Q_k is independent of W^q.

Theorem 4. *Consider the recursions* (78) *and SE equations* (83) *under the above assumptions. Assume additionally that, for all k:*

(i) For $i = 1, 2$, the functions

$$C_i(\alpha_i), \quad \Gamma_i(\gamma_i, \alpha_i),$$

are continuous at the points $(\gamma_i, \alpha_i) = (\overline{\gamma}_{ik}, \overline{\alpha}_{ik})$ from the SE equations; and

(ii) The function $f_p(p, w^p, \gamma_1)$ and its derivative $f_p'(p, w^p, \gamma_1)$ are uniformly Lipschitz continuous in (p, w^p) at $\gamma_1 = \overline{\gamma}_{1k}$.

(iii) The function $f_q(q, w^q, \gamma_2)$ and its derivative $f_q'(q, w^q, \gamma_2)$ are uniformly Lipschitz continuous in (q, w^q) at $\gamma_2 = \overline{\gamma}_{2k}$.

Then,

(a) For any fixed k, almost surely the components of $(\mathbf{w}^p, \mathbf{p}_0, \ldots, \mathbf{p}_k)$ empirically converge as

$$\lim_{N\to\infty}\{(w_n^p, p_{0n}, \ldots, p_{kn})\} \overset{PL(2)}{=} (W^p, P_0, \ldots, P_k), \tag{85}$$

where W^p is the random variable in the limit (81) *and $(P_0, \ldots, P_k)$ is a zero mean Gaussian random vector independent of W^p, with $\mathbb{E}(P_k^2) = \tau_{1k}$. In addition, we have that*

$$\lim_{N\to\infty}(\alpha_{1k}, \gamma_{1k}) = (\overline{\alpha}_{1k}, \overline{\gamma}_{1k}), \tag{86}$$

almost surely.

(b) For any fixed k, almost surely the components of $(\mathbf{w}^q, \mathbf{q}_0, \ldots, \mathbf{q}_k)$ empirically converge as

$$\lim_{N\to\infty}\{(w_n^q, q_{0n}, \ldots, q_{kn})\} \overset{PL(2)}{=} (W^q, Q_0, \ldots, Q_k), \tag{87}$$

where W^q is the random variable in the limit (81) *and $(Q_0, \ldots, Q_k)$ is a zero mean Gaussian random vector independent of W^q, with $\mathbb{E}(P_k^2) = \tau_{2k}$. In addition, we have that*

$$\lim_{N\to\infty}(\alpha_{2k}, \gamma_{2k}) = (\overline{\alpha}_{2k}, \overline{\gamma}_{2k}), \tag{88}$$

almost surely.

Proof: We will prove this in the next Appendix, Appendix F. ∎

APPENDIX F
PROOF OF THEOREM 4

A. Induction Argument

We use an induction argument. Given iterations $k, \ell \geq 0$, define the hypothesis, $H_{k,\ell}$ as the statement:

- Part (a) of Theorem 4 is true up to k; and

- Part (b) of Theorem 4 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 $H_{k+1,k}$.

B. Induction Initialization

We first show that the hypothesis $H_{0,-1}$ is true. That is, we must show (85) and (86) for $k = 0$. This is a special case of Lemma 5. Specifically, for each N, let $\mathbf{U} = \mathbf{I}_N$, the $N \times N$ identity matrix, which trivially satisfies property (i) of Lemma 5 with $s = 0$. Let $\mathbf{x} = \mathbf{p}_0$. Since $\mathbf{p}_0 = \mathbf{V}\mathbf{u}_0$ and $\mathbf{V}$ is Haar distributed independent of $\mathbf{u}_0$, we have that $\mathbf{p}_0$ is orthogonally invariant and satisfies property (ii) of Lemma 5. Also,

$$\lim_{N\to\infty} \|\mathbf{p}_0\|^2 \overset{(a)}{=} \lim_{N\to\infty} \|\mathbf{u}_0\|^2 \overset{(b)}{=} \mathbb{E}[U_0^2] \overset{(c)}{=} \tau_{10},$$

where (a) follows from the fact that $\mathbf{p}_0 = \mathbf{V}\mathbf{u}_0$ and $\mathbf{V}$ is orthogonal; (b) follows from the assumption (80) and (c) follows from the definition (84). This proves property (iii) of Lemma 5. Hence, $\mathbf{p}_0 = \mathbf{U}\mathbf{p}_0$, we have that the components of $\mathbf{p}_0$ converge empirically as

$$\lim_{N\to\infty} \{p_{0n}\} \overset{PL(2)}{=} P_0 \sim \mathcal{N}(0, \tau_{10}),$$

for a Gaussian random variable P_0. Moreover, since $\mathbf{V}$ is independent of $\mathbf{w}^p$, and the components of $\mathbf{w}^p$ converge empirically as (81), we have that the components of $\mathbf{p}_n, \mathbf{w}^p$ almost surely converge empirically as

$$\lim_{N\to\infty} \{w_n^p, p_{0n}\} \overset{PL(2)}{=} (W^p, P_0),$$

where W^p is independent of P_0. This proves (85) for $k = 0$.

Now, we have assumed in (82) that $\gamma_{10} \to \overline{\gamma}_{10}$ as $N \to \infty$. Also, since $f_p'(p, w^p, \gamma_1)$ is uniformly Lipschitz continuous in (p, w^p) at $\gamma_1 = \overline{\gamma}_{10}$, we have that $\alpha_{10} = \langle \mathbf{f}_p'(\mathbf{p}_0, \mathbf{w}^p, \gamma_{10}) \rangle$ converges to $\overline{\alpha}_{10}$ in (83a) almost surely. This proves (86).

C. 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. Since $\Gamma_1(\gamma_i, \alpha_i)$ is continuous at $(\overline{\gamma}_{1k}, \overline{\alpha}_{1k})$, the limits (86) combined with (83c) show that

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

In addition, the induction hypothesis shows that for $\ell = 0, \ldots, k$, the components of $(\mathbf{w}^p, \mathbf{p}_\ell)$ almost surely converge empirically as

$$\lim_{N\to\infty} \{(w_n^p, p_{\ell n})\} \overset{PL(2)}{=} (W^p, P_\ell),$$

where $P_\ell \sim \mathcal{N}(0, \tau_{1\ell})$ for $\tau_{1\ell}$ given by the SE equations. Since $f_p(\cdot)$ is Lipschitz continuous, $C_1(\alpha_{1\ell})$ is continuous at $\alpha_{1\ell} = \overline{\alpha}_{1\ell}$, the definition of $\mathbf{v}_\ell$ in (78c) and the limits (86) show that

$$\lim_{N\to\infty} \{(w_n^p, p_{\ell n}, v_{\ell n})\} \overset{PL(2)}{=} (W^p, P_\ell, V_\ell),$$

where V_ℓ is the random variable,

$$V_\ell = g_p(P_\ell, W^p, \overline{\gamma}_{1\ell}, \overline{\alpha}_{1\ell}), \tag{89}$$

and $g_p(\cdot)$ is the function

$$g_p(p, w^p, \gamma_1, \alpha_1) := C_1(\alpha_1)\left[f_p(p, w^p, \gamma_1) - \alpha_1 p\right]. \tag{90}$$

We next introduce the notation

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

to represent the first $k+1$ values of the vector $\mathbf{u}_\ell$. We define the matrices $\mathbf{V}_k$, $\mathbf{Q}_k$ and $\mathbf{P}_k$ similarly. Using this notation, let G_k be the set of random vectors,

$$G_k := \{\mathbf{U}_k, \mathbf{P}_k, \mathbf{V}_k, \mathbf{Q}_{k-1}\}. \tag{91}$$

With some abuse of notation, we will also use G_k to denote the sigma-algebra generated by these variables. The set (91) contains all the outputs of the algorithm (78) immediately *before* (78d) in iteration k.

Now, the actions of the matrix $\mathbf{V}$ in the recursions (78) are through the matrix-vector multiplications (78a) and (78d). Hence, if we define the matrices

$$\mathbf{A}_k := [\mathbf{P}_k\ \mathbf{V}_{k-1}], \quad \mathbf{B}_k := [\mathbf{U}_k\ \mathbf{Q}_{k-1}], \tag{92}$$

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

$$\mathbf{A}_k = \mathbf{V}\mathbf{B}_k. \tag{93}$$

Hence, the conditional distribution of $\mathbf{V}$ given G_k is precisely the uniform distribution on the set of orthogonal matrices satisfying (93). The matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ are of dimensions $N \times s$ where $s = 2k + 1$. From Lemma 4, this conditional distribution is given by

$$\mathbf{V}|_{G_k} \overset{d}{=} \mathbf{A}_k(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k)^{-1}\mathbf{B}_k^\mathsf{T} + \mathbf{U}_{\mathbf{A}_k^\perp}\widetilde{\mathbf{V}}\mathbf{U}_{\mathbf{B}_k^\perp}^\mathsf{T}, \tag{94}$$

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.

Using (94) we can write $\mathbf{q}_k$ in (78d) as a sum of two terms

$$\mathbf{q}_k = \mathbf{V}^\mathsf{T}\mathbf{v}_k = \mathbf{q}_k^{\mathrm{det}} + \mathbf{q}_k^{\mathrm{ran}}, \tag{95}$$

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

$$\mathbf{q}_k^{\mathrm{det}} = \mathbf{B}_k(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k)^{-1}\mathbf{A}_k^\mathsf{T}\mathbf{v}_k, \tag{96}$$

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

$$\mathbf{q}_k^{\mathrm{ran}} = \mathbf{U}_{\mathbf{B}_k^\perp}\widetilde{\mathbf{V}}^\mathsf{T}\mathbf{U}_{\mathbf{A}_k^\perp}^\mathsf{T}\mathbf{v}_k. \tag{97}$$

The next few lemmas will evaluate the asymptotic distributions of the two terms in (95).

Lemma 6. *Under the induction hypothesis $H_{k,k-1}$, there exists constants $\beta_{k,0}, \ldots, \beta_{k,k-1}$ such that the components of $\mathbf{q}_k^{\mathrm{det}}$ along with $(\mathbf{q}_0, \ldots, \mathbf{q}_{k-1})$ converge empirically as*

$$\lim_{N\to\infty} \left\{w_n^q, q_{0n}, \ldots, q_{k-1,n}, q_{kn}^{\mathrm{det}})\right\} \overset{PL(2)}{=} (W^q, Q_0, \ldots, Q_{k-1}, Q_k^{\mathrm{det}}), \tag{98}$$

where Q_ℓ, $\ell = 0, \ldots, k-1$ are the Gaussian random variables in induction hypothesis (87) *and $Q_k^{\rm det}$ is a linear combination,*

$$Q_k^{\rm det} = \beta_{k0} Q_0 + \cdots + \beta_{k,k-1} Q_{k-1}. \tag{99}$$

Proof: We evaluate the asymptotic values of various terms in (96). Using the definition of $\mathbf{A}_k$ in (92),

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

We can then easily evaluate the asymptotic value of these terms as follows. For example, the asymptotic value of the (i,j) component of the matrix $\mathbf{P}_k^{\mathsf{T}}\mathbf{P}_k$ is given by

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N}\left[\mathbf{P}_k^{\mathsf{T}}\mathbf{P}_k\right]_{ij} &\overset{(a)}{=} \frac{1}{N}\mathbf{p}_i^{\mathsf{T}}\mathbf{p}_j \\ &= \frac{1}{N}\sum_{n=1}^{N} p_{in}p_{jn} \overset{(b)}{=} E(P_iP_j) \overset{(c)}{=} \left[\mathbf{Q}_k^p\right]_{ij}, \end{aligned}$$

where (a) follows since the i-th column of $\mathbf{P}_k$ is precisely the vector $\mathbf{p}_i$; (b) follows due to convergence assumption in (85); and in (c), we use $\mathbf{Q}_k^p$ to denote the covariance matrix of $(P_0, \ldots, P_k)$. Similarly

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

where $\mathbf{Q}_{k-1}^v$ has the components,

$$\left[\mathbf{Q}_{k-1}^v\right]_{ij} = \mathbb{E}\left[V_iV_j\right],$$

where V_i is the random variable in (89). Finally, the expectation for the cross-terms are given by

$$\begin{aligned} \mathbb{E}(V_iX_j) &\overset{(a)}{=} \mathbb{E}(g_p(P_i, W^p, \overline{\gamma}_{1i}, \overline{\alpha}_{1i})X_j) \\ &\overset{(b)}{=} \mathbb{E}\left[g_p'(P_i, W^p, \overline{\gamma}_{1i}, \overline{\alpha}_{1i})\right]\mathbb{E}(X_iX_j) \\ &\overset{(c)}{=} \mathbb{E}(X_iX_j)C_1(\overline{\alpha}_{1i})\left[\mathbb{E}\left(f_p'(P_i, W^p, \overline{\gamma}_{1i})\right) - \overline{\alpha}_{1i}\right] \\ &\overset{(d)}{=} 0, \end{aligned}$$

where (a) follows from (89); (b) follows from Stein's Lemma; and (c) follows from the definition of $g_p(\cdot)$ in (90); and (d) follows from (83a). The above calculations show that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k = \begin{bmatrix} \mathbf{Q}_k^p & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{k-1}^v \end{bmatrix}. \tag{100}$$

A similar calculation shows that

$$\lim_{N\to\infty} \frac{1}{N}\mathbf{A}_k^{\mathsf{T}}\mathbf{s}_k = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_k^s \end{bmatrix}, \tag{101}$$

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

$$\mathbf{b}_k^s = \left[\mathbb{E}(V_0V_k)\ \mathbb{E}(V_1V_k)\ \cdots \mathbb{E}(V_{k-1}V_k)\right]^{\mathsf{T}}. \tag{102}$$

Combining (100) and (101) shows that

$$\lim_{N\to\infty} (\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{A}_k^{\mathsf{T}}\mathbf{s}_k = \begin{bmatrix} \mathbf{0} \\ \beta_k \end{bmatrix}, \tag{103}$$

where

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

Therefore,

$$\begin{aligned} \mathbf{q}_k^{\rm det} &= \mathbf{B}_k(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k)^{-1}\mathbf{A}_k^{\mathsf{T}}\mathbf{v}_k \\ &= \left[\mathbf{U}_k\ \mathbf{Q}_{k-1}\right]\begin{bmatrix} \mathbf{0} \\ \beta_k \end{bmatrix} + O(1/N) \\ &= \sum_{\ell=0}^{k-1} \beta_{k\ell}\mathbf{q}_\ell + O(1/N), \end{aligned} \tag{104}$$

where the term $O(1/N)$ means a vector sequence, $\boldsymbol{\xi}(N) \in \mathbb{R}^N$ such that

$$\lim_{N\to\infty} \frac{1}{N}\|\boldsymbol{\xi}(N)\|^2 = 0.$$

A continuity argument then shows (98). ∎

Lemma 7. *Under the induction hypothesis $H_{k,k-1}$, the following limit holds almost surely*

$$\lim_{N\to\infty} \frac{1}{N}\|\mathbf{U}_{\mathbf{A}_k^\perp}^{\mathsf{T}}\mathbf{s}_k\|^2 = \rho_k, \tag{105}$$

for some constant $\rho_k \geq 0$.

Proof: From (92), the matrix $\mathbf{A}_k$ has $s = 2k+1$ columns. From Lemma 4, $\mathbf{U}_{\mathbf{A}_k^\perp}$ is an orthonormal basis of $N - s$ in the $\mathrm{Range}(\mathbf{A}_k)^\perp$. Hence, the energy $\|\mathbf{U}_{\mathbf{A}_k^\perp}\mathbf{s}_k\|^2$ is precisely

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

Using similar calculations as the previous lemma, we have

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

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

Lemma 8. *Under the induction hypothesis $H_{k,k-1}$, the components of the "random" part $\mathbf{q}_k^{\rm ran}$ along with the components of $(\mathbf{w}^q, \mathbf{q}_0, \ldots, \mathbf{q}_{k-1})$ almost surely converge empirically as*

$$\begin{aligned} &\lim_{N\to\infty} \{(w_n^q, q_{0n}, \ldots, q_{k-1,n}, q_{kn}^{\rm ran}\} \\ &\qquad \overset{PL(2)}{=} (W^q, Q_0, \ldots, Q_{k-1}, U_k), \end{aligned} \tag{106}$$

where $U_k \sim \mathcal{N}(0, \rho_k)$ is a Gaussian random variable independent of $(W^q, Q_0, \ldots, Q_{k-1})$ and ρ_k is the constant in Lemma 7.

Proof: This is a direct application of Lemma 5. Let $\mathbf{x} = \widetilde{\mathbf{V}}^{\mathsf{T}}\mathbf{U}_{\mathbf{A}_k^\perp}^{\mathsf{T}}\mathbf{s}_k$ so that

$$\mathbf{q}_k^{\rm det} = \mathbf{U}_{\mathbf{B}_k^\perp}\mathbf{x}_k.$$

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, the conditional distribution $\mathbf{x}_k$ given G_k is orthogonally invariant in that

$$\mathbf{U}\mathbf{x}_k|_{G_k} \overset{d}{=} \mathbf{x}_k|_{G_k},$$

for any orthogonal matrix $\mathbf{U}$. Lemma 7 also shows that

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

almost surely. The limit (106) now follows from Lemma 5. ∎

Using the partition (95) and Lemmas 6 and 8, the components of $(\mathbf{w}^q, \mathbf{q}_0, \ldots, \mathbf{q}_k)$ almost surely converge empirically as

$$\begin{aligned}
&\lim_{N\to\infty} \{(w_n^q, q_{0n}, \ldots, q_{kn})\} \\
&\overset{PL(2)}{=} \lim_{N\to\infty} \{(w_n^q, q_{0n}, \ldots, q_{kn}^{\rm det} + q_{kn}^{\rm ran})\} \\
&\overset{PL(2)}{=} (W^q, Q_0, \ldots, Q_k),
\end{aligned}$$

where Q_k is the random variable

$$Q_k = \beta_{k0}Q_0 + \cdots + \beta_{k,k-1}Q_{k-1} + U_k.$$

Since $(Q_0, \ldots, Q_{k-1})$ is jointly Gaussian and U_k is Gaussian independent of $(Q_0, \ldots, Q_{k-1})$ we have that $(Q_0, \ldots, Q_k)$ is Gaussian. This proves (87).

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,

$$\lim_{N\to\infty} \gamma_{2k} = \lim_{N\to\infty} \Gamma_1(\gamma_{1k}, \alpha_{1k}) = \overline{\gamma}_{2k}. \tag{107}$$

In addition, since we have assumed that $\mathbf{f}_q'(\mathbf{q}, \mathbf{w}^q, \gamma_1)$ is Lipschitz continuous in $(\mathbf{q}, \mathbf{w}^q)$ and continuous in γ_1,

$$\begin{aligned}
\lim_{N\to\infty} \alpha_{2k} &= \lim_{N\to\infty} \langle \mathbf{f}_q'(\mathbf{q}_k, \mathbf{w}^q, \gamma_{1k})\rangle \\
&= \mathbb{E}\left[f_q'(Q_k, W^q, \overline{\gamma}_{1k})\right] = \overline{\alpha}_{1k}.
\end{aligned} \tag{108}$$

The limits (107) and (108) prove (88).

Finally, we need to show that $\mathbb{E}(Q_k^2) = \tau_{2k}$ is the variance from the SE equations.

$$\begin{aligned}
\mathbb{E}(Q_k^2) &\overset{(a)}{=} \lim_{N\to\infty} \frac{1}{N}\|\mathbf{q}_k\|^2 \\
&\overset{(b)}{=} \lim_{N\to\infty} \frac{1}{N}\|\mathbf{v}_k\|^2 \\
&\overset{(c)}{=} \mathbb{E}\left[g_p(P_k, W^p, \overline{\gamma}_{1k}, \overline{\alpha}_{1k})\right] \\
&\overset{(d)}{=} C_1^2(\overline{\alpha}_{1k})\mathbb{E}\left[\left(f_p(P_k, W^p, \overline{\gamma}_{1k}) - \overline{\alpha}_{1k}P_k\right)^2\right] \\
&= C_1^2(\overline{\alpha}_{1k})\Big\{\mathbb{E}\left[f_p^2(P_k, W^p, \overline{\gamma}_{1k})\right] \\
&\quad - 2\overline{\alpha}_{1k}\mathbb{E}\left[P_k f_p(P_k, W^p, \overline{\gamma}_{1k})\right] + \overline{\alpha}_{1k}^2\mathbb{E}\left[P_k^2\right]\Big\} \\
&\overset{(e)}{=} C_1^2(\overline{\alpha}_{1k})\Big\{\mathbb{E}\left[f_p^2(P_k, W^p, \overline{\gamma}_{1k})\right] \\
&\quad - 2\overline{\alpha}_{1k}\tau_{1k}\mathbb{E}\left[f_p'(P_k, W^p, \overline{\gamma}_{1k})\right] + \overline{\alpha}_{1k}^2\tau_{1k}\Big\} \\
&\overset{(f)}{=} C_1^2(\overline{\alpha}_{1k})\left\{\mathbb{E}\left[f_p^2(P_k, W^p, \overline{\gamma}_{1k})\right] - \overline{\alpha}_{1k}^2\tau_{1k}\right\} \\
&\overset{(g)}{=} \tau_{2k},
\end{aligned} \tag{109}$$

where (a) follows from the fact that the components of $\mathbf{q}_k$ converge empirically to Q_k; (b) follows from (78d) and the fact that $\mathbf{V}$ is orthogonal; (c) follows from the limit (89); and (d) follows from (90); (e) follows from Stein's Lemma and the fact that $\mathbb{E}[P_k^2] = \tau_{1k}$; (f) follows from the definition of $\overline{\alpha}_{1k}$ in (83a); and (g) follows from (83b). Thus, $\mathbb{E}[Q_k^2] = \tau_{2k}$, and we have proven the implication $H_{k,k-1} \Rightarrow H_{k,k}$.

Appendix G
Proof of Theorem 1

Theorem 1 is essentially a special case of Theorem 4. We need to simply rewrite the recursions in Algorithm 3 in the form (78). To this end, define the error terms

$$\mathbf{p}_k := \mathbf{r}_{1k} - \mathbf{x}^0, \quad \mathbf{v}_k := \mathbf{r}_{2k} - \mathbf{x}^0, \tag{110}$$

and their transforms,

$$\mathbf{u}_k := \mathbf{V}^{\mathsf{T}}\mathbf{p}_k, \quad \mathbf{q}_k := \mathbf{V}^{\mathsf{T}}\mathbf{v}_k. \tag{111}$$

Also, define the disturbance terms

$$\mathbf{w}^q := (\boldsymbol{\xi}, \mathbf{s}), \quad \mathbf{w}^p := \mathbf{x}^0, \quad \boldsymbol{\xi} := \mathbf{U}^{\mathsf{T}}\mathbf{w}, \tag{112}$$

and the componentwise update functions

$$f_q(q, (\xi, s), \gamma_2) := \frac{\gamma_w s\xi + \gamma_2 q}{\gamma_w s^2 + \gamma_2}, \tag{113a}$$

$$f_p(p, x^0, \gamma_1) = g_1(p + x^0, \gamma_1) - x^0. \tag{113b}$$

With these definitions, we claim that the outputs satisfy the recursions:

$$\mathbf{p}_k = \mathbf{V}\mathbf{u}_k \tag{114a}$$

$$\alpha_{1k} = \langle \mathbf{f}_p'(\mathbf{p}_k, \mathbf{x}^0, \gamma_{1k})\rangle, \quad \gamma_{2k} = \frac{(1-\alpha_{1k})\gamma_{1k}}{\alpha_{1k}} \tag{114b}$$

$$\mathbf{v}_k = \frac{1}{1-\alpha_{1k}}\left[\mathbf{f}_p(\mathbf{p}_k, \mathbf{x}^0, \gamma_{1k}) - \alpha_{1k}\mathbf{p}_k\right] \tag{114c}$$

$$\mathbf{q}_k = \mathbf{V}^{\mathsf{T}}\mathbf{v}_k \tag{114d}$$

$$\alpha_{2k} = \langle \mathbf{f}_q'(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k})\rangle, \quad \gamma_{1,k+1} = \frac{(1-\alpha_{2k})\gamma_{2k}}{\alpha_{2k}} \tag{114e}$$

$$\mathbf{u}_{k+1} = \frac{1}{1-\alpha_{2k}}\left[\mathbf{f}_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) - \alpha_{2k}\mathbf{q}_k\right] \tag{114f}$$

Before we prove (114), we can see that (114) is a special case of the general recursions in (78) if we define

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

It is also straightforward to verify the continuity assumptions in Theorem 4. The assumption of Theorem 1 states that $\overline{\alpha}_{ik} \in (0,1)$. Since $\overline{\gamma}_{10} > 0$, $\overline{\gamma}_{ik} > 0$ for all k and i. Therefore, $C_i(\alpha_i)$ and $\Gamma_i(\gamma_i, \alpha_i)$ are continuous at all points $(\gamma_i, \alpha_i) = (\overline{\gamma}_{ik}, \overline{\alpha}_{ik})$. Also, since $s \in [0, S_{max}]$ and $\gamma_{2k} > 0$ for all k, the function $f_q(q, (\xi, s), \gamma_2)$ in (113) is uniformly Lipschitz continuous in (q, ξ, s) at all $\gamma_2 = \overline{\gamma}_{2k}$. Similarly, since the denoiser function $g_1(r_1, \gamma_1)$ is assumed be to uniformly Lipschitz continuous in r_1 at all $\gamma_1 = \overline{\gamma}_{1k}$, so is the function $f_p(r_1, x^0, \gamma_1)$ in (113b). Hence all the conditions of Theorem 4 are satisfied. The SE equations (50) immediately from the general SE equations (83). In addition, the limits (45) and and (48) are special cases of the limits (85) and (87). This proves Theorem 1.

So, it remains only to show that the updates in (114) indeed hold. Equations (114a) and (114d) follow immediately from the definitions (110) and (111). Next, observe that we can

rewrite the LMMSE estimation function (24) as

$$\begin{aligned}
&\mathbf{g}_2(\mathbf{r}_{2k},\gamma_{2k})\\
&\overset{(a)}{=}\left(\gamma_w\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_w\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}^0+\gamma_w\mathbf{A}^\mathsf{T}\mathbf{w}+\gamma_{2k}\mathbf{r}_{2k}\right)\\
&\overset{(b)}{=}\mathbf{x}^0+\left(\gamma_w\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_{2k}(\mathbf{r}_{2k}-\mathbf{x}^0)+\gamma_w\mathbf{A}^\mathsf{T}\mathbf{w}\right)\\
&\overset{(d)}{=}\mathbf{x}^0+\mathbf{V}\left(\gamma_w\mathbf{S}^2+\gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_{2k}\mathbf{q}_k+\mathbf{S}\boldsymbol{\xi}\right),\\
&\overset{(d)}{=}\mathbf{x}^0+\mathbf{V}\mathbf{f}_q(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k}),
\end{aligned}\tag{115}$$

where (a) follows by substituting (28) into (24); (b) is a simple algebraic manipulation; (c) follows from the SVD definition (29) and the definitions $\boldsymbol{\xi}$ in (112) and $\mathbf{q}_k$ in (111); and (d) follows from the definition of componentwise function $f_q(\cdot)$ in (113a). Therefore, the divergence α_{2k} satisfies

$$\begin{aligned}
\alpha_{2k}&\overset{(a)}{=}\frac{1}{N}\mathrm{Tr}\left[\frac{\partial\mathbf{g}_2(\mathbf{r}_{2k},\gamma_{2k})}{\partial\mathbf{r}_{2k}}\right]\\
&\overset{(b)}{=}\frac{1}{N}\mathrm{Tr}\left[\mathbf{V}\mathrm{Diag}(\mathbf{f}_q'(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k}))\frac{\partial\mathbf{q}_k}{\partial\mathbf{r}_{2k}}\right]\\
&\overset{(c)}{=}\frac{1}{N}\mathrm{Tr}\left[\mathbf{V}\mathrm{Diag}(\mathbf{f}_q'(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k}))\mathbf{V}^\mathsf{T}\right]\\
&\overset{(d)}{=}\langle\mathbf{f}_q'(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k})\rangle,
\end{aligned}\tag{116}$$

where (a) follows from line 12 of Algorithm 3 and (6)–(7); (b) follows from (115); (c) follows from (111); and (d) follows from $\mathbf{V}^\mathsf{T}\mathbf{V}=\mathbf{I}$ and (6)–(7). Also, from lines 13-14 of Algorithm 3,

$$\gamma_{1,k+1}=\eta_{2k}-\gamma_{2k}=\gamma_{2k}\left[\frac{1}{\alpha_{2k}}-1\right].\tag{117}$$

Equations (116) and (117) prove (114e). In addition,

$$\begin{aligned}
\mathbf{p}_{k+1}&\overset{(a)}{=}\mathbf{r}_{1,k+1}-\mathbf{x}^0\\
&\overset{(b)}{=}\frac{1}{1-\alpha_{2k}}\left[\mathbf{g}_2(\mathbf{r}_{2k},\gamma_{2k})-\alpha_{2k}\mathbf{r}_{2k}\right]-\mathbf{x}^0\\
&\overset{(c)}{=}\frac{1}{1-\alpha_{2k}}\left[\mathbf{x}^0+\mathbf{V}\mathbf{f}_q(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k})-\alpha_{2k}(\mathbf{x}^0+\mathbf{v}_k)\right]-\mathbf{x}^0\\
&\overset{(d)}{=}\frac{1}{1-\alpha_{2k}}\left[\mathbf{V}\mathbf{f}_q(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k})-\alpha_{2k}\mathbf{v}_k\right]\\
&\overset{(e)}{=}\mathbf{V}\left[\frac{1}{1-\alpha_{2k}}\left[\mathbf{f}_q(\mathbf{q}_k,\mathbf{w}^q,\gamma_{2k})-\alpha_{2k}\mathbf{q}_k\right]\right],
\end{aligned}\tag{118}$$

where (a) follows from (110); (b) follows from lines 11-15 of Algorithm 3; (c) follows from (115) and the definition of $\mathbf{v}_k$ in (110); (d) follows from collecting the terms with $\mathbf{x}^0$; and (e) follows from the definition $\mathbf{q}_k=\mathbf{V}^\mathsf{T}\mathbf{v}_k$ in (111). Combining (118) with $\mathbf{u}_{k+1}=\mathbf{V}^\mathsf{T}\mathbf{p}_{k+1}$ proves (114f).

The derivation for the updates for $\mathbf{v}_k$ are similar. First,

$$\alpha_{1k}\overset{(a)}{=}\langle\mathbf{g}_1'(\mathbf{r}_{1k},\gamma_{1k})\rangle\overset{(b)}{=}\langle\mathbf{f}_p'(\mathbf{p}_k,\mathbf{x}^0)\rangle,\tag{119}$$

where (a) follows from line 5 of Algorithm 3 and (b) follows from the vectorization of $\mathbf{f}_p(\cdot)$ in (113b) and the fact that $\mathbf{p}_k=\mathbf{r}_{1k}+\mathbf{x}^0$. Also, from lines 6-7 of Algorithm 3,

$$\gamma_{2k}=\eta_{1k}-\gamma_{1k}=\gamma_{1k}\left[\frac{1}{\alpha_{1k}}-1\right].\tag{120}$$

Equations (119) and (120) prove (114b). Also,

$$\begin{aligned}
\mathbf{v}_k&\overset{(a)}{=}\mathbf{r}_{2k}-\mathbf{x}^0\\
&\overset{(b)}{=}\frac{1}{1-\alpha_{1k}}\left[\mathbf{g}_1(\mathbf{r}_{1k},\gamma_{1k})-\alpha_{1k}\mathbf{r}_{1k}\right]-\mathbf{x}^0\\
&\overset{(c)}{=}\frac{1}{1-\alpha_{1k}}\left[\mathbf{f}_p(\mathbf{p}_k,\mathbf{x}^0,\gamma_{1k})+\mathbf{x}^0-\alpha_{1k}(\mathbf{p}_k+\mathbf{x}^0)\right]-\mathbf{x}^0\\
&\overset{(d)}{=}\frac{1}{1-\alpha_{1k}}\left[\mathbf{f}_p(\mathbf{p}_k,\mathbf{x}^0,\gamma_{1k})-\alpha_{1k}\mathbf{p}_k\right]
\end{aligned}\tag{121}$$

where (a) is the definition of $\mathbf{v}_k$ in (110); (b) follows from lines 4-8 of Algorithm 3; (c) follows from the vectorization of $f_p(\cdot)$ in (113b) and the definition of $\mathbf{p}_k$ in (110); and (d) follows from collecting the terms with $\mathbf{x}^0$. This proves (114c). All together, we have proven (114) and the proof is complete.

Appendix H
Proof of Theorem 2

We use induction. Suppose that, for some k, $\overline{\gamma}_{1k}=\tau_{1k}^{-1}$. From (50a), (44) and (36),

$$\overline{\alpha}_{1k}=\overline{\gamma}_{1k}\mathcal{E}_1(\overline{\gamma}_{1k}).\tag{122}$$

Hence, from (50b), $\overline{\eta}_{1k}^{-1}=\mathcal{E}_1(\overline{\gamma}_{1k})$ and $\overline{\gamma}_{2k}=\overline{\eta}_{1k}-\overline{\gamma}_{1k}$. Also,

$$\begin{aligned}
\tau_{2k}&\overset{(a)}{=}\frac{1}{(1-\overline{\alpha}_{1k})^2}\left[\mathcal{E}_1(\overline{\gamma}_{1k},\tau_{1k})-\overline{\alpha}_{1k}^2\tau_{1k}\right]\\
&\overset{(b)}{=}\frac{1}{(1-\overline{\gamma}_{1k}\mathcal{E}_1(\overline{\gamma}_{1k}))^2}\left[\mathcal{E}_1(\overline{\gamma}_{1k},\tau_{1k})-\overline{\gamma}_{1k}\mathcal{E}_1^2(\overline{\gamma}_{1k})\right]\\
&\overset{(c)}{=}\frac{\mathcal{E}_1(\overline{\gamma}_{1k},\tau_{1k})}{1-\overline{\gamma}_{1k}\mathcal{E}_1(\overline{\gamma}_{1k})}\\
&\overset{(d)}{=}\frac{1}{\overline{\eta}_{1k}-\overline{\gamma}_{1k}},
\end{aligned}$$

where (a) follows from (50c); (b) follows from (122) and the matched condition $\overline{\gamma}_{1k}=\tau_{1k}^{-1}$; (c) follows from canceling terms in the fraction and (d) follows from the fact that $\overline{\eta}_{1k}^{-1}=\mathcal{E}_1(\overline{\gamma}_{1k})$ and $\overline{\gamma}_{1k}=\overline{\eta}_{1k}/\overline{\alpha}_{1k}$. This proves (56a). A similar argument shows that (56b) holds if $\overline{\gamma}_{2k}=\tau_{2k}^{-1}$. Finally, (56c) follows from (56) and (55).

References

[1] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, pp. 1–14, 2016.

[2] 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.

[3] ——, "Message passing algorithms for compressed sensing I: Motivation and construction," in *Proc. Info. Theory Workshop*, Jan. 2010, pp. 1–5.

[4] 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.

[5] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Ann. Appl. Prob.*, vol. 25, no. 2, pp. 753–822, 2015.

[6] C. Rush and R. Venkataramanan, "Finite-sample analysis of approximate message passing," in *Proc. IEEE ISIT*, 2016, pp. 755–759.

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

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

[9] 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.

[10] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Swept approximate message passing for sparse estimation," in *Proc. ICML*, 2015, pp. 1123–1132.

[11] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, "Inference for generalized linear models via alternating directions and Bethe free energy minimization," in *Proc. IEEE ISIT*, 2015, pp. 1640–1644.

[12] M. Opper and O. Winther, "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling," *Physical Review E*, vol. 64, no. 5, p. 056131, 2001.

[13] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Machine Learning Research*, vol. 9, pp. 759–813, Sep. 2008.

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

[15] Y. Kabashima and M. Vehkaperä, "Signal recovery using expectation consistent approximation for linear observations," in *Proc. IEEE ISIT*, 2014, pp. 226–230.

[16] A. K. Fletcher, M. Sahraee-Ardakan, S. Rangan, and P. Schniter, "Expectation consistent approximate inference: Generalizations and convergence," *Proc. ISIT*, 2016.

[17] B. Çakmak, O. Winther, and B. H. Fleury, "S-AMP: Approximate message passing for general matrix ensembles," in *Proc. IEEE Information Theory Workshop (ITW)*, 2014, pp. 192–196.

[18] ——, "S-AMP for non-linear observation models," in *Proc. IEEE ISIT*, 2015.

[19] J. Ma and L. Ping, "Orthogonal AMP," *arXiv:1602.06509*, 2016.

[20] A. M. Tulino, G. Caire, S. Verdú, and S. Shamai, "Support recovery with sparsely sampled free random matrices," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4243–4271, 2013.

[21] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.

[22] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Theory*, Saint Petersburg, Russia, Jul.–Aug. 2011, pp. 2174–2178.

[23] A. Montanari, "Graphical model concepts in compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge Univ. Press, Jun. 2012, pp. 394–438.

[24] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.

[25] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[26] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of 'solvable model of a spin glass'," *Phil. Mag.*, vol. 35, pp. 983–990, 1977.

[27] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publ., 1988.

[28] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring Artificial Intelligence in the New Millennium*. San Francisco, CA: Morgan Kaufmann Publishers, 2003, pp. 239–269.

[29] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[30] M. Seeger, "Expectation propagation for exponential families," *EPFL-REPORT-161464*, 2005.

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

[32] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborová, "Variational free energies for compressed sensing," in *Proc. IEEE ISIT*, Jul. 2014, pp. 1499–1503.

[33] A. K. Fletcher and P. Schniter, "Learning and free energies for vector approximate message passing," *arXiv preprint*, 2016.

[34] "Generalized approximate message passing," SourceForge.net project GAMPmatlab, available on-line at `http://gampmatlab.sourceforge.net/`.

[35] H. J. Brascamp and E. H. Lieb, "On extensions of the Brunn-Minkowski and Prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation," in *Inequalities*. Springer, 2002, pp. 441–464.

[36] A. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Found. Trends Commun. Info. Thy.*, vol. 1, pp. 1–182, 2004.

[37] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact," *Proc. ISIT*, 2016.

[38] J. Barbier, M. Dia, N. Macris, and F. Krzakala, "The mutual information in random linear estimation," *arXiv preprint arXiv:1607.02335*, 2016.