Estimation in Rotationally Invariant Generalized Linear Models via Approximate Message Passing

Ramji Venkataramanan [1] **Kevin Kögler** [2] **Marco Mondelli** [2]

Abstract

We consider the problem of signal estimation in generalized linear models defined via rotationally invariant design matrices. Since these matrices can have an arbitrary spectral distribution, this model is well suited for capturing complex correlation structures which often arise in applications. We propose a novel family of approximate message passing (AMP) algorithms for signal estimation, and rigorously characterize their performance in the high-dimensional limit via a state evolution recursion. Our rotationally invariant AMP has complexity of the same order as the existing AMP derived under the restrictive assumption of a Gaussian design; our algorithm also recovers this existing AMP as a special case. Numerical results showcase a performance close to Vector AMP (which is conjectured to be Bayes-optimal in some settings), but obtained with a much lower complexity, as the proposed algorithm does not require a computationally expensive singular value decomposition.

1. Introduction

We consider the problem of estimating a $d-$dimensional signal $\boldsymbol{x}^* \in \mathbb{R}^d$ from an observation $\boldsymbol{y} \in \mathbb{R}^n$ obtained via a generalized linear model (GLM) (McCullagh & Nedler, 1989). Specifically, given a design matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ with rows $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{R}^d$, the observation $\boldsymbol{y} \equiv (y_1, \dots, y_n)$ is generated as

$$y_i = q(\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle, \varepsilon_i), \quad \text{for } i = 1, \dots, n, \tag{1}$$

where $\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle = \boldsymbol{a}_i^\mathsf{T} \boldsymbol{x}^*$ denotes the Euclidean inner product, $\boldsymbol{\varepsilon} \equiv (\varepsilon_1, \dots, \varepsilon_n)$ is a noise vector and $q : \mathbb{R}^2 \to \mathbb{R}$ is a known function. The model (1) covers many widely studied problems in statistical estimation and signal processing: examples include linear regression (Donoho, 2006; Eldar & Kutyniok, 2012) ($y_i = \langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle + \varepsilon_i$), phase retrieval (Shechtman et al., 2015; Fannjiang & Strohmer, 2020) ($y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle|^2 + \varepsilon_i$), and 1-bit compressed sensing (Boufounos & Baraniuk, 2008) ($y_i = \text{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle + \varepsilon_i)$).

A range of estimators based on convex relaxations, spectral methods, and non-convex methods have been proposed for specific instances of GLMs, such as sparse linear regression (Tibshirani, 1996; Candés & Tao, 2007; Hastie et al., 2019), phase retrieval (Netrapalli et al., 2013; Candès et al., 2013; 2015; Mondelli & Montanari, 2019; Luo et al., 2019; Lu & Li, 2020) and one-bit compressed sensing (Plan & Vershynin, 2012; 2013; Jacques et al., 2013). Most of these techniques are generic and can incorporate certain constraints like sparsity, but they are not well-equipped to exploit specific information about $\boldsymbol{x}^*$, e.g., a known signal prior.

Approximate message passing (AMP) is a family of iterative algorithms that can be tailored to take advantage of structural information known about the signal. AMP algorithms were first proposed for estimation in linear models (Kabashima, 2003; Bayati & Montanari, 2012; 2011; Donoho et al., 2009; Krzakala et al., 2012; Maleki et al., 2013), but have since been applied to a range of statistical estimation problems, including generalized linear models (Barbier et al., 2019; Ma et al., 2019; Maillard et al., 2020; Mondelli & Venkataramanan, 2021a; Rangan, 2011; Schniter & Rangan, 2014; Sur & Candès, 2019) and low-rank matrix estimation (Deshpande & Montanari, 2014; Fletcher & Rangan, 2018; Kabashima et al., 2016; Lesieur et al., 2017; Montanari & Venkataramanan, 2021; Barbier et al., 2020). An attractive feature of AMP is that under suitable model assumptions, its performance in the high-dimensional limit is precisely characterized by a succinct deterministic recursion called *state evolution* (Bayati & Montanari, 2011; Bolthausen, 2014; Javanmard & Montanari, 2013). Using the state evolution analysis, it has been proved that AMP achieves Bayes-optimal performance for some models (Deshpande & Montanari, 2014; Donoho et al., 2013; Montanari & Venkataramanan, 2021; Barbier et al., 2019),

[1]University of Cambridge, United Kingdom [2]ISTA, Austria. Correspondence to: Ramji Venkataramanan <rv285@cam.ac.uk>, Marco Mondelli <marco.mondelli@ist.ac.at>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

and a conjecture from statistical physics posits that AMP is optimal among polynomial-time algorithms. The above works, including the original GAMP algorithm (Rangan, 2011) for estimation in GLMs, all assume that the matrix $\boldsymbol{A}$ defining the model is i.i.d. Gaussian. While some of these results can be generalized to the broader class of i.i.d. sub-Gaussian matrices via universality arguments (Bayati et al., 2015; Chen & Lam, 2021), the i.i.d. assumption limits the applicability of AMP in practice. In this paper, we present an AMP algorithm for generalized linear models defined via a rotationally invariant design matrix $\boldsymbol{A}$. The class of rotational invariant matrices includes Gaussian matrices, but is much bigger. Rotational invariance only imposes that the orthogonal matrices in the singular value decomposition of $\boldsymbol{A}$ are uniformly random, and allows for arbitrary singular values. Hence, $\boldsymbol{A}$ is able to capture a complex correlation structure, which is typical in applications.

Main contributions. We propose an AMP algorithm for GLMs with rotationally invariant design matrices. The algorithm, which we call RI-GAMP, uses a pair of multivariate 'denoising' functions to produce an updated signal estimate in each iteration. The iterates depend on the free cumulants of the spectral distribution of the design matrix. Assuming that these free cumulants are known (e.g., via the spectral distribution), the complexity of RI-GAMP is of the same order as that of the standard GAMP algorithm (Rangan, 2011). Moreover, when the design matrix is i.i.d. Gaussian, RI-GAMP reduces to standard GAMP. Our main technical contribution is a state evolution result for RI-GAMP (Theorem 3.1), which gives a rigorous characterization of its performance in the high-dimensional limit as $n, d \to \infty$ with a fixed ratio $\delta = n/d$, for a constant $\delta > 0$. We also present numerical simulation results for linear regression and 1-bit compressed sensing, showcasing the performance of RI-GAMP on both synthetic data and images. The performance of RI-GAMP closely matches that of Vector AMP (Rangan et al., 2019; Pandit et al., 2020) (which is conjectured to be Bayes-optimal in some settings), but obtained with much lower complexity, as RI-GAMP does not require computing the singular value decomposition of $\boldsymbol{A}$.

RI-GAMP offers a flexible framework to analyze other estimators for GLMs, e.g., spectral methods. Standard GAMP has been used as a proof technique to study the distributional properties of linear and spectral estimators under Gaussian model assumptions (Mondelli et al., 2021; Mondelli & Venkataramanan, 2021a). An exciting research direction is to use RI-GAMP to analyze spectral estimators for rotationally invariant GLMs in the high-dimensional limit.

Proof idea. The key idea in establishing the state evolution result is to design an auxiliary AMP algorithm whose iterates are close to those of RI-GAMP. The auxiliary AMP is an instance of the abstract AMP iteration for rotationally invariant matrices analyzed in (Fan, 2021; Zhong et al., 2021), hence a state evolution result can be obtained for it. We then show that each iterate of RI-GAMP is close to one of the auxiliary AMP, and use this fact to translate the state evolution result for the latter to the RI-GAMP. We emphasize that the auxiliary AMP only serves as a proof technique. Indeed, it is initialized using the unknown signal $\boldsymbol{x}^*$, and therefore cannot be used for estimation.

RI-GAMP vs. Vector AMP. Vector AMP (VAMP) is an iterative algorithm (based on Expectation Propagation) for estimation in rotationally invariant linear (Rangan et al., 2019; Takeuchi, 2020; 2021b) and generalized linear models (Schniter et al., 2016; Pandit et al., 2020). Like RI-GAMP, VAMP can be tailored to take advantage of prior information about the signal and its performance can be characterized by a state evolution recursion. It is also shown in Rangan et al. (2019); Pandit et al. (2020) that the asymptotic estimation error of VAMP (with optimal denoising functions) coincides with the replica prediction for the Bayes-optimal error whenever the state evolution recursion has a unique fixed point.

The RI-GAMP algorithm proposed here is distinct from VAMP, and the associated state evolution recursions are also different. Let us highlight some attractive features of RI-GAMP. First, RI-GAMP does not require the computationally expensive ($O(d^3)$) singular value decomposition used in VAMP. Instead, it uses the free cumulants of the design matrix which can be estimated in $O(d^2)$ time (details on p.4). We confirm via numerical simulations that the complexity advantage of RI-GAMP over VAMP is significant and increases with the problem dimension (see Figure 4).

The state evolution result (Theorem 3.1) for RI-GAMP holds under mild assumptions on the denoising functions (see **(A1)** on p.5), while the analysis for VAMP requires slightly stronger conditions, e.g., the denoising functions and their derivatives need to be uniformly Lipschitz continuous. The numerical results in Section 4 show that the performance of RI-GAMP matches that of VAMP, except near parameter values corresponding to a phase transition in the estimation error.

Other related work. Orthogonal AMP (Ma & Ping, 2017) is an algorithm equivalent to VAMP for estimation in rotationally invariant linear models. Çakmak et al. (2016) proposed a variant of Expectation Propagation (an algorithm closely related to VAMP) for rotationally invariant GLMs. Ma et al. (2021) recently studied the performance of Expectation Propagation for rotationally invariant GLMs, and analyzed the impact of the spectrum on the estimation performance. VAMP has also been used to obtain the asymptotic risk of convex penalized estimators for rotationally invariant

GLMs (Gerbelot et al., 2020a;b). A few lower complexity alternatives to VAMP have been proposed recently, including convolutional AMP (Takeuchi, 2021a), Memory AMP for linear models (Liu et al., 2020), and Generalized Memory AMP for GLMs (Tian et al., 2021). The phase retrieval problem (which is a special case of a GLM) has been studied for design matrices with orthogonal columns, a model distinct from the rotationally invariant one considered here (Dudeja et al., 2020a;b). Finally, we mention that AMP has also been applied to low-rank matrix estimation with rotationally invariant noise (Opper et al., 2016; Çakmak & Opper, 2019; Fan, 2021; Zhong et al., 2021; Mondelli & Venkataramanan, 2021b).

2. Preliminaries

Notation and definitions. For $n \in \mathbb{N}$, we use the shorthand $[n]$ to denote the set $\{1, \dots, n\}$. Given a vector $\boldsymbol{x}$, we write $\|\boldsymbol{x}\|$ for its ℓ_2 norm. All vectors are treated as column vectors. Given $\boldsymbol{x} = (x_1, \dots, x_d)$, we denote by $\langle \boldsymbol{x} \rangle$ its empirical average $\frac{1}{d}\sum_{i=1}^d x_i$. The empirical distribution of $\boldsymbol{x}$ is given by $\frac{1}{d}\sum_{i=1}^d \delta_{x_i}$, where δ_{x_i} denotes a Dirac delta mass on x_i. Similarly, the joint empirical distribution of the rows of a matrix $(\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^t) \in \mathbb{R}^{d\times t}$ is $\frac{1}{d}\sum_{i=1}^d \delta_{(x_i^1, \dots, x_i^t)}$. Given a matrix $\boldsymbol{A}$, we denote by $(\boldsymbol{A})_{i,j}$ its (i,j)-th element. For a square matrix $\boldsymbol{M}$, we follow the convention that $\boldsymbol{M}^0$ is the identity matrix.

W_2 convergence. We write $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^k) \xrightarrow{W_2} (X_1, \dots, X_k)$ for the convergence in Wasserstein-2 distance of the joint empirical distribution of the rows of $(\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^k) \in \mathbb{R}^{d\times k}$ to the law of the random vector $(X_1, \dots, X_k)$. Equivalently (Feng et al., 2021)[Corollary 7.4], for any $L > 0$ and function $\psi : \mathbb{R}^k \to \mathbb{R}$ that satisfies

$$|\psi(\boldsymbol{u}) - \psi(\boldsymbol{v})| \le L\|\boldsymbol{u} - \boldsymbol{v}\|\,(1 + \|\boldsymbol{u}\| + \|\boldsymbol{v}\|) \tag{2}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^k$, we have

$$\lim_{d\to\infty} \frac{1}{d}\sum_{i=1}^d \psi(x_i^1, \dots, x_i^k) = \mathbb{E}\{\psi(X_1, \dots, X_k)\}. \tag{3}$$

A function satisfying (2) for some fixed $L > 0$ is called a pseudo-Lipschitz function of order 2.

Rotationally invariant generalized linear models. The $n \times d$ design matrix $\boldsymbol{A}$ is formed by stacking the sensing vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$, i.e., $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]^\mathsf{T}$. We assume that $\boldsymbol{A}$ is bi-rotationally invariant in law, i.e., $\boldsymbol{A} = \boldsymbol{O}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{Q}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ is an $n \times d$ diagonal matrix containing the singular values of $\boldsymbol{A}$, and $\boldsymbol{O}$, $\boldsymbol{Q}$ are Haar orthogonal matrices independent of one another and also of $\boldsymbol{\Lambda}$. As $d \to \infty$, we assume that $n/d = \delta$, for some constant $\delta > 0$. The matrix $\boldsymbol{A}$ is independent of the signal $\boldsymbol{x}^* \in \mathbb{R}^d$, and the noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. The observation $\boldsymbol{y} \in \mathbb{R}^n$ is generated according to (1). We assume that there exist random variables X_*, ε with finite second moments such that $\boldsymbol{x}^* \xrightarrow{W_2} X_*$, and $\boldsymbol{\varepsilon} \xrightarrow{W_2} \varepsilon$ as $n \to \infty$. Furthermore, we assume that the empirical distribution of $\boldsymbol{\lambda}$ converges weakly almost surely to a compactly supported random variable Λ. We denote by $\{\kappa_{2k}\}_{k\ge1}$ the rectangular free cumulants associated with the moments $\{m_{2k}\}_{k\ge1}$, where m_{2k} is the k-th moment of the empirical eigenvalue distribution of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ (for details, see (34)-(35) in Appendix A). For $\delta > 1$, let $\tilde{\Lambda}$ be a mixture of Λ (w.p. $1/\delta$) and a point mass at 0 (w.p. $1 - 1/\delta$); for $\delta \le 1$, we set $\tilde{\Lambda} = \Lambda$. Then, the assumptions above imply that as $n, d \to \infty$, $m_{2k} \to \bar{m}_{2k} = \mathbb{E}\{\tilde{\Lambda}^{2k}\}$ and $\kappa_{2k} \to \bar{\kappa}_{2k}$ almost surely, where $\{\bar{m}_{2k}\}_{k\ge1}$ and $\{\bar{\kappa}_{2k}\}_{k\ge1}$ are the even moments and rectangular free cumulants of $\tilde{\Lambda}$.

3. Rotationally Invariant Generalized AMP

Algorithm. We propose the following rotationally invariant generalized AMP (RI-GAMP) to estimate $\boldsymbol{x}^* \in \mathbb{R}^d$ from the observation $\boldsymbol{y} \in \mathbb{R}^n$ and the design matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$. For $t \ge 1$, compute:

$$\boldsymbol{x}^t = \boldsymbol{A}^\mathsf{T}\boldsymbol{s}^t - \sum_{i=1}^{t-1} \beta_{ti}\,\hat{\boldsymbol{x}}^i, \qquad \hat{\boldsymbol{x}}^t = f_t(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t), \tag{4}$$

$$\boldsymbol{r}^t = \boldsymbol{A}\hat{\boldsymbol{x}}^t - \sum_{i=1}^{t} \alpha_{ti}\,\boldsymbol{s}^i, \qquad \boldsymbol{s}^{t+1} = h_{t+1}(\boldsymbol{r}^1, \dots, \boldsymbol{r}^t, \boldsymbol{y}). \tag{5}$$

The iteration is initialized with $\boldsymbol{s}^1 = h_1(\boldsymbol{y})$ and $\boldsymbol{x}^1 = \boldsymbol{A}^\mathsf{T}\boldsymbol{s}^1$. For $t \ge 1$, the functions $f_t : \mathbb{R}^t \to \mathbb{R}$ and $h_{t+1} : \mathbb{R}^{t+1} \to \mathbb{R}$ are applied row-wise to vectors and matrices. The scalars $\{\alpha_{ti}\}_{i=1}^t$ and $\{\beta_{ti}\}_{i=1}^{t-1}$ are obtained in terms of two lower-triangular matrices $\boldsymbol{\Psi}_{t+1}, \boldsymbol{\Phi}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$. These matrices are defined as

$$\boldsymbol{\Psi}_{t+1} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & \langle \partial_1 \hat{\boldsymbol{x}}^1 \rangle & 0 & \dots & 0 \\ 0 & \langle \partial_1 \hat{\boldsymbol{x}}^2 \rangle & \langle \partial_2 \hat{\boldsymbol{x}}^2 \rangle & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \langle \partial_1 \hat{\boldsymbol{x}}^t \rangle & \langle \partial_2 \hat{\boldsymbol{x}}^t \rangle & \dots & \langle \partial_t \hat{\boldsymbol{x}}^t \rangle \end{pmatrix},$$

$$\boldsymbol{\Phi}_{t+1} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \langle \partial_g \boldsymbol{s}^1 \rangle & 0 & 0 & \dots & 0 \\ \langle \partial_g \boldsymbol{s}^2 \rangle & \langle \partial_1 \boldsymbol{s}^2 \rangle & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \langle \partial_g \boldsymbol{s}^t \rangle & \langle \partial_1 \boldsymbol{s}^t \rangle & \dots & \langle \partial_{t-1} \boldsymbol{s}^t \rangle & 0 \end{pmatrix}, \tag{6}$$

where for $k \in [t]$, the vector $\partial_k \hat{\boldsymbol{x}}^t \in \mathbb{R}^d$ denotes the partial derivative $\partial_{x_k} f_t(x_1, \dots, x_t)$ applied row-wise to $\hat{\boldsymbol{x}}^t = f_t(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t)$. Similarly, the vector $\partial_k \boldsymbol{s}^t \in \mathbb{R}^n$ denotes the partial derivative $\partial_{r_k} h_t(r_1, \dots, r_{t-1}, y)$ applied

row-wise to $\boldsymbol{s}^t = h_t(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{t-1}, \boldsymbol{y})$. Recalling that $y = q(g, \varepsilon)$, we can view $h_t(r_1, \dots, r_{t-1}, q(g, \varepsilon))$ as a function of $(t+1)$ variables, with $\partial_g h_t(r_1, \dots, r_{t-1}, q(g, \varepsilon))$ being the partial derivative with respect to g. For $t > 1$, the vector $\partial_g \boldsymbol{s}^t$ denotes this partial derivative applied row-wise to $\boldsymbol{s}^t = h_t(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{t-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon}))$, and the vector $\partial_g \boldsymbol{s}^1$ is defined via the partial derivative $\partial_g h_1(q(g, \varepsilon))$. Next, recalling that $\{\kappa_{2k}\}_{k \geq 1}$ denote the rectangular free cumulants, define matrices $\boldsymbol{M}^{\alpha}_{t+1}, \boldsymbol{M}^{\beta}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ as:

$$\boldsymbol{M}^{\alpha}_{t+1} = \sum_{j=0}^{t+1} \kappa_{2(j+1)}\, \boldsymbol{\Psi}_{t+1} \Big(\boldsymbol{\Phi}_{t+1} \boldsymbol{\Psi}_{t+1} \Big)^j, \tag{7}$$

$$\boldsymbol{M}^{\beta}_{t+1} = \delta \sum_{j=0}^{t} \kappa_{2(j+1)}\, \boldsymbol{\Phi}_{t+1} \Big(\boldsymbol{\Psi}_{t+1} \boldsymbol{\Phi}_{t+1} \Big)^j. \tag{8}$$

Then, the coefficients $\{\alpha_{ti}\}_{i=1}^{t}$ and $\{\beta_{ti}\}_{i=1}^{t-1}$ in (4)-(5) are obtained from the last row of $\boldsymbol{M}^{\alpha}_{t+1}$ and $\boldsymbol{M}^{\beta}_{t+1}$ as:

$$(\alpha_{t1}, \dots, \alpha_{tt}) = \big((\boldsymbol{M}^{\alpha}_{t+1})_{t+1,2}\,, \dots, (\boldsymbol{M}^{\alpha}_{t+1})_{t+1,t+1} \big), \tag{9}$$

$$(\beta_{t1}, \dots, \beta_{t,t-1}) = \big((\boldsymbol{M}^{\beta}_{t+1})_{t+1,2}, \dots, (\boldsymbol{M}^{\beta}_{t+1})_{t+1,t} \big). \tag{10}$$

Estimating the free cumulants. From the definitions of $\{\alpha_{ti}\}_{i=1}^{t}$ and $\{\beta_{ti}\}_{i=1}^{t-1}$ above, it follows that we need the free cumulants $\{\kappa_{2(j+1)}\}_{j=0}^{t+1}$ to compute the first t iterates of RI-GAMP in (5). These free cumulants can be recursively computed from the moments $\{m_{2(j+1)}\}_{j=0}^{t+1}$ of the spectral distribution of $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}$, using the formula (35) in Appendix A. The moments $\{m_{2(j+1)}\}_{j=0}^{t+1}$ can be estimated in $O(d^2)$ time via the following simple algorithm proposed by Liu et al. (2020). Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, pick an independent standard Gaussian vector $\boldsymbol{s}^0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and for $k \geq 1$, compute $\boldsymbol{s}^k = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{s}^{k-1}$ for odd k and $\boldsymbol{s}^k = \boldsymbol{A}\boldsymbol{s}^{k-1}$ for even k. Then, $\frac{\|\boldsymbol{s}^k\|^2}{d}$ is a consistent estimate of the k-th moment of the spectral distribution of $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}$. Thus, the complexity of estimating the free cumulants is of the same order as one iteration of RI-GAMP.

State evolution. The coefficients $\{\alpha_{ti}\}_{i=1}^{t}$ and $\{\beta_{ti}\}_{i=1}^{t-1}$ play a crucial role in debiasing the AMP iterates, ensuring that their limiting empirical distributions are accurately captured by state evolution. Indeed, Theorem 3.1 shows that the joint empirical distribution of $(\boldsymbol{g}, \boldsymbol{r}^1, \dots, \boldsymbol{r}^t)$ converges to a $(t+1)$-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{\Sigma}}_{t+1})$. Similarly, the joint empirical distribution of $(\boldsymbol{x}^1 - \bar{\mu}_1 \boldsymbol{x}, \dots, \boldsymbol{x}^t - \bar{\mu}_t \boldsymbol{x})$ converges to a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{\Omega}}_t)$. We define the covariance matrices $\bar{\boldsymbol{\Omega}}_t, \bar{\boldsymbol{\Sigma}}_t \in \mathbb{R}^{t\times t}$ and the vector $\bar{\boldsymbol{\mu}}_t \equiv (\bar{\mu}_1, \dots, \bar{\mu}_t)$ recursively for $t \geq 1$, starting with

$$\begin{aligned} &\bar{\boldsymbol{\Sigma}}_1 = \bar{\kappa}_2 \mathbb{E}\{X_*^2\}, \quad \bar{\boldsymbol{\mu}}_1 = \delta \bar{\kappa}_2\, \mathbb{E}\{\partial_g h_1(q(G, \varepsilon))\}, \\ &\bar{\boldsymbol{\Omega}}_1 = \delta \bar{\kappa}_2 \mathbb{E}\{h_1(q(G, \varepsilon))^2\} \\ &\qquad + \delta \bar{\kappa}_4 \mathbb{E}\{X_*^2\} (\mathbb{E}\{\partial_g h_1(q(G, \varepsilon))\})^2, \end{aligned} \tag{11}$$

where $G \sim \mathcal{N}(0,\, \bar{\kappa}_2 \mathbb{E}\{X_*^2\})$ is independent of ε. Here, X_* is the law of the limiting empirical distribution of the signal, as defined in Section 2. For $t \geq 1$, given $\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Omega}}_t, \bar{\boldsymbol{\Sigma}}_t$, let

$$\begin{aligned} &(G, R_1, \dots, R_{t-1}) \sim \mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{\Sigma}}_t), \\ &S_t = h_t(R_1, \dots, R_{t-1}, Y), \ \text{ where } Y = q(G, \varepsilon), \\ &(X_1, \dots X_t) = \bar{\boldsymbol{\mu}}_t X_* + (W_1, \dots, W_t), \\ &\text{where } (W_1, \dots, W_t) \sim \mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{\Omega}}_t) \text{ is independent of } X_*, \end{aligned} \tag{12}$$

$$\hat{X}_t = f_t(X_1, \dots X_t). \tag{13}$$

Let $\bar{\boldsymbol{\Delta}}_{t+1}, \bar{\boldsymbol{\Gamma}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ be symmetric matrices with entries given by

$$\begin{aligned} &(\bar{\boldsymbol{\Delta}}_{t+1})_{1,i} = (\bar{\boldsymbol{\Delta}}_{t+1})_{i,1} = 0, \\ &(\bar{\boldsymbol{\Delta}}_{t+1})_{i+1,j+1} = \mathbb{E}\{S_i S_j\}, \quad i, j \in [t], \end{aligned} \tag{14}$$

$$\begin{aligned} &(\bar{\boldsymbol{\Gamma}}_{t+1})_{1,1} = \mathbb{E}\{X_*^2\}, \\ &(\bar{\boldsymbol{\Gamma}}_{t+1})_{1,i+1} = (\bar{\boldsymbol{\Gamma}}_{t+1})_{i+1,1} = \mathbb{E}\{X_* \hat{X}_i\}, \\ &(\bar{\boldsymbol{\Gamma}}_{t+1})_{i+1,j+1} = \mathbb{E}\{\hat{X}_i \hat{X}_j\}, \quad i, j \in [t]. \end{aligned} \tag{15}$$

Furthermore, let $\bar{\boldsymbol{\Psi}}_{t+1}, \bar{\boldsymbol{\Phi}}_{t+1}$ denote the deterministic versions of the matrices $\boldsymbol{\Psi}_{t+1}, \boldsymbol{\Phi}_{t+1}$ in (6), obtained by replacing the empirical averages by expectations. Specifically, to obtain $\bar{\boldsymbol{\Psi}}_{t+1}, \bar{\boldsymbol{\Phi}}_{t+1}$ we replace the entries as follows:

$$\begin{aligned} &\langle \partial_k \hat{\boldsymbol{x}}^t \rangle \to \mathbb{E}\{\partial_k \hat{X}_t\} = \mathbb{E}\{\partial_{X_k} f_t(X_1, \dots, X_t)\}, \\ &\langle \partial_k \boldsymbol{s}^t \rangle \to \mathbb{E}\{\partial_k S_t\} = \mathbb{E}\{\partial_{R_k} h_t(R_1, \dots, R_{t-1}, q(G, \varepsilon))\}, \\ &\langle \partial_g \boldsymbol{s}^t \rangle \to \mathbb{E}\{\partial_g S_t\} = \mathbb{E}\{\partial_g h_t(R_1, \dots, R_{t-1}, q(g, \varepsilon))|_{g=G}\}. \end{aligned} \tag{16}$$

We now describe how $\bar{\boldsymbol{\Sigma}}_{t+1}, \bar{\boldsymbol{\Omega}}_{t+1}, \bar{\boldsymbol{\mu}}_{t+1}$ are computed from $\bar{\boldsymbol{\Sigma}}_t, \bar{\boldsymbol{\Omega}}_t, \bar{\boldsymbol{\mu}}_t$. Given $\bar{\boldsymbol{\Sigma}}_t, \bar{\boldsymbol{\Omega}}_t, \bar{\boldsymbol{\mu}}_t$, we can evaluate the matrices $\bar{\boldsymbol{\Delta}}_{t+1}, \bar{\boldsymbol{\Gamma}}_{t+1}, \bar{\boldsymbol{\Psi}}_{t+1}, \bar{\boldsymbol{\Phi}}_{t+1}$. From these, we compute $\bar{\boldsymbol{\Sigma}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ as

$$\bar{\boldsymbol{\Sigma}}_{t+1} = \sum_{j=0}^{2t+1} \bar{\kappa}_{2(j+1)}\, \boldsymbol{\Xi}^{(j)}_{t+1}, \tag{17}$$

where $\boldsymbol{\Xi}^{(0)}_{t+1} = \bar{\boldsymbol{\Gamma}}_{t+1}$, and for $j \geq 1$:

$$\begin{aligned} \boldsymbol{\Xi}^{(j)}_{t+1} &= \sum_{i=0}^{j} (\bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Phi}}_{t+1})^i\, \bar{\boldsymbol{\Gamma}}_{t+1} \Big((\bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Phi}}_{t+1})^{\mathsf{T}} \Big)^{j-i} \\ &+ \sum_{i=0}^{j-1} (\bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Phi}}_{t+1})^i \bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Delta}}_{t+1} \bar{\boldsymbol{\Psi}}^{\mathsf{T}}_{t+1} \Big((\bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Phi}}_{t+1})^{\mathsf{T}} \Big)^{j-i-1}. \end{aligned} \tag{18}$$

Recalling that $\bar{\mathbf{\Sigma}}_{t+1}$ is the covariance of $(G, R_1, \dots, R_t)$, we can now compute $\bar{\mathbf{\Delta}}_{t+2}, \bar{\mathbf{\Phi}}_{t+2} \in \mathbb{R}^{(t+2)\times(t+2)}$. Using these, we define a symmetric $(t+2)\times(t+2)$ matrix $\mathbf{\Omega}'_{t+2}$, whose first row and column equal zero and whose lower right $(t+1)\times(t+1)$ submatrix equals $\bar{\mathbf{\Omega}}_{t+1}$. Specifically,

$$\mathbf{\Omega}'_{t+2} = \delta \sum_{j=0}^{2(t+1)} \bar{\kappa}_{2(j+1)} \mathbf{\Theta}^{(j)}_{t+2}, \tag{19}$$

where $\mathbf{\Theta}^{(0)}_{t+2} = \bar{\mathbf{\Delta}}_{t+2}$, and for $j \geq 1$:

$$\mathbf{\Theta}^{(j)}_{t+2} = \sum_{i=0}^{j} (\bar{\mathbf{\Phi}}_{t+2}\bar{\mathbf{\Psi}}_{t+2})^i \, \bar{\mathbf{\Delta}}_{t+2} \Big((\bar{\mathbf{\Phi}}_{t+2}\bar{\mathbf{\Psi}}_{t+2})^\mathsf{T} \Big)^{j-i} + \sum_{i=0}^{j-1} (\bar{\mathbf{\Phi}}_{t+2}\bar{\mathbf{\Psi}}_{t+2})^i \, \bar{\mathbf{\Phi}}_{t+2}\bar{\mathbf{\Gamma}}_{t+2}\bar{\mathbf{\Phi}}^\mathsf{T}_{t+2} \Big((\bar{\mathbf{\Phi}}_{t+2}\bar{\mathbf{\Psi}}_{t+2})^\mathsf{T} \Big)^{j-i-1}. \tag{20}$$

Then, the entries of the covariance matrix $\bar{\mathbf{\Omega}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ are given by:

$$(\bar{\mathbf{\Omega}}_{t+1})_{ij} = (\mathbf{\Omega}'_{t+2})_{i+1,j+1}, \quad i,j \in [t+1]. \tag{21}$$

Finally, we compute the mean parameter $\bar{\mu}_{t+1} = \left(\bar{\boldsymbol{M}}^{\beta}_{t+2}\right)_{t+2,1}$, where

$$\bar{\boldsymbol{M}}^{\beta}_{t+2} = \delta \sum_{j=0}^{t+1} \bar{\kappa}_{2(j+1)} \, \bar{\mathbf{\Phi}}_{t+2} \Big(\bar{\mathbf{\Psi}}_{t+2}\bar{\mathbf{\Phi}}_{t+2} \Big)^j. \tag{22}$$

Though the formulas for $\mathbf{\Theta}^{(j)}_{t+2}$ in (20) and $\bar{\boldsymbol{M}}^{\beta}_{t+2}$ in (22) contain $\bar{\mathbf{\Gamma}}_{t+2}$ and $\bar{\mathbf{\Psi}}_{t+2}$, the last rows and columns of these matrices are zeroed out in the computation (due to the form of $\bar{\mathbf{\Delta}}_{t+2}$ and $\bar{\mathbf{\Phi}}_{t+2}$). Therefore the formulas depend only on the top left submatrices of $\bar{\mathbf{\Gamma}}_{t+2}$ and $\bar{\mathbf{\Psi}}_{t+2}$, namely, $\bar{\mathbf{\Gamma}}_{t+1}$ and $\bar{\mathbf{\Psi}}_{t+1}$. We also note that the matrices $\bar{\mathbf{\Omega}}_t$ and $\bar{\mathbf{\Sigma}}_t$ are the top left submatrices of $\bar{\mathbf{\Omega}}_{t+1}$ and $\bar{\mathbf{\Sigma}}_{t+1}$, respectively. Similarly, the mean vector $\bar{\boldsymbol{\mu}}_{t+1}$ is obtained by appending $\bar{\mu}_{t+1}$ to $\bar{\boldsymbol{\mu}}_t$.

Main result. Having defined the state evolution recursion to compute $\bar{\boldsymbol{\mu}}_t, \bar{\mathbf{\Omega}}_t, \bar{\mathbf{\Sigma}}_t$ (which specify the joint distributions in (12)-(13)), we are ready to state our main result. We make the following assumption on the functions f_t, h_t used in the AMP (4)-(5), for $t \geq 1$:

(A1) The functions $f_t(X_1, \dots, X_t)$ and $h_t(R_1, \dots, R_{t-1}, q(G, \varepsilon))$ are Lipschitz in each of their arguments. The partial derivatives $\partial_{X_k} f_t((X_1, \dots, X_t))$, $\partial_G h_t(R_1, \dots, R_{t-1}, q(G, \varepsilon))$, and $\partial_{R_\ell} h_t(R_1, \dots, R_{t-1}, q(G, \varepsilon))$ are all continuous on sets of probability 1, under the laws of $(X_1 \dots, X_t)$ and $(G, R_1, \dots, R_{t-1})$ given in (12)-(13).

Theorem 3.1. *Consider a rotationally invariant generalized linear model with the assumptions in Section 2 and the AMP (4)-(5) with the assumption **(A1)** above. Let $\psi : \mathbb{R}^{2t+1} \to \mathbb{R}$ and $\phi : \mathbb{R}^{2t+2} \to \mathbb{R}$ be any pseudo-Lipschitz functions of order 2. Then for each $t \geq 1$, we almost surely have*

$$\lim_{n\to\infty} \frac{1}{d} \sum_{i=1}^{d} \psi(x_i^1, \dots, x_i^t, \hat{x}_i^1, \dots, \hat{x}_i^t, x_i^*) = \mathbb{E}\{\psi(X_1, \dots, X_t, \hat{X}_1, \dots, \hat{X}_t, X_*)\}, \tag{23}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(r_i^1, \dots, r_i^t, s_i^1, \dots, s_i^{t+1}, y_i) = \mathbb{E}\{\phi(R_1, \dots, R_t, S_1, \dots, S_{t+1}, Y)\}, \tag{24}$$

where the random variables on the right are defined in (12)-(13). Equivalently, as $n \to \infty$, the joint empirical distributions of $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t, \hat{\boldsymbol{x}}^1, \dots, \hat{\boldsymbol{x}}^t, \boldsymbol{x}^*)$ and $(\boldsymbol{r}^1, \dots, \boldsymbol{r}^t, \boldsymbol{s}^1, \dots, \boldsymbol{s}^{t+1}, \boldsymbol{y})$ converge almost surely in Wasserstein-2 distance to $(X_1, \dots, X_t, \hat{X}_1, \dots, \hat{X}_t, X_*)$ and $(R_1, \dots, R_t, S_1, \dots, S_{t+1}, Y)$, respectively.

The proof of the theorem is given in Appendix D, and we provide a proof sketch in Section 5. When the design matrix $\boldsymbol{A}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries, we have $\bar{\kappa}_2 = \delta$ and $\bar{\kappa}_{2k} = 0$ for $k \geq 2$. In this case, the RI-GAMP (4)-(5), with denoising functions of the form $f_t(\boldsymbol{x}^t)$ and $h_{t+1}(\boldsymbol{r}^t, \boldsymbol{y})$, reduces to the existing GAMP algorithm (Rangan, 2011). The state evolution recursion also reduces to that of GAMP (see, e.g., Section 4 of (Feng et al., 2021)). This opens up an exciting research direction on using RI-GAMP to generalize results where GAMP has been used as a proof technique under Gaussian model assumptions. One example is to determine the distributional properties of spectral estimators for rotationally invariant GLMs.

MSE and correlation. The result (23) readily leads to the evaluation of the usual quantities of interest, such as the mean squared error (MSE) and the normalized squared correlation. Indeed, by taking $\psi(\hat{x}_i^t, x_i^*) = (\hat{x}_i^t - x_i^*)^2$, we have that $\frac{1}{d}\|\hat{\boldsymbol{x}}^t - \boldsymbol{x}^*\|_2^2 \to \mathbb{E}\{(\hat{X}_t - X_*)^2\}$ for each $t \geq 1$. Furthermore, by taking $\psi(\hat{x}_i^t, x_i^*) = \hat{x}_i^t \cdot x_i^*$, $\psi(\hat{x}_i^t) = (\hat{x}_i^t)^2$ and $\psi(x_i^*) = (x_i^*)^2$, we have that $|\langle \hat{\boldsymbol{x}}^t, \boldsymbol{x}^* \rangle|^2 / (\|\hat{\boldsymbol{x}}^t\|^2 \|\boldsymbol{x}^*\|^2)$ tends to $(\mathbb{E}\{X_t\, X_*\})^2 / (\mathbb{E}\{X_t^2\}\mathbb{E}\{X_*^2\})$.

Empirical state evolution parameters. We can define empirical versions of the state evolution parameters, denoted by $\mathbf{\Sigma}_{t+1}, \mathbf{\Omega}_{t+1}, \boldsymbol{\mu}_{t+1}$, by replacing $\bar{\mathbf{\Psi}}_k, \bar{\mathbf{\Phi}}_k, \bar{\mathbf{\Delta}}_k, \bar{\mathbf{\Gamma}}_k$ ($k \in \{t+1, t+2\}$) with $\mathbf{\Psi}_k, \mathbf{\Phi}_k, \mathbf{\Delta}_k, \mathbf{\Gamma}_k$ in (18), (20), and (22). The latter matrices are computed using empirical averages instead of expectations: $\mathbf{\Phi}_k, \mathbf{\Psi}_k$ are defined in (6) and for $\mathbf{\Delta}_k, \mathbf{\Gamma}_k$, we replace the expectations $\mathbb{E}\{S_i S_j\}$ and $\mathbb{E}\{\hat{X}_i \hat{X}_j\}$ in (14)-(15) by $\langle \boldsymbol{s}^i, \boldsymbol{s}^j \rangle / n$ and $\langle \hat{\boldsymbol{x}}^i, \hat{\boldsymbol{x}}^j \rangle / d$. The expectations $\mathbb{E}\{X_* \hat{X}_i\}$ in $\bar{\mathbf{\Gamma}}_k$ can be estimated for the case

of posterior mean denoisers using the identity in (57). Theorem 3.1 gives that the empirical versions of these matrices converge to the deterministic ones almost surely, and therefore, $\boldsymbol{\Sigma}_{t+1} \to \bar{\boldsymbol{\Sigma}}_{t+1}$, $\boldsymbol{\Omega}_{t+1} \to \bar{\boldsymbol{\Omega}}_{t+1}$, and $\boldsymbol{\mu}_{t+1} \to \bar{\boldsymbol{\mu}}_{t+1}$. For the simulations in Section 4, RI-GAMP is implemented with state evolution parameters empirically estimated, as this choice leads to more stable numerical results.

Initialization. Note that the algorithm is initialized with $\boldsymbol{x}^1 = \boldsymbol{A}^{\mathsf{T}} h_1(\boldsymbol{y})$. If this initialization is not effective, in the sense that $\frac{1}{d}\langle \boldsymbol{x}^1, \boldsymbol{x}^* \rangle \to 0$, then state evolution remains stuck at a trivial fixed point, i.e., $\bar{\mu}_t = 0$ for all t, and all the iterates produced by RI-GAMP are not correlated with the signal.[1] To address this issue, we can assume to be given an initialization $\boldsymbol{x}^1$, which is correlated with $\boldsymbol{x}^*$, i.e., $\frac{1}{d}\langle \boldsymbol{x}^1, \boldsymbol{x}^* \rangle \to \alpha > 0$, and independent of $\boldsymbol{A}$. Theorem 3.1 still holds for such an initialization, with the only change being in the initialization of the state evolution recursion. Suppose that $\boldsymbol{x}^1 \overset{W_2}{\longrightarrow} X_1$ for a random variable X_1 satisfying $\mathbb{E}\{X_1 X_*\} = \alpha$. Then, in the state evolution initialization (11), we set $\bar{\boldsymbol{\mu}}_1 = \mathbb{E}\{X_1 X_*\} = \alpha$ and $\bar{\boldsymbol{\Omega}}_1 = \mathbb{E}\{(X_1 - \bar{\boldsymbol{\mu}}_1 X_*)^2\}$ (the parameter $\bar{\boldsymbol{\Sigma}}_1$ is unchanged). This ensures that state evolution is not stuck at a trivial fixed point. A practical alternative to assuming an informative initialization is to initialize AMP with a spectral estimator. Analyzing RI-GAMP with spectral initialization is an interesting direction for future research.

Choice of denoisers. The performance of RI-GAMP is determined by the functions $\{f_t, h_{t+1}\}_{t \geq 1}$. A key question is how to choose these functions to optimize the estimation performance. Given any choice of $\{f_t, h_{t+1}\}$ satisfying Assumption **(A1)**, Theorem 3.1 implies

$$\begin{aligned} &\boldsymbol{r}^{t-1} \overset{W_2}{\longrightarrow} R_{t-1} \equiv \frac{(\bar{\boldsymbol{\Sigma}}_t)_{t,1}}{(\bar{\boldsymbol{\Sigma}}_t)_{1,1}} G + W'_{t-1}, \\ &W'_{t-1} \sim \mathcal{N}\left(0, (\bar{\boldsymbol{\Sigma}}_t)_{t,t} - \frac{((\bar{\boldsymbol{\Sigma}}_t)_{t,1})^2}{(\bar{\boldsymbol{\Sigma}}_t)_{1,1}}\right), \quad t \geq 2, \end{aligned} \tag{25}$$

$$\begin{aligned} &\boldsymbol{x}^t \overset{W_2}{\longrightarrow} X_t \equiv \bar{\mu}_t X_* + W_t, \\ &W_t \sim \mathcal{N}(0, (\bar{\boldsymbol{\Omega}}_t)_{t,t}) \text{ independent of } X_*, \quad t \geq 1, \end{aligned} \tag{26}$$

where the RHS of (25) and (26) follow from the joint distributions specified in (12)-(13). From (26), we see that the quality of the estimate in each iteration t is governed by the ratio $\bar{\mu}_t^2/(\bar{\boldsymbol{\Omega}}_t)_{t,t}$. Thus, having fixed $\{f_k, h_k\}_{k \leq t-1}$, the Bayes-optimal choice for $h_t(r_1, \ldots, r_{t-1}, y)$ maximizes $\bar{\mu}_t^2/(\bar{\boldsymbol{\Omega}}_t)_{t,t}$. Similarly, given $\{f_k\}_{k \leq t-1}$ and $\{h_k\}_{k \leq t}$, the Bayes-optimal choice for $f_t(x_1, \ldots, x_t)$ maximizes the normalized squared correlation of R_t and G, which is proportional to $\frac{(\bar{\boldsymbol{\Sigma}}_{t+1})^2_{t+1,1}}{(\bar{\boldsymbol{\Sigma}}_{t+1})_{t+1,t+1}}$. We remark that, even when the signal prior (i.e., the law of X_*) is known, finding these optimal denoisers is challenging due to the complicated nature of the state evolution recursion (12)-(22). However, for the special case of an i.i.d. Gaussian design, the state evolution is considerably simpler and the Bayes-optimal choices are (cf. Section 4.2 of (Feng et al., 2021)):

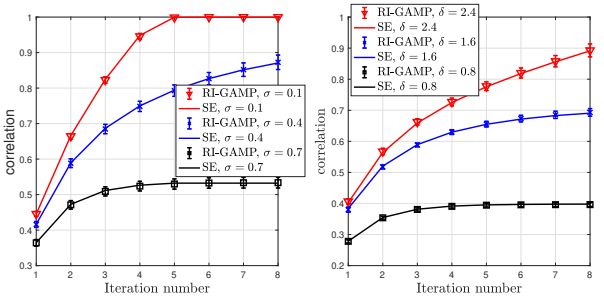
$$\begin{aligned} f_t(x_t) &= c_1 \mathbb{E}\{X_* \mid X_t = x_t\}, \\ h_{t+1}(r_t, y) &= c_2(\mathbb{E}\{G | R_t = r_t, Y = y\} - \mathbb{E}\{G | R_t = r_t\}), \end{aligned} \tag{27}$$

where c_1, c_2 are arbitrary non-zero constants. Here, for a general rotationally invariant $\boldsymbol{A}$, we propose the following denoisers:

$$f_t(x_1, \ldots, x_t) = \mathbb{E}\{X_* \mid X_1 = x_1, \ldots, X_t = x_t\}, \tag{28}$$

$$\begin{aligned} h_{t+1}(r_1, \ldots, r_t, y) = &\mathbb{E}\{G | R_1 = r_1, \ldots, R_t = r_t, Y = y\} \\ &- \mathbb{E}\{G \mid R_1 = r_1, \ldots, R_t = r_t\}. \end{aligned} \tag{29}$$

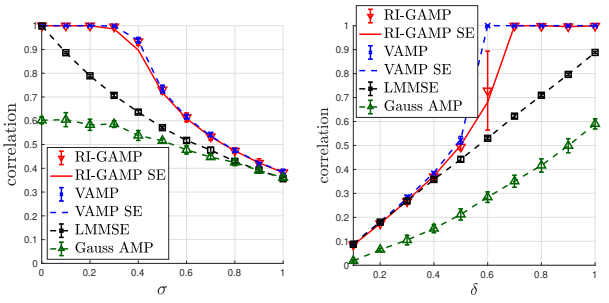
For an i.i.d. Gaussian $\boldsymbol{A}$, (28)-(29) reduce to (27), which is provably Bayes-optimal. When $\boldsymbol{A}$ is not Gaussian, using denoisers that depend on all the preceding iterates (instead of only the most recent one) can have a remarkable impact on the performance of RI-GAMP. In fact, in the setting of Section 4 where the eigenvalues of $\boldsymbol{A}$ follow a Beta distribution, taking (27) does not improve much over the performance of the existing GAMP algorithm that assumes $\boldsymbol{A}$ to be Gaussian (green curves in Figures 2-3). In contrast, taking (28)-(29) leads to a performance close to VAMP (blue curves). Though we do not expect the choices in (28)-(29) to be optimal iteration-by-iteration, based on the simulation results we conjecture that they achieve the same fixed point as the Bayes-optimal denoisers.



(a) Linear regression (b) 1-bit compressed sensing

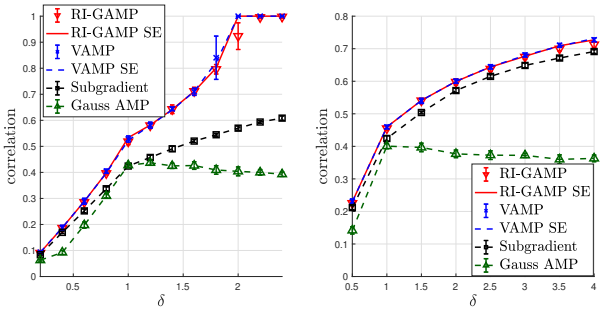
Figure 1: Normalized squared correlation between iterate $\hat{\boldsymbol{x}}^t$ and signal $\boldsymbol{x}^*$, as a function of the number of iterations t. Solid lines represent state evolution predictions, and the markers represent the empirical performance of RI-GAMP.

[1] For both linear regression and 1-bit compressed sensing, the standard initialization $\boldsymbol{x}^1 = \boldsymbol{A}^{\mathsf{T}} h_1(\boldsymbol{y})$ is effective. For a characterization of the GLMs for which this initialization is not effective for a Gaussian design matrix $\boldsymbol{A}$, see (3.13) in (Mondelli & Venkataramanan, 2021a) and the discussion therein. One important example is phase retrieval ($y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle|^2 + \varepsilon_i$).



(a) $\delta = 1$ (b) $\sigma = 0.1$

Figure 2: Linear regression with a Rademacher prior: normalized squared correlation vs. noise level σ (on the left) and vs. aspect ratio δ (on the right).



(a) Rademacher prior (b) Gaussian prior

Figure 3: Noiseless 1-bit compressed sensing: normalized squared correlation vs. aspect ratio δ, for two priors.

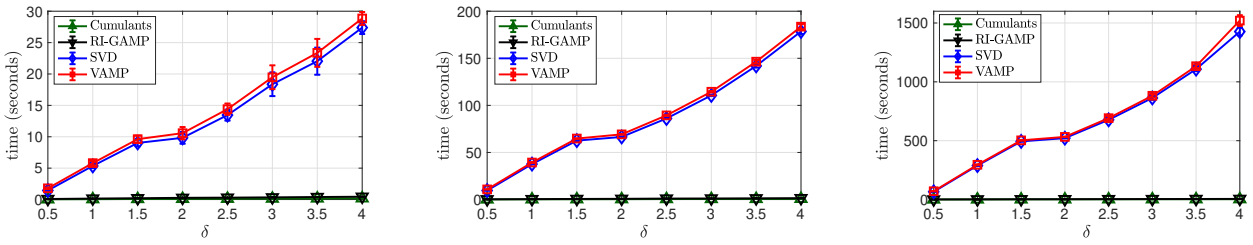
4. Numerical Simulations

For synthetic data, we consider two models: *(i)* linear regression, i.e., $y_i = \langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$; *(ii)* noiseless 1-bit compressed sensing, i.e., $y_i = \mathrm{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle)$. The design matrix $\boldsymbol{A}$ is rotationally invariant in law, i.e., $\boldsymbol{A} = \boldsymbol{O}^\mathsf{T} \boldsymbol{\Lambda} \boldsymbol{Q}$, where $\boldsymbol{O}$, $\boldsymbol{Q}$ are Haar orthogonal matrices, and $\boldsymbol{\Lambda}$ has i.i.d. $\sqrt{6} \cdot \mathrm{Beta}(1,2)$ diagonal entries. (The normalization of the $\mathrm{Beta}(1,2)$ distribution is chosen to ensure a unit second moment.) In the simulations, the free cumulants κ_{2k} are replaced by their limits $\bar{\kappa}_{2k}$, which can be obtained in closed-form (see Appendix A). We set $d = 8000$, repeat each experiment for 10 independent runs, and report the average and error bars at 1 standard deviation.

We implement the RI-GAMP given in (4)-(5), with initialization $\boldsymbol{s}^1 = \boldsymbol{y}$ and $\boldsymbol{x}^1 = \boldsymbol{A}^\mathsf{T} \boldsymbol{s}^1$. The denoisers f_t and h_{t+1}, for $t \geq 1$, are given by (28)-(29). The expressions for these denoisers and the associated calculations are given in Appendix B. The denoisers f_t, h_{t+1} and their derivatives depend on the state evolution parameters, which can be estimated consistently from the data. The implementation details are described at the end of Appendix B.

Figure 1 shows that the state evolution predictions closely match the performance of RI-GAMP for practical values of d and n, validating the result of Theorem 3.1. We plot the normalized squared correlation $\langle \hat{\boldsymbol{x}}^t, \boldsymbol{x}^* \rangle^2 / (\|\hat{\boldsymbol{x}}^t\|^2 \|\boldsymbol{x}^*\|^2)$ as a function of the iteration number. In (a), we consider linear regression with a Rademacher prior, $\delta = 1$ and $\sigma \in \{0.1, 0.4, 0.7\}$; and in (b), noiseless 1-bit compressed sensing with a Rademacher prior and $\delta \in \{0.8, 1.6, 2.4\}$. In all cases, the agreement between RI-GAMP and its SE is excellent. The next two figures show that the performance of RI-GAMP closely matches that of VAMP in a variety of settings. The results for linear regression with a Rademacher prior are shown in Figure 2: on the left, we plot the normalized squared correlation as a function of σ, for $\delta = 1$; on the right, we plot the same metric as a function of δ, for $\sigma = 0.1$. Additional results for a different choice of δ and σ are reported in Figure 6 in Appendix C. Figure 3 shows the performance for noiseless 1-bit compressed sensing: we plot the normalized squared correlation as a function of δ, for two signal priors (Rademacher in (a), and Gaussian in (b)). The *red curve* in each plot corresponds to *RI-GAMP*, together with the related SE. The *blue curve* corresponds to *VAMP*, together with the related SE. The implementation details for VAMP are given at the end of Appendix B. The *green curve* corresponds to the standard *GAMP* algorithm which is derived based on the (incorrect) assumption that $\boldsymbol{A}$ is i.i.d. Gaussian. The denoisers f_t and h_{t+1} are given by (27), which would be Bayes-optimal were the design matrix $\boldsymbol{A}$ Gaussian. The implementation of GAMP is a special case of our proposed RI-GAMP (obtained by setting all the rectangular free cumulants except $\bar{\kappa}_2$ to 0). The GAMP state evolution predictions (not shown in the plots) do not match the performance of the algorithm, since $\boldsymbol{A}$ is not Gaussian. Finally, the *black curve* corresponds to *(i)* the *linear minimum mean squared error* (LMMSE) estimator $\hat{\boldsymbol{x}} = \boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{A}^\mathsf{T} + \sigma \boldsymbol{I})^{-1} \boldsymbol{y}$ for linear regression, and *(ii)* a *subgradient* method for 1-bit compressed sensing. This last method minimizes $\|[y \odot \mathrm{sign}(\boldsymbol{A}\boldsymbol{x})]_-\|_1$ via subgradient descent (here, $\odot$ denotes the Hadamard product and $[a]_- = \max\{-a, 0\}$ is applied component-wise). The algorithm was proposed in (Jacques et al., 2013) for the recovery of sparse signals, and the original version includes a sparsity enforcing step. For our setup (with no sparsity), we run it without the sparsity enforcing step, and the method reduces to subgradient descent.

Performance of RI-GAMP vs. VAMP. Taking the results of Figures 2-3 together, we highlight that RI-GAMP exhibits a performance close to VAMP. Recall that the fixed points of the VAMP state evolution satisfy the replica equation, whose solution is conjectured to give the Bayes-



(a) $d = 2000$ (b) $d = 4000$ (c) $d = 8000$

Figure 4: Running time of RI-GAMP vs. VAMP. The green curve shows the time for computing the free cumulants in RI-GAMP, and the blue curve the time for the SVD in the initial step of VAMP.

optimal mean squared error; see Theorem 3 of Rangan et al. (2019) for linear regression, and Theorem 4 of Pandit et al. (2020) for the general case. Thus, this conjecture implies that VAMP is Bayes-optimal when its state evolution has a unique fixed point, and our numerical simulations suggest that RI-GAMP is also near-optimal. We remark that, close to the phase transition for exact recovery (see $\delta \approx 0.6$ in Figure 2b, and $\delta \approx 1.8$ in Figure 3a), there is a (small) performance gap between RI-GAMP and VAMP. This is due to the fact that, for both RI-GAMP and VAMP, the number of iterations needed for convergence grows when the algorithm operates close to this phase transition. We note that VAMP shows a larger error bar for $\delta \approx 1.8$ in Figure 3a. For RI-GAMP, the issue is that the expressions of f_t and h_{t+1} depend on covariance matrices whose dimension grows with t. For large t, these covariance matrices become ill-conditioned and RI-GAMP is unstable. However, the stability displayed by VAMP comes at the cost of requiring the computationally expensive SVD of $\boldsymbol{A}$. As shown next, RI-GAMP is significantly faster than VAMP, and is therefore an appealing alternative in many practical settings.

Complexity of RI-GAMP vs. VAMP. The computational complexity of VAMP is dominated by the initial SVD which has $O(d^3)$ running time. In contrast, the free cumulants required for RI-GAMP can be estimated in $O(d^2)$ time, as described on p.4. Each iteration of RI-GAMP also takes $O(d^2)$ time, and the algorithm typically converges in a few tens of iterations.

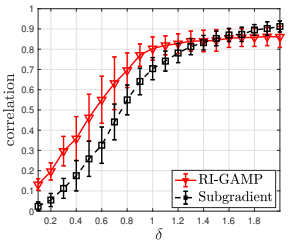
Figure 4 shows the running times for VAMP (including the initial SVD) and RI-GAMP (including the estimation of the free cumulants from the data), for noiseless 1-bit compressed sensing. The running time of VAMP is dominated by the SVD, and the computational advantage of RI-GAMP increases quickly with the problem dimension. For $\delta := \frac{n}{d} \in [0.5, 4]$, RI-GAMP is 20-60$\times$ faster than VAMP at $d = 2000$, 40-120$\times$ faster at $d = 4000$, and 80-240$\times$ faster at $d = 8000$.

Impact of eigenvalue distribution and prior. RI-GAMP exploits the spectral distribution of $\boldsymbol{A}$, which gives a large performance improvement over the GAMP designed for a Gaussian $\boldsymbol{A}$. Furthermore, RI-GAMP also takes advantage of the signal prior, which cannot be exploited by either the LMMSE estimator (for linear regression) or the subgradient method (for 1-bit compressed sensing). In fact, the subgradient method has roughly the same correlation for the two choices of the prior (cf. the black curves in Figure 3a and 3b), and is outperformed by RI-GAMP in both settings.

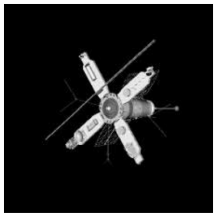
1-bit compressed sensing on a sparse image. In Figure 5, we consider noiseless 1-bit compressed sensing with the input $\boldsymbol{x}^*$ being the sparse grayscale image considered in (Schniter & Rangan, 2014), with $d = 225^2 = 50625$ and a sparsity (fraction of non-black pixels) of $8645/50625$. The design matrix $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{Q}_n \boldsymbol{\Pi}_n \boldsymbol{\Lambda} \boldsymbol{\Pi}_d \boldsymbol{Q}_d$, where $\boldsymbol{Q}_n$, $\boldsymbol{Q}_d$ are orthonormal Discrete Cosine Transform (DCT) matrices in n, d dimensions, $\boldsymbol{\Pi}_n, \boldsymbol{\Pi}_d$ are random permutation matrices, and $\boldsymbol{\Lambda}$ has i.i.d. $\sqrt{6} \cdot \mathrm{Beta}(1,2)$ diagonal entries. This choice of $\boldsymbol{A}$ significantly speeds up matrix multiplications, as in (Tian et al., 2021). We report the average and error bars at 1 standard deviation for 100 independent trials. For RI-GAMP, we use a non-negative Bernoulli-Gaussian prior (cf. (Schniter & Rangan, 2014) and (Vila & Schniter, 2014)); the expression for the corresponding denoiser f_t is in Appendix B. As shown in Figure 5a, RI-GAMP improves on the subgradient method in (Jacques et al., 2013) up until $\delta = 1.5$ (this improvement is clearly visibile in the reconstructions for $\delta = 0.8$, see Figures 5c and 5d). For larger δ, the performance of RI-GAMP does not improve further, due to the aforementioned numerical instabilities. Additional experiments on RGB images when the input $\boldsymbol{x}^*$ is obtained via a wavelet transform are reported in Appendix C.

5. Proof Sketch of Theorem 3.1

The proof is based on an auxiliary AMP algorithm whose iterates mimic the true AMP in (4)-(5). The iterates of the

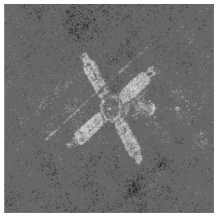


(a) Correlation vs. δ



(b) Original image

(c) RI-GAMP, $\delta = 0.8$



(d) Subgradient, $\delta = 0.8$

Figure 5: RI-GAMP versus subgradient method for the recovery of a sparse image from 1-bit measurements.

auxiliary AMP, denoted by $\boldsymbol{z}^t, \boldsymbol{v}^t \in \mathbb{R}^d$ and $\boldsymbol{v}^t, \boldsymbol{u}^{t+1} \in \mathbb{R}^n$, are computed as follows, for $t \geq 1$:

$$\boldsymbol{z}^t = \boldsymbol{A}^\mathsf{T}\boldsymbol{u}^t - \sum_{i=1}^{t-1} \mathsf{b}_{ti}\boldsymbol{v}^i, \ \boldsymbol{v}^t = \tilde{f}_t(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^t, \boldsymbol{x}^*), \quad (30)$$

$$\boldsymbol{m}^t = \boldsymbol{A}\boldsymbol{v}^t - \sum_{i=1}^{t} \mathsf{a}_{ti}\boldsymbol{u}^i, \ \boldsymbol{u}^{t+1} = \tilde{h}_{t+1}(\boldsymbol{m}^1, \ldots, \boldsymbol{m}^t, \boldsymbol{\varepsilon}). \quad (31)$$

The iteration is initialized with $\boldsymbol{u}^1 = \boldsymbol{0}, \quad \boldsymbol{z}^1 = \boldsymbol{0}$. The function $\tilde{f}_1 : \mathbb{R}^2 \to \mathbb{R}$ is defined as $\tilde{f}_1(z_1, x) = x$, which yields $\boldsymbol{v}^1 = \boldsymbol{x}^*$ and $\boldsymbol{m}^1 = \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{g}$. For $t \geq 1$, the functions $\tilde{f}_{t+1} : \mathbb{R}^{t+2} \to \mathbb{R}$ and $\tilde{h}_{t+1} : \mathbb{R}^{t+1} \to \mathbb{R}$, which act row-wise on matrices, are defined as

$$\begin{aligned} &\tilde{f}_{t+1}(z_1, \ldots, z_{t+1}, x) \\ &\quad = f_t(z_2 + \bar{\mu}_1 x, \ z_3 + \bar{\mu}_2 x, \ \ldots, \ z_{t+1} + \bar{\mu}_t x), \quad (32) \\ &\tilde{h}_{t+1}(m_1, \ \ldots, m_t, \ \varepsilon) = h_t(m_2, \ldots, m_t, \ q(m_1, \varepsilon)). \end{aligned}$$

Here, f_t, h_t are the functions from the original AMP, and $(\bar{\mu}_1, \ldots, \bar{\mu}_t)$ are the state evolution parameters computed according to (22). The debiasing coefficients $\{\mathsf{a}_{ti}\}_{i=1}^t$ and $\{\mathsf{b}_{ti}\}_{i=1}^{t-1}$ for the auxiliary AMP are given in Appendix D.1.

Idea of the proof. The auxiliary AMP (30)-(31) is an instance of the abstract AMP recursion for non-symmetric rotationally invariant matrices, which was analyzed in (Zhong et al., 2021). The state evolution result in Theorem 2.6 of (Zhong et al., 2021) implies that the joint empirical distribution of $(\boldsymbol{m}^1, \ldots, \boldsymbol{m}^t)$ converges to a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Sigma}}_t)$. Similarly, the joint empirical distribution of $(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^t)$ also converges to a a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Omega}}_t)$. The covariance matrices $\widetilde{\boldsymbol{\Sigma}}_t, \widetilde{\boldsymbol{\Omega}}_t$ are recursively defined via the state evolution for the auxiliary AMP (for details, see Section D.2).

The proof of Theorem 3.1 consists of two steps. First, we show that the state evolution parameters of the auxiliary AMP match those of the true AMP. In particular, Lemma D.2 proves that $\widetilde{\boldsymbol{\Sigma}}_t = \bar{\boldsymbol{\Sigma}}_t$ and $\widetilde{\boldsymbol{\Omega}}_{t+1} = \boldsymbol{\Omega}'_{t+1}$, where the matrices on the right are defined in (17) and (19). Next, we show in Lemma D.3 that the true AMP iterates (4)-(5) are close to the auxiliary AMP iterates (30)-(31) in the following sense. For $t \geq 1$:

$$\begin{aligned} &\frac{\|\boldsymbol{x}^t - (\boldsymbol{z}^{t+1} + \bar{\mu}_t \boldsymbol{x}^*)\|^2}{d} \to 0, \quad \frac{\|\hat{\boldsymbol{x}}^t - \boldsymbol{v}^{t+1}\|^2}{d} \to 0, \\ &\frac{\|\boldsymbol{r}^t - \boldsymbol{m}^{t+1}\|^2}{n} \to 0, \quad \frac{\|\boldsymbol{s}^t - \boldsymbol{u}^{t+1}\|^2}{n} \to 0. \quad (33) \end{aligned}$$

Lemma D.3 actually proves a more general convergence statement which implies (33). It shows that the empirical joint distribution of the iterates of the true AMP converges to that of the auxiliary AMP. The result of Theorem 3.1 then follows from Lemmas D.2 and D.3.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. KK and MM were partially supported by the 2019 Lopez-Loreta Prize.

References

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Barbier, J., Macris, N., and Rush, C. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57: 764–785, 2011.

Bayati, M. and Montanari, A. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58: 1997–2017, 2012.

Bayati, M., Lelarge, M., Montanari, A., et al. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2): 753–822, 2015.

Benaych-Georges, F. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144 (3-4):471–515, 2009.

Bolthausen, E. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *Conference on Information Sciences and Systems (CISS)*, pp. 16–21, 2008.

Çakmak, B. and Opper, M. Memory-free dynamics for the Thouless-Anderson-Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices. *Physical Review E*, 99(6):062140, 2019.

Çakmak, B., Opper, M., Fleury, B. H., and Winther, O. Self-averaging expectation propagation. arXiv:1608.06602, 2016.

Candés, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

Candès, E. J., Strohmer, T., and Voroninski, V. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Candès, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Chen, W.-K. and Lam, W.-K. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1 – 44, 2021.

Deshpande, Y. and Montanari, A. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2197–2201, 2014.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52:489–509, April 2006.

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

Donoho, D. L., Javanmard, A., and Montanari, A. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, Nov. 2013.

Dudeja, R., Bakhshizadeh, M., Ma, J., and Maleki, A. Analysis of spectral methods for phase retrieval with random orthogonal matrices. *IEEE Transactions on Information Theory*, 66(8):5182–5203, 2020a.

Dudeja, R., Ma, J., and Maleki, A. Information theoretic limits for phase retrieval with subsampled haar sensing matrices. *IEEE Transactions on Information Theory*, 66: 8002 – 8045, Decemeber 2020b.

Eldar, Y. C. and Kutyniok, G. *Compressed sensing: Theory and applications*. Cambridge University Press, 2012.

Fan, Z. Approximate message passing algorithms for rotationally invariant matrices. *Annals of Statistics*, 2021. to appear.

Fannjiang, A. and Strohmer, T. The numerics of phase retrieval. *Acta Numerica*, 29:125–228, 2020.

Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. A unifying tutorial on Approximate Message Passing. arXiv:2105.02180, 2021.

Fletcher, A. K. and Rangan, S. Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562, 2018.

Gerbelot, C., Abbara, A., and Krzakala, F. Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices. In *Conference on Learning Theory (COLT)*, pp. 1682–1713, 2020a.

Gerbelot, C., Abbara, A., and Krzakala, F. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula). arXiv:2006.06581, 2020b.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the Lasso and generalizations*. Chapman and Hall/CRC, 2019.

Jacques, L., Laska, J. N., Boufounos, P. T., and Baraniuk, R. G. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, pp. 115–144, 2013.

Kabashima, Y. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121, Oct 2003.

Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A., and Zdeborová, L. Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

Lee, G. R., Gommers, R., Waselewski, F., Wohlfahrt, K., and O'Leary, A. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36): 1237, 2019. doi: 10.21105/joss.01237. URL https://doi.org/10.21105/joss.01237.

Lesieur, T., Krzakala, F., and Zdeborová, L. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7): 073403, 2017.

Liu, L., Huang, S., and Kurkoski, B. M. Memory approximate message passing. arXiv:2012.10861, 2020.

Lu, Y. M. and Li, G. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3): 507–541, 2020.

Luo, W., Alghamdi, W., and Lu, Y. M. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9): 2347–2356, 2019.

Ma, J. and Ping, L. Orthogonal AMP. *IEEE Access*, 5: 2020–2033, 2017.

Ma, J., Xu, J., and Maleki, A. Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629, 2019.

Ma, J., Xu, J., and Maleki, A. Impact of sensing spectrum for signal recovery under a generalized linear model. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. Asymptotic analysis of complex lasso via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308, 2013.

McCullagh, P. and Nedler, J. *Generalized linear models*. Chapman and Hall/CRC, 1989.

Mondelli, M. and Montanari, A. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19:703–773, 2019.

Mondelli, M. and Venkataramanan, R. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 397–405. PMLR, 2021a.

Mondelli, M. and Venkataramanan, R. PCA initialization for approximate message passing in rotationally invariant models. In *Neural Information Processing Systems (NeurIPS)*, 2021b.

Mondelli, M., Thrampoulidis, C., and Venkataramanan, R. Optimal combination of linear and spectral estimators for generalized linear models. *Foundations of Computational Mathematics*, pp. 1–54, 2021.

Montanari, A. and Venkataramanan, R. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 45(1):321–345, 2021.

Netrapalli, P., Jain, P., and Sanghavi, S. Phase retrieval using alternating minimization. In *Neural Information Processing Systems (NIPS)*, pp. 2796–2804, 2013.

Opper, M., Cakmak, B., and Winther, O. A theory of solving TAP equations for Ising models with general invariant random matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11):114002, 2016.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):336–347, 2020.

Plan, Y. and Vershynin, R. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59 (1):482–494, 2012.

Plan, Y. and Vershynin, R. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Rangan, S. Generalized Approximate Message Passing for Estimation with Random Linear Mixing. In *IEEE International Symposium on Information Theory (ISIT)*, 2011.

Rangan, S., Schniter, P., and Fletcher, A. K. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

Schniter, P. and Rangan, S. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

Schniter, P., Rangan, S., and Fletcher, A. K. Vector approximate message passing for the generalized linear model. In *50th Asilomar Conference on Signals, Systems and Computers*, pp. 1525–1529. IEEE, 2016.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Takeuchi, K. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386, 2020.

Takeuchi, K. Bayes-optimal convolutional AMP. *IEEE Transactions on Information Theory*, 67(7):4405–4428, 2021a.

Takeuchi, K. On the convergence of Orthogonal/Vector AMP: Long-memory message-passing strategy. arXiv:2111.05522, 2021b.

Tian, F., Liu, L., and Chen, X. Generalized memory approximate message passing. arXiv:2110.06069, 2021.

Tibshirani, R. Regression shrinkage and selection with the Lasso. *J. Royal. Statist. Soc B*, 58:267–288, 1996.

Vila, J. P. and Schniter, P. An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals. *IEEE Transactions on Signal Processing*, 62(18): 4689–4703, 2014.

Zhong, X., Su, C., and Fan, Z. Approximate Message Passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. arXiv:2110.02318, 2021.

A. Background on Rectangular Free Cumulants

Let X be a random variable of finite moments of all orders, and denote its even moments by $m_{2k} = \mathbb{E}\{X^{2k}\}$. In this paper, X^2 represents either the empirical eigenvalue distribution of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T} \in \mathbb{R}^{n\times n}$, or its limit law Λ^2 (in the latter case, the moments and rectangular free cumulants are denoted by $\{\bar{m}_{2k}\}_{k\geq 1}$ and $\{\bar{\kappa}_{2k}\}_{k\geq 1}$, respectively). The rectangular free cumulants $\{\kappa_{2k}\}_{k\geq 1}$ of X are defined recursively by the moment-cumulant relations (cf. Section 3 of (Benaych-Georges, 2009))

$$m_{2k} = \delta \sum_{\pi\in \mathrm{NC}'(2k)} \prod_{\substack{S\in\pi \\ \min S \text{ is odd}}} \kappa_{|S|} \prod_{\substack{S\in\pi \\ \min S \text{ is even}}} \kappa_{|S|}, \tag{34}$$

where $\mathrm{NC}'(2k)$ is the set of non-crossing partitions π of $\{1,\dots,2k\}$ such that each set $S \in \pi$ has even cardinality. Furthermore, by exploiting the connection between the formal power series with coefficients $\{m_{2k}\}_{k\geq 1}$ and $\{\kappa_{2k}\}_{k\geq 1}$, each rectangular free cumulant κ_{2k} can be computed from $m_2,\dots,m_{2k}$ and $\kappa_2,\dots,\kappa_{2(k-1)}$ as (cf. Lemma 3.4 of (Benaych-Georges, 2009))

$$\kappa_{2k} = m_{2k} - [z^k] \sum_{j=1}^{k-1} \kappa_{2j} \left(z(\delta M(z)+1)(M(z)+1)\right)^j, \tag{35}$$

where $M(z) = \sum_{k=1}^{\infty} m_{2k} z^k$ and $[z^k](q(z))$ denotes the coefficient of z^k in the polynomial $q(z)$.

In the numerical simulations of Section 4, the singular values of $\boldsymbol{A}$ are i.i.d. $\sqrt{6}\cdot\mathrm{Beta}(1,2)$. Hence, for $\delta \in (0,1)$, X has distribution $\sqrt{6}\cdot\mathrm{Beta}(1,2)$ and consequently $\bar{m}_{2k} = \frac{6^k}{(k+1)(2k+1)}$; for $\delta \geq 1$, X has distribution $\sqrt{6}\cdot\mathrm{Beta}(1,2)$ w.p. $1/\delta$ and it is equal to 0 w.p. $1-1/\delta$, and consequently $\bar{m}_{2k} = \frac{1}{\delta}\frac{6^k}{(k+1)(2k+1)}$. Then, given the moments $\{\bar{m}_{2k}\}_{k\geq 1}$, the rectangular free cumulants $\{\bar{\kappa}_{2k}\}_{k\geq 1}$ are computed recursively using (35).

B. Computation of Denoisers, and Implementation Details

Computation of f_t for Rademacher prior. Here, $\mathbb{P}(X_* = 1) = \mathbb{P}(X_* = -1) = 1/2$. Hence, (28) can be specialized as:

$$f_t(x_1,\dots,x_t) = \mathbb{E}\{X_* \mid X_1 = x_1,\dots,X_t = x_t\} = 2\cdot\mathbb{P}(X_* = 1 \mid X_1 = x_1,\dots,X_t = x_t) - 1. \tag{36}$$

From (13), we have that $(X_1,\dots,X_t) = \bar{\boldsymbol{\mu}}_t X + (W_1,\dots,W_t)$, with $(W_1,\dots,W_t) \sim \mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{\Omega}}_t)$. Thus,

$$P(X_* = 1 \mid X_1 = x_1,\dots,X_t = x_t) = \frac{\exp\left(\frac{-(\boldsymbol{x}-\bar{\boldsymbol{\mu}}_t)^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}(\boldsymbol{x}-\bar{\boldsymbol{\mu}}_t)}{2}\right)}{\exp\left(\frac{-(\boldsymbol{x}-\bar{\boldsymbol{\mu}}_t)^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}(\boldsymbol{x}-\bar{\boldsymbol{\mu}}_t)}{2}\right) + \exp\left(\frac{-(\boldsymbol{x}+\bar{\boldsymbol{\mu}}_t)^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}(\boldsymbol{x}+\bar{\boldsymbol{\mu}}_t)}{2}\right)}, \tag{37}$$

where $\boldsymbol{x} = (x_1,\dots,x_t)^\mathsf{T}$. (All vectors in this section, including $\boldsymbol{x}$ and $\bar{\boldsymbol{\mu}}_t$, are treated as column vectors, unless otherwise mentioned.) Combining (36) and (37), we obtain

$$f_t(x_1,\dots,x_t) = \tanh\left(\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{x}\right). \tag{38}$$

Furthermore, the partial derivatives of f_t can be expressed in the following compact form:

$$\partial_{x_i} f_t(x_1,\dots,x_t) = \left(1 - \tanh^2\left(\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{x}\right)\right)\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{e}_i, \quad \text{for } i\in[t], \tag{39}$$

where $\boldsymbol{e}_i$ is the vector corresponding to the i-th element of the canonical basis of $\mathbb{R}^t$.

Computation of f_t for Gaussian prior. Here, $X_* \sim \mathcal{N}(0,1)$. By evaluating explicitly the conditional expectation, one readily obtains that

$$f_t(x_1,\dots,x_t) = \mathbb{E}\{X_* \mid X_1 = x_1,\dots,X_t = x_t\} = \frac{\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{x}}{1+\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\bar{\boldsymbol{\mu}}_t}, \tag{40}$$

which leads to the following expressions for the partial derivatives:

$$\partial_{x_i} f_t(x_1,\dots,x_t) = \frac{\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{e}_i}{1+\bar{\boldsymbol{\mu}}_t^\mathsf{T}(\bar{\boldsymbol{\Omega}}_t)^{-1}\bar{\boldsymbol{\mu}}_t}, \quad \text{for } i\in[t], \tag{41}$$

Computation of f_t for non-negative Bernoulli-Gaussian prior. Here, X_* is equal to 0 with probability $1-\lambda$ and it is distributed according to the modulus of a Gaussian with 0 mean and variance σ^2 with probability λ, i.e., $X_* \sim (1-\lambda)\delta_0 + \lambda \mathcal{N}_+(0,\sigma^2)$. The parameter λ is taken to be $1/6$, which is close to the actual sparsity of the image given by $8645/50625$; the parameter σ^2 is taken to be $1/\lambda$, which gives $\mathbb{E}\{X_*^2\} = 1$, and the image is normalized to have unit second moment. Now, we can write

$$f_t(x_1,\dots,x_t) = \mathbb{E}\{X_* \mid X_1 = x_1,\dots,X_t = x_t\} = \frac{\mathbb{E}_{X_*}\{X_*\,\mathbb{P}(X_1 = x_1,\dots,X_t = x_t \mid X_*)\}}{\mathbb{E}_{X_*}\{\mathbb{P}(X_1 = x_1,\dots,X_t = x_t \mid X_*)\}}, \tag{42}$$

where $\mathbb{E}_{X_*}$ denotes the expected value with respect to X_*. Using that $(X_1,\dots,X_t) = \bar{\boldsymbol{\mu}}_t X_* + (W_1,\dots,W_t)$ with $(W_1,\dots,W_t) \sim \mathcal{N}(\mathbf{0},\bar{\boldsymbol{\Omega}}_t)$, it is straightforward to compute the expectations on the RHS, which yields

$$f_t(x_1,\dots,x_t) = \frac{\frac{\lambda}{\sqrt{2\pi\sigma^2}}\left[\frac{\sqrt{\pi}b}{\sqrt{2a^3}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)+\frac{2}{a}\right]}{1-\lambda+\frac{\lambda}{\sqrt{a\sigma^2}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)}, \tag{43}$$

where $a = 1/\sigma^2 + \bar{\boldsymbol{\mu}}_t^{\mathsf{T}}(\bar{\boldsymbol{\Omega}}_t)^{-1}\bar{\boldsymbol{\mu}}_t$, $b = 2\bar{\boldsymbol{\mu}}_t^{\mathsf{T}}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{x}$, $\boldsymbol{x} = (x_1,\dots,x_t)$, and Erf is the error function.
To compute the derivative, we write $\partial_{x_i} f_t(x_1,\dots,x_t) = \partial_b f_t(x_1,\dots,x_t)\partial_{x_i} b$. Since $\partial_{x_i} b = 2\bar{\boldsymbol{\mu}}_t^{\mathsf{T}}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{e}_i$, after some manipulations, one obtains

$$\begin{aligned}\partial_{x_i} f_t(x_1,\dots,x_t) = {} & \frac{2\frac{\lambda}{\sqrt{2\pi\sigma^2}}\left[\frac{b}{2a^2}+\frac{\sqrt{\pi}}{\sqrt{2a^3}}\left(1+\frac{b^2}{4a}\right)\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)\right]}{1-\lambda+\frac{\lambda}{\sqrt{a\sigma^2}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)}\bar{\boldsymbol{\mu}}_t^{\mathsf{T}}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{e}_i \\ & - \frac{\frac{\lambda}{\sqrt{2\pi\sigma^2}}\left(\frac{\sqrt{\pi}b}{\sqrt{2a^3}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)+\frac{2}{a}\right)}{\left(1-\lambda+\frac{\lambda}{\sqrt{a\sigma^2}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)\right)^2} \\ & \cdot 2\left(\frac{\lambda}{a\sqrt{2\pi\sigma^2}}+\frac{\lambda b}{4\sqrt{a^3\sigma^2}}\exp\left(\frac{b^2}{8a}\right)\left(1+\mathrm{Erf}\left(\frac{b}{\sqrt{8a}}\right)\right)\right)\bar{\boldsymbol{\mu}}_t^{\mathsf{T}}(\bar{\boldsymbol{\Omega}}_t)^{-1}\boldsymbol{e}_i.\end{aligned} \tag{44}$$

Computation of h_{t+1} for linear regression. In this case, we have $Y = G + W$, where $W \sim \mathcal{N}(0,\sigma^2)$. For $t = 0$, we set $h_1(y) = y$ and consequently

$$\partial_g h_1(y) = 1. \tag{45}$$

For $t > 0$, h_{t+1} is defined as in (29). From (12), we have that $(G, R_1,\dots,R_t) \sim \mathcal{N}(\mathbf{0},\bar{\boldsymbol{\Sigma}}_{t+1})$. Thus, the second conditional expectation in (29) can be expressed as

$$\mathbb{E}\{G \mid R_1 = r_1,\dots,R_t = r_t\} = (\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,2:t+1]}\right)^{-1}\boldsymbol{r}, \tag{46}$$

where $\boldsymbol{r} = (r_1,\dots,r_t)^{\mathsf{T}}$. Note that $(G, R_1,\dots,R_t, Y) \sim \mathcal{N}(\mathbf{0},\bar{\boldsymbol{S}}_{t+2})$, where

$$\bar{\boldsymbol{S}}_{t+2} := \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_{t+1} & (\bar{\boldsymbol{\Sigma}}_{t+1})_{[1:t+1,1]} \\ (\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,1:t+1]} & \mathbb{E}\{Y^2\} \end{bmatrix}.$$

Here, we denote by $(\boldsymbol{A})_{[i_1:i_2,j_1:j_2]}$ the submatrix obtained by taking the rows of $\boldsymbol{A}$ from i_1 to i_2 and the columns of $\boldsymbol{A}$ from j_1 to j_2 (if $i_1 = i_2$ or $j_1 = j_2$, the second index is omitted). Thus, the first conditional expectation in (29) can be expressed as

$$\mathbb{E}\{G \mid R_1 = r_1,\dots,R_t = r_t, Y = y\} = (\bar{\boldsymbol{S}}_{t+2})_{[1,2:t+2]}\left((\bar{\boldsymbol{S}}_{t+2})_{[2:t+2,2:t+2]}\right)^{-1}\begin{bmatrix}\boldsymbol{r}\\ y\end{bmatrix}. \tag{47}$$

By combining (46) and (47), we obtain

$$\begin{aligned} h_{t+1}(r_1,\dots,r_t,y) &= E\{G \mid R_1 = r_1,\dots,R_t = r_t, Y = y\} - E\{G \mid R_1 = r_1,\dots,R_t = r_t\} \\ &= (\bar{\boldsymbol{S}}_{t+2})_{[1,2:t+2]}\left((\bar{\boldsymbol{S}}_{t+2})_{[2:t+2,2:t+2]}\right)^{-1}\begin{bmatrix}\boldsymbol{r}\\ y\end{bmatrix} - (\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,\,2:t+1]}\right)^{-1}\boldsymbol{r}. \end{aligned} \tag{48}$$

Furthermore, the partial derivatives of h_{t+1} can be expressed in the following compact form:

$$\begin{aligned}\partial_{r_i} h_{t+1}(r_1,\dots,r_t,y) &= (\bar{\boldsymbol{S}}_{t+2})_{[1,2:t+2]}\left((\bar{\boldsymbol{S}}_{t+2})_{[2:t+2,2:t+2]}\right)^{-1}\begin{bmatrix}\boldsymbol{e}_i\\0\end{bmatrix}\\ &\quad-(\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,2:t+1]}\right)^{-1}\boldsymbol{e}_i,\ \text{ for } i\in[t],\\ \partial_g h_{t+1}(r_1,\dots,r_t,y) &= (\bar{\boldsymbol{S}}_{t+2})_{[1,2:t+2]}\left((\bar{\boldsymbol{S}}_{t+2})_{[2:t+2,2:t+2]}\right)^{-1}\boldsymbol{e}_{t+1}.\end{aligned} \tag{49}$$

Computation of h_{t+1} for noiseless 1-bit compressed sensing. In this case, we have that $Y=\operatorname{sign}(G)$, which implies that

$$\mathbb{P}(Y=1\mid G=g)=\frac{1+\operatorname{sign}(g)}{2},\qquad \mathbb{P}(Y=-1\mid G=g)=\frac{1-\operatorname{sign}(g)}{2}. \tag{50}$$

For $t=0$, we set $h_1(y)=y$ and consequently

$$\mathbb{E}\{\partial_g h_1(Y)\}=\mathbb{E}\{\partial_g \operatorname{sign}(G)\}=\frac{\mathbb{E}\{G\operatorname{sign}(G)\}}{\mathbb{E}\{G^2\}}=\sqrt{\frac{2}{\pi\mathbb{E}\{G^2\}}}, \tag{51}$$

where the first equality follows from the definition of h_1 and y, and the second equality is obtained by recalling that G is Gaussian with zero mean and by applying Stein's lemma. For $t>0$, h_{t+1} is defined as in (29). Since $(G,R_1,\dots,R_t)\sim\mathcal{N}(\boldsymbol{0},\bar{\boldsymbol{\Sigma}}_{t+1})$, the conditional distribution of G given $(R_1=r_1,\dots,R_t=r_t)$ is $\mathcal{N}(\hat{r}_t,\hat{\sigma}_t^2)$ where

$$\begin{aligned}\hat{r}_t &= \mathbb{E}\{G\mid R_1=r_1,\dots,R_t=r_t\}=(\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,2:t+1]}\right)^{-1}\boldsymbol{r},\\ \hat{\sigma}_t^2 &= \mathbb{E}\{G^2\}-(\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,\,2:t+1]}\right)^{-1}(\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,1]}.\end{aligned} \tag{52}$$

We therefore have

$$\mathbb{E}\{G\mid R_1=r_1,\dots,R_t=r_t,Y=y\}=\mathbb{E}\{G\mid \hat{R}_t=\hat{r}_t,Y=y\}=\frac{\mathbb{E}_Z\left\{(\hat{r}_t+\hat{\sigma}_tZ)\dfrac{1+\operatorname{sign}(y(\hat{r}_t+\hat{\sigma}_tZ))}{2}\right\}}{\mathbb{E}_Z\left\{\dfrac{1+\operatorname{sign}(y(\hat{r}_t+\hat{\sigma}_tZ))}{2}\right\}}, \tag{53}$$

where $Z\sim\mathcal{N}(0,1)$, $\mathbb{E}_Z$ indicates that the expectation is taken over Z, and in the second line we use (50). As $\mathbb{E}\{G\mid R_1=r_1,\dots,R_t=r_t\}=\hat{r}_t$, (53) readily implies that

$$h_{t+1}(r_1,\dots,r_t,y)=\frac{\hat{\sigma}_t\mathbb{E}_Z\left\{Z\dfrac{1+\operatorname{sign}(y(\hat{r}_t+\hat{\sigma}_tZ))}{2}\right\}}{\mathbb{E}_Z\left\{\dfrac{1+\operatorname{sign}(y(\hat{r}_t+\hat{\sigma}_tZ))}{2}\right\}}=\frac{\hat{\sigma}_t\,\phi\left(\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)}{\dfrac{y+1}{2}-\Phi\left(-\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)}, \tag{54}$$

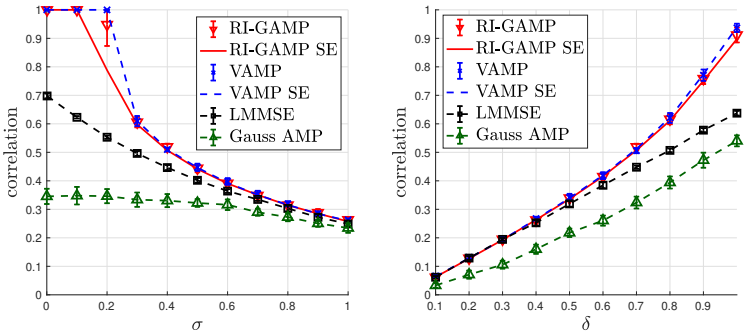
where $\phi(x)=\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, and $\Phi(x)=\int_{-\infty}^{x}\phi(t)\,\mathrm{d}t$. The second equality in (54) is obtained by computing the expectations and using the fact that $y\in\{-1,1\}$.

For the partial derivatives of h_{t+1}, we note that $\partial_{r_i}h_{t+1}=\frac{\partial h_{t+1}}{\partial\hat{r}_t}\frac{\partial\hat{r}_t}{\partial r_i}$, and

$$\frac{\partial\hat{r}_t}{\partial r_i}=(\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,2:t+1]}\right)^{-1}\boldsymbol{e}_i,\quad i\in[t].$$

Thus, by using (54) to compute $\frac{\partial h_{t+1}}{\partial\hat{r}_t}$, after some manipulations, we have that

$$\partial_{r_i}h_{t+1}(r_1,\dots,r_t,y)=\frac{-\phi^2\left(\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)-\dfrac{\hat{r}_t}{\hat{\sigma}_t}\phi\left(\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)\left(\dfrac{y+1}{2}-\Phi\left(-\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)\right)}{\left(\dfrac{y+1}{2}-\Phi\left(-\dfrac{\hat{r}_t}{\hat{\sigma}_t}\right)\right)^2}\cdot(\bar{\boldsymbol{\Sigma}}_{t+1})_{[1,2:t+1]}\left((\bar{\boldsymbol{\Sigma}}_{t+1})_{[2:t+1,2:t+1]}\right)^{-1}\boldsymbol{e}_i. \tag{55}$$



(a) $\delta = 0.7$ (b) $\sigma = 0.4$

Figure 6: Additional numerical results for linear regression with a Rademacher prior: normalized squared correlation vs. noise level σ (on the left) and vs. aspect ratio δ (on the right).

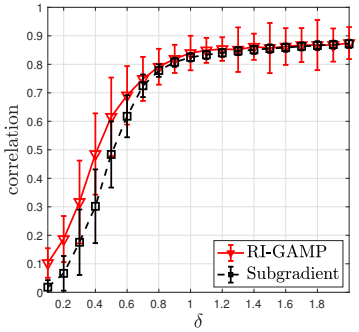


Figure 7: Comparison between the normalized squared correlation of RI-GAMP and of a subgradient method for the recovery of an image from 1-bit measurements of its wavelet transform.

Finally, for the partial derivative with respect to g, we have

$$\mathbb{E}\left\{\partial_g h_{t+1}(R_1,\dots,R_t,Y)\right\} = \mathbb{E}\left\{h_{t+1}(R_1,\dots,R_t,Y)\,\frac{\mathbb{E}\left\{G \mid R_1,\dots,R_t,Y\right\} - \mathbb{E}\left\{G \mid R_1,\dots,R_t\right\}}{\mathrm{Var}(G \mid R_1,\dots R_t)}\right\}$$
$$= \frac{1}{\hat{\sigma}_t^2}\mathbb{E}\left\{h_{t+1}(R_1,\dots,R_t,Y)^2\right\}, \tag{56}$$

where the first equality follows from Stein's lemma (see e.g. (A.8) of (Mondelli & Venkataramanan, 2021a)), and in the second equality we use the definition (29) of h_{t+1} and that $\mathrm{Var}(G \mid R_1,\dots R_t) = \hat{\sigma}_t^2$ (see (52)).

Implementation details. *RI-GAMP*: We use consistent empirical estimates for the state evolution parameters required for the posterior mean denoisers and their partial derivatives. These estimates are computed as described on p.6. To estimate the first row and column of $\bar{\mathbf{\Gamma}}_{t+1}$, we use the definition (28) and the tower property of conditional expectation:

$$(\bar{\mathbf{\Gamma}}_{t+1})_{1,i+1} = \mathbb{E}\left\{X_* f_i(X_1,\dots,X_i)\right\} = \mathbb{E}\left\{f_i(X_1,\dots,X_i)^2\right\}, \quad i \in [t]. \tag{57}$$

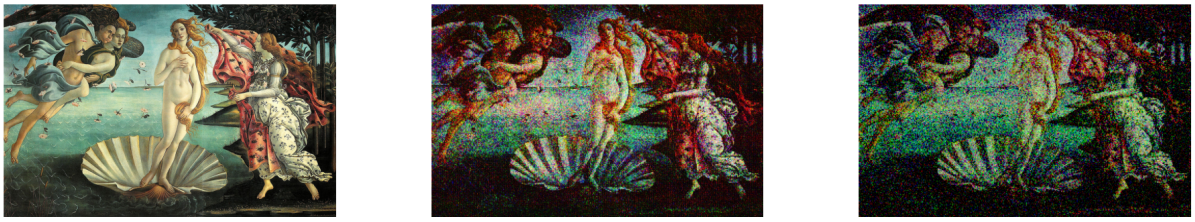
Therefore, one can consistently estimate $(\bar{\mathbf{\Gamma}}_{t+1})_{1,i+1}$ via $\|f_i(\boldsymbol{x}_1,\dots,\boldsymbol{x}_i)\|^2/d$. The partial derivatives for the matrix $\mathbf{\Psi}_{t+1}$ (defined in (6)) are computed using (39) (for Rademacher prior) and (41) (for Gaussian prior). The partial derivatives for the matrix $\mathbf{\Phi}_{t+1}$ (again, defined in (6)) are computed using (49) (for linear regression) and (55)-(56) (for 1-bit compressed sensing). For the quantity $\langle \partial_g h_1(\boldsymbol{y})\rangle$, we use the deterministic limit $\mathbb{E}\{\partial_g h_1(Y)\}$ which is given in the two settings by (45) and (51), respectively.

VAMP: Our implementation of VAMP is based on Algorithm 2 in (Schniter et al., 2016) and the corresponding state evolution is derived from (Pandit et al., 2020)[2]. To ensure numerical stability, we clipped the α_i and β_i in Algorithm 2 to lie in $[\mathtt{tol}, 1-\mathtt{tol}]$, where $\mathtt{tol} = 10^{-11}$.

C. Additional Numerical Results

In Figure 6, we provide additional numerical results for the model of linear regression with a Rademacher signal prior: on the left, we plot the normalized squared correlation as a function of σ, for $\delta = 0.7$; and on the right, we plot the same metric as function of δ, for $\sigma = 0.4$. The results showcase a similar qualitative behavior as discussed in Section 4: the performance of RI-GAMP is close to that of VAMP, except when approaching the phase transition for exact recovery ($\sigma \approx 0.2$ in Figure 6a); and RI-GAMP significantly improves upon algorithms that do not take into account the signal prior (LMMSE) or the spectrum of the noise (Gauss AMP).

[2]See also the code available at `https://sourceforge.net/projects/gampmatlab/`.



(a) Original image (b) RI-GAMP, $\delta = 0.5$ (c) Subgradient, $\delta = 0.5$

Figure 8: Reconstruction provided by RI-GAMP and by a subgradient method from 1-bit measurements of the wavelet transform of an RGB image.

In Figure 7, we consider noiseless 1-bit compressed sensing. The input $\boldsymbol{x}^*$ is the Haar wavelet transform of the RGB image in Figure 8a. We process the three channels (R, G, B) separately and the input dimension is $d = 820 \times 1280 = 1049600$. The design matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{Q}_n \boldsymbol{\Pi}_n \boldsymbol{\Lambda} \boldsymbol{\Pi}_d \boldsymbol{Q}_d$, where $\boldsymbol{Q}_n$, $\boldsymbol{Q}_d$ are orthonormal Discrete Cosine Transform (DCT) matrices in n, d dimensions, $\boldsymbol{\Pi}_n, \boldsymbol{\Pi}_d$ are random permutation matrices, and $\boldsymbol{\Lambda}$ has i.i.d. $\sqrt{6} \cdot \mathrm{Beta}(1,2)$ diagonal entries. We compare the performance of RI-GAMP against the subgradient method from (Jacques et al., 2013). For the wavelet transform we use the implementation given in (Lee et al., 2019). For RI-GAMP we use the same non-negative Bernoulli-Gaussian prior employed for the satellite image (cf. Figure 5 in Section 4). Since there is no clear way to define the true sparsity of the signal in this setting, we fix $\delta = 0.5$ and optimize over the sparsity rate for both algorithms, which yields a sparsity of $1/10$ (i.e., assuming that $1/10$ entries are non-zero) for RI-GAMP and of $1/20$ for the subgradient method. We also note that the performance of the algorithms around these values is quite stable, so we don't expect the precise choice of the sparsity rate to matter much for the chosen range of δ. In Figure 7, we report the normalized squared correlation averaged over the 3 channels and error bars at 1 standard deviation for 100 random trials. We remark that RI-GAMP improves upon the subgradient method for δ up to 1 and, for larger δ, its performance does not increase noticeably due to the already discussed numerical instabilities. The reconstructions provided by RI-GAMP and by the subgradient method for $\delta = 0.5$ are also compared in Figure 8.

D. Proof of Theorem 3.1

D.1. Debiasing Coefficients for the Auxiliary AMP

The debiasing coefficients $\{\mathsf{a}_{ti}\}_{i=1}^{t}$ and $\{\mathsf{b}_{ti}\}_{i=1}^{t-1}$ for the auxiliary AMP in (30)-(31) are defined in terms of two $t \times t$ lower triangular matrices, $\hat{\boldsymbol{\Psi}}_t$ and $\hat{\boldsymbol{\Phi}}_t$, given by

$$\hat{\boldsymbol{\Psi}}_t = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ 0 & \langle \partial_2 \boldsymbol{v}^2 \rangle & 0 & \dots & 0 \\ 0 & \langle \partial_2 \boldsymbol{v}^3 \rangle & \langle \partial_3 \boldsymbol{v}^3 \rangle & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \langle \partial_2 \boldsymbol{v}^t \rangle & \langle \partial_3 \boldsymbol{v}^t \rangle & \dots & \langle \partial_t \boldsymbol{v}^t \rangle \end{pmatrix}, \qquad \hat{\boldsymbol{\Phi}}_t = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \langle \partial_1 \boldsymbol{u}^2 \rangle & 0 & \dots & 0 & 0 \\ \langle \partial_1 \boldsymbol{u}^3 \rangle & \langle \partial_2 \boldsymbol{u}^3 \rangle & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \langle \partial_1 \boldsymbol{u}^t \rangle & \langle \partial_2 \boldsymbol{u}^t \rangle & \dots & \langle \partial_{t-1} \boldsymbol{u}^t \rangle & 0 \end{pmatrix}, \tag{58}$$

where

$$\begin{aligned} &\partial_k \boldsymbol{v}^t = \partial_{z_k} \tilde{f}_t(\boldsymbol{z}^1, \dots, \boldsymbol{z}^t, \boldsymbol{x}^*) = \partial_{k-1} f_{t-1}(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}, \dots, \boldsymbol{z}^t + \bar{\mu}_{t-1} \boldsymbol{x}^*), \qquad k \geq 2, \\ &\partial_1 \boldsymbol{u}^t = \partial_g h_{t-1}(\boldsymbol{m}^2, \dots, \boldsymbol{m}^{t-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon}))\Big|_{\boldsymbol{g}=\boldsymbol{m}^1}, \\ &\partial_k \boldsymbol{u}^t = \partial_{m_k} \tilde{h}_t(\boldsymbol{m}^1, \dots, \boldsymbol{m}^{t-1}, \boldsymbol{\varepsilon}) = \partial_{k-1} h_{t-1}(\boldsymbol{m}^2, \dots, \boldsymbol{m}^{t-1}, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})), \qquad k \geq 2. \end{aligned} \tag{59}$$

Here, $\partial_{k-1} f_{t-1}$ and $\partial_{k-1} h_{t-1}$ denote the partial derivatives with respect to the $(k-1)$-th input variable. Next, define matrices $\boldsymbol{M}_t^{\mathsf{a}}, \boldsymbol{M}_t^{\mathsf{b}} \in \mathbb{R}^{t \times t}$ as:

$$\begin{aligned} \boldsymbol{M}_t^{\mathsf{a}} &= \sum_{j=0}^{t} \kappa_{2(j+1)} \hat{\boldsymbol{\Psi}}_t (\hat{\boldsymbol{\Phi}}_t \hat{\boldsymbol{\Psi}}_t)^j, \\ \boldsymbol{M}_t^{\mathsf{b}} &= \delta \sum_{j=0}^{t-1} \kappa_{2(j+1)} \hat{\boldsymbol{\Phi}}_t (\hat{\boldsymbol{\Psi}}_t \hat{\boldsymbol{\Phi}}_t)^j. \end{aligned} \tag{60}$$

Then, the coefficients $\{\mathsf{a}_{ti}\}_{i=1}^t$ and $\{\mathsf{b}_{ti}\}_{i=1}^{t-1}$ are obtained from the last row of $\boldsymbol{M}_t^{\mathsf{a}}$ and $\boldsymbol{M}_t^{\mathsf{b}}$ as:

$$\begin{aligned} (\mathsf{a}_{t1}, \dots, \mathsf{a}_{tt}) &= (\, (\boldsymbol{M}_t^{\mathsf{a}})_{t,1}, \dots, (\boldsymbol{M}_t^{\mathsf{a}})_{t,t} \,), \\ (\mathsf{b}_{t1}, \dots, \mathsf{b}_{t,t-1}) &= (\, (\boldsymbol{M}_t^{\mathsf{b}})_{t,1}, \dots, (\boldsymbol{M}_t^{\mathsf{b}})_{t,t-1} \,). \end{aligned} \tag{61}$$

D.2. State Evolution for Auxiliary AMP

Using Theorem 2.6 of (Zhong et al., 2021), we first establish a state evolution result for the auxiliary AMP (30)-(31). We will show in Proposition D.1 that the joint empirical distribution of $(\boldsymbol{m}^1, \dots, \boldsymbol{m}^t)$ converges to a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Sigma}}_t)$, and the joint empirical distribution of $(\boldsymbol{z}^1, \dots, \boldsymbol{z}^t)$ converges to a a t-dimensional Gaussian $\mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Omega}}_t)$.

The covariance matrices are defined recursively for $t \geq 1$, starting with $\widetilde{\boldsymbol{\Omega}}_1 = 0$ and $\widetilde{\boldsymbol{\Sigma}}_1 = \bar{\kappa}_2 \mathbb{E}\{X_*^2\}$. Given $(\widetilde{\boldsymbol{\Omega}}_t, \widetilde{\boldsymbol{\Sigma}}_t)$, let

$$(Z_1 = 0, Z_2, \dots, Z_t) \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Omega}}_t), \quad V_t = f_{t-1}(Z_2 + \bar{\mu}_1 X_*,\, Z_3 + \bar{\mu}_2 X_*,\, \dots,\, Z_t + \bar{\mu}_{t-1}, X_*), \tag{62}$$

$$(M_1, \dots, M_t) \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Sigma}}_t), \quad U_{t+1} = h_t(M_2, \dots, M_t,\, q(M_1, \varepsilon)). \tag{63}$$

In (62), $(Z_2, \dots, Z_t)$ and X_* are independent, and we define $V_1 = X_*$ and $U_1 = 0$.

Let $\widetilde{\boldsymbol{\Delta}}_{t+1}, \widetilde{\boldsymbol{\Gamma}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ be symmetric matrices with entries given by

$$(\widetilde{\boldsymbol{\Delta}}_{t+1})_{ij} = \mathbb{E}\{U_i U_j\}, \quad (\widetilde{\boldsymbol{\Gamma}}_{t+1})_{ij} = \mathbb{E}\{V_i V_j\}, \quad 1 \leq i, j \leq (t+1). \tag{64}$$

Furthermore, we define $\widetilde{\boldsymbol{\Psi}}_{t+1}, \widetilde{\boldsymbol{\Phi}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ as the deterministic versions of the matrices $\hat{\boldsymbol{\Psi}}_{t+1}, \hat{\boldsymbol{\Phi}}_{t+1}$. Specifically,

$$\widetilde{\boldsymbol{\Psi}}_{t+1} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & \mathbb{E}\{\partial_1 f_1\} & 0 & \dots & 0 \\ 0 & \mathbb{E}\{\partial_1 f_2\} & \mathbb{E}\{\partial_2 f_2\} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mathbb{E}\{\partial_1 f_t\} & \mathbb{E}\{\partial_2 f_t\} & \dots & \mathbb{E}\{\partial_t f_t\} \end{pmatrix}, \quad \widetilde{\boldsymbol{\Phi}}_{t+1} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \mathbb{E}\{\partial_g h_1\} & 0 & \dots & 0 & 0 \\ \mathbb{E}\{\partial_g h_2\} & \mathbb{E}\{\partial_1 h_2\} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbb{E}\{\partial_g h_t\} & \mathbb{E}\{\partial_1 h_t\} & \dots & \mathbb{E}\{\partial_{t-1} h_t\} & 0 \end{pmatrix}, \tag{65}$$

where we have used the shorthand

$$\begin{aligned} \mathbb{E}\{\partial_k f_\ell\} &\equiv \mathbb{E}\{\partial_k f_\ell(Z_2 + \bar{\mu}_1 X_*,\, Z_3 + \bar{\mu}_2 X_*,\, \dots,\, Z_{\ell+1} + \bar{\mu}_\ell X_*)\}, \\ \mathbb{E}\{\partial_g h_\ell\} &\equiv \mathbb{E}\{\partial_g h_\ell(M_2, \dots, M_\ell, q(g, \varepsilon))|_{g=M_1}\}, \\ \mathbb{E}\{\partial_k h_\ell\} &\equiv \mathbb{E}\{\partial_k h_\ell(M_2, \dots, M_\ell, q(M_1, \varepsilon))\}. \end{aligned} \tag{66}$$

From these matrices, we compute the covariances $\widetilde{\boldsymbol{\Sigma}}_{t+1}, \widetilde{\boldsymbol{\Omega}}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ as:

$$\widetilde{\boldsymbol{\Sigma}}_{t+1} = \sum_{j=0}^{2t+1} \bar{\kappa}_{2(j+1)}\, \widetilde{\boldsymbol{\Xi}}_{t+1}^{(j)}, \qquad \widetilde{\boldsymbol{\Omega}}_{t+1} = \delta \sum_{j=0}^{2t} \bar{\kappa}_{2(j+1)}\, \widetilde{\boldsymbol{\Theta}}_{t+1}^{(j)}, \tag{67}$$

where $\widetilde{\boldsymbol{\Xi}}_{t+1}^{(0)} = \widetilde{\boldsymbol{\Gamma}}_{t+1}$, $\widetilde{\boldsymbol{\Theta}}_{t+1}^{0} = \widetilde{\boldsymbol{\Delta}}_{t+1}$, and for $j \geq 1$:

$$\begin{aligned} \widetilde{\boldsymbol{\Xi}}_{t+1}^{(j)} &= \sum_{i=0}^{j} (\widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1})^i\, \widetilde{\boldsymbol{\Gamma}}_{t+1} \Big((\widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1})^{\mathsf{T}} \Big)^{j-i} + \sum_{i=0}^{j-1} (\widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1})^i\, \widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Delta}}_{t+1} \widetilde{\boldsymbol{\Psi}}_{t+1}^{\mathsf{T}} \Big((\widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1})^{\mathsf{T}} \Big)^{j-i-1}, \\ \widetilde{\boldsymbol{\Theta}}_{t+1}^{(j)} &= \sum_{i=0}^{j} (\widetilde{\boldsymbol{\Phi}}_{t+1} \widetilde{\boldsymbol{\Psi}}_{t+1})^i\, \widetilde{\boldsymbol{\Delta}}_{t+1} \Big((\widetilde{\boldsymbol{\Phi}}_{t+1} \widetilde{\boldsymbol{\Psi}}_{t+1})^{\mathsf{T}} \Big)^{j-i} + \sum_{i=0}^{j-1} (\widetilde{\boldsymbol{\Phi}}_{t+1} \widetilde{\boldsymbol{\Psi}}_{t+1})^i\, \widetilde{\boldsymbol{\Phi}}_{t+1} \widetilde{\boldsymbol{\Gamma}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1}^{\mathsf{T}} \Big((\widetilde{\boldsymbol{\Phi}}_{t+1} \widetilde{\boldsymbol{\Psi}}_{t+1})^{\mathsf{T}} \Big)^{j-i-1}. \end{aligned} \tag{68}$$

We note that computing the last column and row of $\widetilde{\mathbf{\Gamma}}_{t+1}, \widetilde{\mathbf{\Psi}}_{t+1}$ requires knowledge of $\widetilde{\mathbf{\Omega}}_{t+1}$. However, these entries are zeroed out in the computation of $\widetilde{\mathbf{\Xi}}^{(j)}_{t+1}, \widetilde{\mathbf{\Theta}}^{(j)}_{t+1}$ in (68), and hence only $\widetilde{\mathbf{\Gamma}}_t, \widetilde{\mathbf{\Psi}}_t$ are required to compute $\widetilde{\mathbf{\Sigma}}_{t+1}, \widetilde{\mathbf{\Omega}}_{t+1}$ via (67).

Proposition D.1 (State evolution for auxiliary AMP)**.** *Consider the auxiliary AMP in* (30)-(31) *and the state evolution random variables defined in* (62)-(63)*. Let $\tilde{\psi} : \mathbb{R}^{2t+1} \to \mathbb{R}$ and $\tilde{\phi} : \mathbb{R}^{2t+2} \to \mathbb{R}$ be any pseudo-Lipschitz functions of order 2. Then for each $t \geq 1$, we almost surely have*

$$\lim_{n\to\infty} \frac{1}{d} \sum_{i=1}^{d} \tilde{\psi}(z_i^2, \dots, z_i^{t+1}, v_i^2, \dots, v_i^{t+1}, x_i^*) = \mathbb{E}\{\tilde{\psi}(Z_2, \dots Z_{t+1}, V_2, \dots, V_{t+1}, X_*)\}, \tag{69}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(m_i^1, \dots, m_i^t, u_i^1, \dots, u_i^{t+1}, \varepsilon_i) = \mathbb{E}\{\tilde{\phi}(M_1, \dots, M_t, U_1, \dots, U_{t+1}, \varepsilon)\}. \tag{70}$$

Equivalently, as $n \to \infty$, almost surely:

$$(\boldsymbol{z}^2, \dots, \boldsymbol{z}^{t+1}, \boldsymbol{v}^2, \dots, \boldsymbol{v}^{t+1}, \boldsymbol{x}^*) \xrightarrow{W_2} (Z_2, \dots, Z_t, V_2, \dots, V_{t+1}, X_*),$$
$$(\boldsymbol{m}^1, \dots, \boldsymbol{m}^t, \boldsymbol{u}^1, \dots, \boldsymbol{u}^{t+1}, \boldsymbol{\varepsilon}) \xrightarrow{W_2} (M_1, \dots, M_t, U_1, \dots, U_{t+1}, \varepsilon).$$

The proposition follows directly from Theorem 2.6 in (Zhong et al., 2021) as the auxiliary AMP in (30)-(31) is of the standard form for which that state evolution result applies. That result is proved for $\delta = \frac{n}{d} \leq 1$ under two sets of assumptions (cf. Assumptions 2.4 and 2.5 in (Zhong et al., 2021)). The first set of assumptions concerns the design matrix, and these coincide with the ones we describe in Section 2. The second set concerns the empirical distribution of the signal and noise vectors, and the functions $\tilde{f}_t, \tilde{h}_t$ used in the auxiliary AMP. This set of assumptions is also satisfied since $\boldsymbol{x}^* \in \mathbb{R}^d$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ are both independent of the design matrix $\boldsymbol{A}$ and satisfy $\boldsymbol{x}^* \xrightarrow{W_2} X_*$ and $\boldsymbol{\varepsilon} \xrightarrow{W_2} \varepsilon$. Furthermore, our assumption **(A1)** (see p.5) ensures that the required Lipschitz and continuity conditions on $\tilde{f}_t, \tilde{h}_t$ and their partial derivatives are satisfied. Therefore, for $\delta \leq 1$, the iteration in (30)-(31) satisfies all the assumptions under which Theorem 2.6 in (Zhong et al., 2021) holds. Finally, for the case $\delta > 1$, we can rewrite the auxiliary AMP in terms of $\boldsymbol{A}' \equiv \boldsymbol{A}^\mathsf{T}$ and then apply Theorem 2.6 in (Zhong et al., 2021).

We conclude this section by showing that the state evolution of the auxiliary AMP described above is equivalent to the state evolution of the proposed AMP algorithm described in Section 3.

Lemma D.2 (Equivalence of state evolution between true and auxiliary AMP)**.** *For $t \geq 1$, we have that*

$$\begin{aligned}(M_1, \dots, M_t) &\overset{\mathrm{d}}{=} (G, R_1, \dots, R_{t-1}) \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{\Sigma}}_t), \\ (Z_2, \dots, Z_{t+1}) &\overset{\mathrm{d}}{=} (X_1 - \bar{\mu}_1 X_*, \dots, X_t - \bar{\mu}_t X_*) \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{\Omega}}_t),\end{aligned} \tag{71}$$

where the random variables on the left are defined in (62)-(63)*, and the random variables on the right are defined in* (12)-(13)*.*

Proof. We will prove by induction that $\widetilde{\mathbf{\Sigma}}_t = \bar{\mathbf{\Sigma}}_t$ and $\widetilde{\mathbf{\Omega}}_{t+1} = \mathbf{\Omega}'_{t+1}$, where the matrices on the left are defined via (67) and the matrices on the right are defined via (17) and (19). The result of the lemma then follows since $(M_1, \dots, M_t) \sim \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_t)$, $(Z_1 = 0, Z_2, \dots, Z_{t+1}) \sim \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Omega}}_{t+1})$ and $\bar{\mathbf{\Omega}}_t$ is the lower right $t \times t$ submatrix of $\mathbf{\Omega}'_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$.

For $t = 1$, by the initialization in (63), $M_1 \sim \mathcal{N}(0, \widetilde{\mathbf{\Sigma}}_1)$ where $\widetilde{\mathbf{\Sigma}}_1 = \kappa_2 \mathbb{E}\{X_*^2\} = \bar{\mathbf{\Sigma}}_1$, where the last equality is from (11). From (67), the matrix $\widetilde{\mathbf{\Omega}}_2$ can be computed as

$$\widetilde{\mathbf{\Omega}}_2 = \begin{pmatrix} 0 & 0 \\ 0 & \delta\bar{\kappa}_2 \mathbb{E}\{h_1(q(M_1, \varepsilon))^2\} + \delta\bar{\kappa}_4 \mathbb{E}\{V_1^2\}(\mathbb{E}\{\partial_g h_1(q(M_1, \varepsilon))\})^2 \end{pmatrix}. \tag{72}$$

Since $V_1 \overset{\mathrm{d}}{=} X_*$ and $M_1 \overset{\mathrm{d}}{=} G$, the matrix above equals $\mathbf{\Omega}'_2$ (defined via (19)).

Assume towards induction that $\widetilde{\mathbf{\Sigma}}_k = \bar{\mathbf{\Sigma}}_k$ and $\widetilde{\mathbf{\Omega}}_{k+1} = \mathbf{\Omega}'_{k+1}$ for some $k \geq 1$. Recalling that $(M_1, \dots, M_k) \sim \mathcal{N}(0, \widetilde{\mathbf{\Sigma}}_k)$ and $(Z_1 = 0, \dots, Z_{k+1}) \sim \widetilde{\mathbf{\Omega}}_{k+1}$, using the induction hypothesis in the definitions of $\widetilde{\mathbf{\Delta}}_{k+1}, \widetilde{\mathbf{\Gamma}}_{k+1}, \widetilde{\mathbf{\Phi}}_{k+1}, \widetilde{\mathbf{\Psi}}_{k+1}$ in

(64)-(66), we obtain that

$$\widetilde{\bar{\mathbf{\Delta}}}_{k+1} = \bar{\mathbf{\Delta}}_{k+1}, \quad \widetilde{\bar{\mathbf{\Gamma}}}_{k+1} = \bar{\mathbf{\Gamma}}_{k+1}, \quad \widetilde{\bar{\mathbf{\Phi}}}_{k+1} = \bar{\mathbf{\Phi}}_{k+1}, \quad \widetilde{\bar{\mathbf{\Psi}}}_{k+1} = \bar{\mathbf{\Psi}}_{k+1}, \tag{73}$$

where $\bar{\mathbf{\Delta}}_{k+1}, \bar{\mathbf{\Gamma}}_{k+1}, \bar{\mathbf{\Phi}}_{k+1}, \bar{\mathbf{\Psi}}_{k+1}$ are defined via $\bar{\mathbf{\Sigma}}_k$ and $\bar{\mathbf{\Omega}}_k$ in (14)-(16). It follows from the definitions in (17) and (67) that $\widetilde{\bar{\mathbf{\Sigma}}}_{k+1} = \bar{\mathbf{\Sigma}}_{k+1}$. This then implies that $\widetilde{\bar{\mathbf{\Phi}}}_{k+2} = \bar{\mathbf{\Phi}}_{k+2}$ and $\widetilde{\bar{\mathbf{\Delta}}}_{k+2} = \bar{\mathbf{\Delta}}_{k+2}$. Using these in the definitions in (68) and (20), we obtain that $\widetilde{\mathbf{\Theta}}^{(j)}_{k+2} = \mathbf{\Theta}^{(j)}_{k+2}$, and consequently, $\widetilde{\mathbf{\Omega}}_{k+2} = \mathbf{\Omega}'_{k+2}$. This completes the proof of the induction step, and gives the desired result. □

D.3. Proof of Theorem 3.1

At this point, Theorem 3.1 follows from the following intermediate result, whose proof is deferred to Section D.4.

Lemma D.3. *For any order 2 pseudo-Lipschitz functions $\psi : \mathbb{R}^{2t+1} \to \mathbb{R}$ and $\phi : \mathbb{R}^{2t+2} \to \mathbb{R}$, the following limits hold almost surely for $t \geq 1$:*

$$\lim_{n\to\infty} \left| \frac{1}{d}\sum_{i=1}^{d} \psi(x_i^1, \dots, x_i^t, \hat{x}_i^1, \dots, \hat{x}_i^t, x_i^*) \; - \; \frac{1}{d}\sum_{i=1}^{d} \psi(z_i^2 + \bar{\mu}_1 x_i^*, \dots, z_i^{t+1} + \bar{\mu}_t x_i^*, v_i^2, \dots, v_i^{t+1}, x_i^*) \right| = 0, \tag{74}$$

$$\lim_{n\to\infty} \left| \frac{1}{n}\sum_{i=1}^{n} \phi(r_i^1, \dots, r_i^t, s_i^1, \dots, s_i^{t+1}, y_i) \; - \; \frac{1}{n}\sum_{i=1}^{n} \phi(m_i^2, \dots, m_i^{t+1}, u_i^2, \dots, u_i^{t+2}, q(m_i^1, \varepsilon_i)) \right| = 0. \tag{75}$$

Proof of Theorem 3.1. Applying (69) to the pseudo-Lipschitz function

$$\tilde{\psi}(z_2, \dots, z_{t+1}, v_2, \dots, v_{t+1}, x_*) = \psi(z_2 + \bar{\mu}_1 x_*, \dots, z_{t+1} + \bar{\mu}_t x_*, v_2, \dots, v_{t+1}, x_*),$$

we obtain that almost surely

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(z_i^2 + \bar{\mu}_1 x_i^*, \dots, z_i^{t+1} + \bar{\mu}_t x_i^*, v_i^1, \dots, v_i^t, x^*) &= \mathbb{E}\{\psi(Z_2 + \bar{\mu}_1 X_*, \dots, Z_{t+1} + \bar{\mu}_t X_*, V_2, \dots, V_{t+1}, X_*)\} \\ &= \mathbb{E}\{\psi(X_1, \dots, X_t, \hat{X}_1, \dots, \hat{X}_t, X_*)\}, \end{aligned} \tag{76}$$

where the last equality follows from Lemma D.2, by recalling that $V_{\ell+1} = f_\ell(Z_2 + \bar{\mu}_1 X_*, \dots, Z_{\ell+1} + \bar{\mu}_\ell X_*)$ and $\hat{X}_\ell = f_\ell(X_1, \dots, X_\ell)$, for $\ell \geq 1$. Combining (76) with (74) yields (23) of Theorem 3.1. The result (24) is obtained similarly, using (70), Lemma D.2, and (75). □

D.4. Proof of Lemma D.3

Throughout, we use C to denote a generic positive constant. All the limits in the proof hold almost surely, so we don't explicitly state this each time.

Since ψ is pseudo-Lipschitz, we have

$$\begin{aligned} &\left| \frac{1}{d}\sum_{i=1}^{d} \psi(x_i^1, \dots, x_i^t, \hat{x}_i^1, \dots, \hat{x}_i^t, x_i^*) \; - \; \frac{1}{d}\sum_{i=1}^{d} \psi(z_i^2 + \bar{\mu}_1 x_i^*, \dots, z_i^{t+1} + \bar{\mu}_t x_i^*, v_i^2, \dots, v_i^{t+1}, x_i^*) \right| \\ &\leq \frac{C}{d}\sum_{i=1}^{d} \left(1 + |x_i^*| + \sum_{\ell=1}^{t} \Big(|x_i^\ell| + |\hat{x}_i^\ell| + |z_i^{\ell+1} + \bar{\mu}_\ell x_i^*| + |v_i^{\ell+1}| \Big) \right) \cdot \left(\sum_{\ell=1}^{t} \big(|x_i^\ell - z_i^{\ell+1} - \bar{\mu}_\ell x_i^*|^2 \, + \, |\hat{x}_i^\ell - v_i^{\ell+1}|^2 \big) \right)^{\frac{1}{2}} \\ &\leq C(4t+2) \left[1 + \frac{\|\boldsymbol{x}^*\|^2}{d} + \sum_{\ell=1}^{t} \Big(\frac{\|\boldsymbol{x}^\ell\|^2}{d} + \frac{\|\hat{\boldsymbol{x}}^\ell\|^2}{d} + \frac{\|\boldsymbol{z}^{\ell+1} + \bar{\mu}_\ell \, \boldsymbol{x}^*\|^2}{d} + \frac{\|\boldsymbol{v}^{\ell+1}\|^2}{d} \Big) \right]^{\frac{1}{2}} \\ &\qquad \cdot \left(\sum_{\ell=1}^{t} \left(\frac{\|\boldsymbol{x}^\ell - \boldsymbol{z}^{\ell+1} - \bar{\mu}_\ell \boldsymbol{x}^*\|^2}{d} + \frac{\|\hat{\boldsymbol{x}}^\ell - \boldsymbol{v}^{\ell+1}\|^2}{d} \right) \right)^{\frac{1}{2}}, \end{aligned} \tag{77}$$

where the last step uses Cauchy-Schwarz inequality (twice). Similarly, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^{n}\phi(r_i^1,\dots,r_i^t,s_i^1,\dots,s_i^{t+1},\,y_i)\;-\;\frac{1}{n}\sum_{i=1}^{n}\phi(m_i^2,\dots,m_i^{t+1},u_i^2,\dots,u_i^{t+2},\,q(m_i^1,\varepsilon_i))\right|$$
$$\le C(4t+5)\left[1+\frac{\|\boldsymbol{y}\|^2}{n}+\frac{\|q(\boldsymbol{m}^1,\boldsymbol{\varepsilon})\|^2}{n}+\sum_{\ell=1}^{t}\left(\frac{\|\boldsymbol{r}^\ell\|^2}{n}+\frac{\|\boldsymbol{m}^{\ell+1}\|^2}{n}\right)+\sum_{\ell=1}^{t+1}\left(\frac{\|\boldsymbol{s}^\ell\|^2}{n}+\frac{\|\boldsymbol{u}^{\ell+1}\|^2}{n}\right)\right]^{\frac{1}{2}}$$
$$\cdot\left(\sum_{\ell=1}^{t}\Big(\frac{\|\boldsymbol{r}^\ell-\boldsymbol{m}^{\ell+1}\|^2}{n}+\frac{\|\boldsymbol{s}^\ell-\boldsymbol{u}^{\ell+1}\|^2}{n}\Big)+\frac{\|\boldsymbol{s}^{t+1}-\boldsymbol{u}^{t+2}\|^2}{n}+\frac{\|\boldsymbol{y}-q(\boldsymbol{m}^1,\boldsymbol{\varepsilon})\|^2}{n}\right)^{\frac{1}{2}}. \tag{78}$$

We will inductively show that as $n\to\infty$: *(i)* each of the terms in the last line of (77) and (78) converges to zero, and *(ii)* the terms within the square brackets in (77) and (78) all converge to finite, deterministic limits.

Base case $t=1$: Consider (77) for $t=1$. From the AMP initialization, we have $\boldsymbol{x}^1=\boldsymbol{A}^\mathsf{T}h_1(\boldsymbol{y})$, and from (30), we have

$$\boldsymbol{z}^2=\boldsymbol{A}^\mathsf{T}h_1(q(\boldsymbol{m}^1,\boldsymbol{\varepsilon}))\;-\;\mathsf{b}_{21}\boldsymbol{v}^1=\boldsymbol{A}^\mathsf{T}h_1(\boldsymbol{y})\;-\;\delta\kappa_2\langle\partial_g h_1(q(\boldsymbol{g},\boldsymbol{\varepsilon}))\rangle\,\boldsymbol{x}^*, \tag{79}$$

where the last equality is obtained by recalling that $\boldsymbol{m}^1=\boldsymbol{g}$, $\boldsymbol{v}^1=\boldsymbol{x}^*$, and computing the matrix $\boldsymbol{M}_2^{\mathsf{b}}$ in (60) to verify that $\mathsf{b}_{21}=\delta\kappa_2\langle\partial_g h_1(q(\boldsymbol{g},\boldsymbol{\varepsilon}))\rangle$. We therefore have

$$\begin{aligned}\frac{\|\boldsymbol{x}^1-\boldsymbol{z}^2-\bar\mu_1\boldsymbol{x}^*\|^2}{d}&=\frac{\|\boldsymbol{x}^*\|^2}{d}\left(\delta\kappa_2\langle\partial_g h_1(q(\boldsymbol{g},\boldsymbol{\varepsilon}))\rangle-\bar\mu_1\right)^2\\&=\frac{\|\boldsymbol{x}^*\|^2}{d}\delta^2\left[\kappa_2(\langle\partial_g h_1(q(\boldsymbol{g},\boldsymbol{\varepsilon}))\rangle-\mathbb{E}\{\partial_g h_1(q(G,\varepsilon))\})\;+\;\mathbb{E}\{\partial_g h_1(q(G,\varepsilon)\}(\kappa_2-\bar\kappa_2)\right]^2,\end{aligned} \tag{80}$$

where for the last equality, we use the definition of $\bar\mu_1$ in (11). By the assumptions of the theorem, $\|\boldsymbol{x}^*\|^2/d\to\mathbb{E}\{X_*^2\}$. Since $\boldsymbol{m}^1=\boldsymbol{g}$, applying Proposition D.1 for $t=1$ gives

$$(\boldsymbol{g},\boldsymbol{\varepsilon})\xrightarrow{W_2}(G,\varepsilon). \tag{81}$$

Since $h_1(q(g,\varepsilon))$ is Lipschitz in each argument, (81) together with Lemma E.1 implies that $\langle\partial_g h_1(q(\boldsymbol{g},\boldsymbol{\varepsilon}))\rangle\to\mathbb{E}\{\partial_g h_1(q(G,\varepsilon))\}$. Furthermore, by the model assumptions $\kappa_2\to\bar\kappa_2$. Therefore,

$$\lim_{n\to\infty}\frac{1}{d}\|\boldsymbol{x}^1-\boldsymbol{z}^2-\bar\mu_1\boldsymbol{x}^*\|^2=0. \tag{82}$$

Since $\hat{\boldsymbol{x}}^1=f_1(\boldsymbol{x}^1)$ and $\boldsymbol{v}^2=f_1(\boldsymbol{z}^2+\bar\mu_1\boldsymbol{x}^*)$ with f_1 being Lipschitz, we have

$$\frac{\|\hat{\boldsymbol{x}}^1-\boldsymbol{v}^2\|}{d}\le C\,\frac{1}{d}\|\boldsymbol{x}^1-\boldsymbol{z}^2-\bar\mu_1\boldsymbol{x}^*\|^2\to 0,\;\;\text{as } n\to\infty. \tag{83}$$

Now consider the terms inside the square brackets in (77). Using Proposition D.1 and Lemma D.2, we have the following limits for $t\ge 1$:

$$\lim_{n\to\infty}\frac{\|\boldsymbol{z}^{t+1}+\bar\mu_t\boldsymbol{x}^*\|^2}{d}=\mathbb{E}\{(Z_{t+1}+\bar\mu_t X_*)^2\}=\mathbb{E}\{X_t^2\},\qquad\lim_{n\to\infty}\frac{\|\boldsymbol{v}^{t+1}\|^2}{d}=\mathbb{E}\{V_{t+1}^2\}=\mathbb{E}\{\hat X_t^2\}. \tag{84}$$

Using the triangle inequality, we have the following lower and upper bounds, for $t\ge 1$:

$$\begin{aligned}\|\boldsymbol{z}^{t+1}+\bar\mu_t\boldsymbol{x}^*\|\;-\;\|\boldsymbol{x}^t-\boldsymbol{z}^{t+1}-\bar\mu_t\boldsymbol{x}^*\|\le\|\boldsymbol{x}^t\|\le\|\boldsymbol{z}^{t+1}+\bar\mu_t\boldsymbol{x}^*\|\;+\;\|\boldsymbol{x}^t-\boldsymbol{z}^{t+1}-\bar\mu_t\boldsymbol{x}^*\|,\\\|\boldsymbol{v}^{t+1}\|\;-\;\|\hat{\boldsymbol{x}}^t-\boldsymbol{v}^{t+1}\|\le\|\hat{\boldsymbol{x}}^t\|\le\|\boldsymbol{v}^{t+1}\|\;+\;\|\hat{\boldsymbol{x}}^t-\boldsymbol{v}^{t+1}\|.\end{aligned} \tag{85}$$

Using (82)-(85), we obtain

$$\lim_{n\to\infty}\frac{\|\boldsymbol{x}^1\|^2}{d}=\mathbb{E}\{X_1^2\},\qquad\lim_{n\to\infty}\frac{\|\hat{\boldsymbol{x}}^1\|^2}{d}=\mathbb{E}\{\hat X_1^2\}. \tag{86}$$

Using (82)-(86) in (77), we obtain (74) for $t=1$.

Next consider (78) for $t = 1$. From the definition of auxiliary AMP in (30)-(32), we have $\boldsymbol{m}^1 = \boldsymbol{g}$ and

$$\boldsymbol{u}^2 = h_1(q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})) = h_1(q(\boldsymbol{g}, \boldsymbol{\varepsilon})) = h_1(\boldsymbol{y}) = \boldsymbol{s}^1,$$

where the last equality holds due to the initialization of the true AMP (below (5)). We therefore have

$$\frac{\|\boldsymbol{y} - q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})\|^2}{n} = 0, \qquad \frac{\|\boldsymbol{s}^1 - \boldsymbol{u}^2\|^2}{n} = 0. \tag{87}$$

Next, from (31) we have

$$\boldsymbol{m}^2 = \boldsymbol{A}\boldsymbol{v}^2 - \mathsf{a}_{22}\, \boldsymbol{u}^2 = \boldsymbol{A}\boldsymbol{v}^2 - \kappa_2 \langle \partial_1 f_1(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*) \rangle\, h_1(\boldsymbol{y}), \tag{88}$$

where we have used $\boldsymbol{u}^1 = \boldsymbol{0}$ and the value of a_{22} obtained via $\boldsymbol{M}_2^{\mathsf{a}}$ in (60). For the true AMP, from (5) we have

$$\boldsymbol{r}^1 = \boldsymbol{A}\hat{\boldsymbol{x}}^1 - \alpha_{11}\boldsymbol{s}^1 = \boldsymbol{A}\hat{\boldsymbol{x}}^1 - \kappa_2 \langle \partial_1 f_1(\boldsymbol{x}^1) \rangle\, h_1(\boldsymbol{y}). \tag{89}$$

Combining (88) and (89), we obtain

$$\begin{aligned} \frac{\|\boldsymbol{r}^1 - \boldsymbol{m}^2\|^2}{n} &\le 2\frac{\|\boldsymbol{A}(\hat{\boldsymbol{x}}^1 - \boldsymbol{v}^2)\|^2}{n} + 2\kappa_2^2 \frac{\|h_1(\boldsymbol{y})\|^2}{n} \left(\langle \partial_1 f_1(\boldsymbol{x}^1) \rangle - \langle \partial_1 f_1(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*) \rangle \right)^2 \\ &\le 2\|\boldsymbol{A}\|_{\text{op}}^2 \frac{\|\hat{\boldsymbol{x}}^1 - \boldsymbol{v}^2\|^2}{n} + 2\kappa_2^2 \frac{\|h_1(\boldsymbol{y})\|^2}{n} \left(\langle \partial_1 f_1(\boldsymbol{x}^1) \rangle - \langle \partial_1 f_1(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*) \rangle \right)^2. \end{aligned} \tag{90}$$

By assumption, the empirical distribution of $\boldsymbol{\lambda}$, the vector of singular values, converges to Λ which has compact support. Therefore, $\|\boldsymbol{A}\|_{\text{op}} \le C$, and by (83), the first term above tends to zero. We also have $\kappa_2 \to \bar{\kappa}_2$ and $\|h_1(\boldsymbol{y})\|^2/n \to \mathbb{E}\{h_1(Y)^2\}$. Since f_1 is Lipschitz and we have shown above that $\boldsymbol{x}^1 \xrightarrow{W_2} X_1$, Lemma E.1 implies that

$$\lim_{n\to\infty} \langle \partial_1 f_1(\boldsymbol{x}^1) \rangle = \mathbb{E}\{\partial_1 f_1(X_1)\}. \tag{91}$$

Similarly, since by Proposition D.1 we have $\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^* \xrightarrow{W_2} (Z_2 + \bar{\mu}_1 X_*)$, Lemma E.1 implies

$$\lim_{n\to\infty} \langle \partial_1 f_1(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*) \rangle = E\{\partial_1 f_1(Z_2 + \bar{\mu}_1 \boldsymbol{x}^*) = E\{\partial_1 f_1(X_1)\}, \tag{92}$$

where the last equality follows from Lemma D.2. Using (91) and (92) in (90), we have

$$\lim_{n\to\infty} \frac{\|\boldsymbol{r}^1 - \boldsymbol{m}^2\|^2}{n} = 0. \tag{93}$$

Since h_2 is Lipschitz in each argument, (93) also implies that

$$\lim_{n\to\infty} \frac{\|\boldsymbol{s}^2 - \boldsymbol{u}^3\|^2}{n} = \lim_{n\to\infty} \frac{\|h_2(\boldsymbol{r}^1,\, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) - h_2(\boldsymbol{m}^2,\, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon}))\|^2}{n} = 0, \tag{94}$$

where we have used $\boldsymbol{m}^1 = \boldsymbol{g}$. Eqs. (87), (93) and (94) show that for $t = 1$, each term on the last line of (78) tends to zero. Using Proposition D.1 and Lemma D.2, we have for $t \ge 1$:

$$\begin{aligned} &\lim_{n\to\infty} \frac{\|\boldsymbol{m}^{t+1}\|^2}{n} = \mathbb{E}\{M_{t+1}^2\} = \mathbb{E}\{R_t^2\}, \qquad \lim_{n\to\infty} \frac{\|\boldsymbol{u}^{t+1}\|^2}{n} = \mathbb{E}\{U_{t+1}^2\} = \mathbb{E}\{S_t^2\}, \\ &\lim_{n\to\infty} \frac{\|q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})\|^2}{n} = \lim_{n\to\infty} \frac{\|\boldsymbol{y}\|^2}{n} = \mathbb{E}\{q(M_1, \varepsilon)^2\} = \mathbb{E}\{Y^2\}. \end{aligned} \tag{95}$$

Using the triangle inequality, we have the following lower and upper bounds:

$$\begin{aligned} \|\boldsymbol{m}^{t+1}\| - \|\boldsymbol{r}^t - \boldsymbol{m}^{t+1}\| \le \|\boldsymbol{r}^t\| \le \|\boldsymbol{m}^{t+1}\| + \|\boldsymbol{r}^t - \boldsymbol{m}^{t+1}\|, \\ \|\boldsymbol{u}^{t+2}\| - \|\boldsymbol{s}^{t+1} - \boldsymbol{u}^{t+2}\| \le \|\boldsymbol{s}^{t+1}\| \le \|\boldsymbol{u}^{t+2}\| + \|\boldsymbol{s}^{t+1} - \boldsymbol{u}^{t+2}\|. \end{aligned} \tag{96}$$

Combining (96) with (93)-(95), we obtain the following limits:

$$\lim_{n\to\infty} \frac{\|\boldsymbol{r}^1\|^2}{n} = \mathbb{E}\{R_1^2\}, \qquad \lim_{n\to\infty} \frac{\|\boldsymbol{s}^2\|^2}{n} = \mathbb{E}\{S_2^2\}. \tag{97}$$

Using the limits in (95) and (97) in (78) yields the result (75) for $t = 1$.

Induction step: Assume towards induction that the results (74)-(75) hold with t replaced by $(t-1)$ for some $(t-1) \geq 1$, and that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\|\boldsymbol{x}^\ell - \boldsymbol{z}^{\ell+1} - \bar{\mu}_\ell \boldsymbol{x}^*\|^2}{d} = 0, \quad \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{x}}^\ell - \boldsymbol{v}^{\ell+1}\|^2}{d} = 0, \\
&\lim_{n\to\infty} \frac{\|\boldsymbol{r}^\ell - \boldsymbol{m}^{\ell+1}\|^2}{n} = 0, \quad \lim_{n\to\infty} \frac{\|\boldsymbol{s}^{\ell+1} - \boldsymbol{u}^{\ell+2}\|^2}{n} = 0, \\
&\lim_{n\to\infty} \frac{\|\boldsymbol{x}^\ell\|^2}{d} = \mathbb{E}\{X_\ell^2\}, \ \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{x}}^\ell\|^2}{d} = \mathbb{E}\{\hat{X}_\ell^2\}, \ \lim_{n\to\infty} \frac{\|\boldsymbol{r}^\ell\|^2}{n} = \mathbb{E}\{R_\ell^2\}, \ \lim_{n\to\infty} \frac{\|\boldsymbol{s}^{\ell+1}\|^2}{n} = \mathbb{E}\{S_{\ell+1}^2\}, \quad \text{for } \ell \in [t-1].
\end{aligned} \tag{98}$$

From the definitions of $\boldsymbol{x}^t$ and $\boldsymbol{z}^{t+1}$ (see (4) and (30)), we have

$$\begin{aligned}
\boldsymbol{x}^t - \boldsymbol{z}^{t+1} - \bar{\mu}_t \boldsymbol{x}^* &= \boldsymbol{A}^\mathsf{T}(\boldsymbol{s}^t - \boldsymbol{u}^{t+1}) + \sum_{i=1}^{t-1} (\mathsf{b}_{t+1,i+1} \boldsymbol{v}^{i+1} - \beta_{ti} \hat{\boldsymbol{x}}^i) + \mathsf{b}_{t+1,1} \boldsymbol{v}^1 - \bar{\mu}_t \boldsymbol{x}^* \\
&= \boldsymbol{A}^\mathsf{T}(\boldsymbol{s}^t - \boldsymbol{u}^{t+1}) + \sum_{i=1}^{t-1} \mathsf{b}_{t+1,i+1}(\boldsymbol{v}^{i+1} - \hat{\boldsymbol{x}}^i) + \sum_{i=1}^{t-1} (\mathsf{b}_{t+1,i+1} - \beta_{ti}) \hat{\boldsymbol{x}}^i + (\mathsf{b}_{t+1,1} - \bar{\mu}_t) \boldsymbol{x}^*,
\end{aligned} \tag{99}$$

where we have used the fact that $\boldsymbol{v}^1 = \boldsymbol{x}^*$. Using Cauchy-Schwarz inequality, we then have

$$\begin{aligned}
\frac{\|\boldsymbol{x}^t - \boldsymbol{z}^{t+1} - \bar{\mu}_t \boldsymbol{x}^*\|^2}{n} &\leq 2t \Bigg[\|\boldsymbol{A}\|_{\text{op}}^2 \frac{\|\boldsymbol{s}^t - \boldsymbol{u}^{t+1}\|^2}{n} + \sum_{i=1}^{t-1} \mathsf{b}_{t+1,i+1}^2 \frac{\|\boldsymbol{v}^{i+1} - \hat{\boldsymbol{x}}^i\|^2}{n} \\
&\qquad + \sum_{i=1}^{t-1} (\mathsf{b}_{t+1,i+1} - \beta_{ti})^2 \frac{\|\hat{\boldsymbol{x}}^i\|^2}{n} + (\mathsf{b}_{t+1,1} - \bar{\mu}_t)^2 \frac{\|\boldsymbol{x}^*\|^2}{n} \Bigg] \\
&:= 2t(T_1 + T_2 + T_3 + T_4).
\end{aligned} \tag{100}$$

Since $\|\boldsymbol{A}\|_{\text{op}} \leq C$, the induction hypothesis (98) implies that $T_1 \to 0$. By the induction hypothesis, we also have $\|\boldsymbol{v}^{i+1} - \hat{\boldsymbol{x}}^i\|^2/d \to 0$ and $\|\hat{\boldsymbol{x}}^i\|^2/d \to \mathbb{E}\{\hat{X}^2\}$ for $i \leq (t-1)$. Furthermore, $\|\boldsymbol{x}^*\|^2/d \to \mathbb{E}\{X_*^2\}$. Hence, we can prove that T_2, T_3, T_4 each tend to zero by showing that:

$$\lim_{n\to\infty} \mathsf{b}_{t+1,1} = \bar{\mu}_t, \tag{101}$$

$$\lim_{n\to\infty} \mathsf{b}_{t+1,i+1} = \lim_{n\to\infty} \beta_{t,i} = \bar{\beta}_{t,i}, \qquad i \in [t-1], \tag{102}$$

where the limiting values $(\bar{\beta}_{t,i})$ in (102) will be defined below (see (108)).

Recall from (60)-(61) that the coefficients $(\mathsf{b}_{t+1,j})_{j\leq t}$ are determined by the entries of the matrices $\hat{\boldsymbol{\Psi}}_t$ and $\hat{\boldsymbol{\Phi}}_{t+1}$, defined in (58). (Though the definition of $\boldsymbol{M}_{t+1}^\mathsf{b}$ in (60) involves $\hat{\boldsymbol{\Psi}}_{t+1}$, it can be verified that its last row does not affect the computation, so the formula depends only on $\hat{\boldsymbol{\Psi}}_t, \hat{\boldsymbol{\Phi}}_{t+1}$.) From (59), the non-zero entries of these matrices are of the form

$$\begin{aligned}
&\langle \partial_k f_\ell(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*, \ldots, \boldsymbol{z}^{\ell+1} + \bar{\mu}_\ell \boldsymbol{x}^*) \rangle, \quad 1 \leq k \leq \ell \leq (t-1), \\
&\langle \partial_g h_\ell(\boldsymbol{m}^2, \ldots, \boldsymbol{m}^\ell, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})) \rangle, \quad \langle \partial_k h_\ell(\boldsymbol{m}^2, \ldots, \boldsymbol{m}^\ell, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})) \rangle, \quad 1 \leq k < \ell \leq t,
\end{aligned}$$

where we recall that ∂_k denotes the partial derivative with respect to the k-th argument. By Proposition D.1 and Lemma D.2, we have that for $\ell \geq 1$:

$$\begin{aligned}
&(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*, \ldots, \boldsymbol{z}^{\ell+1} + \bar{\mu}_\ell \boldsymbol{x}^*) \xrightarrow{W_2} (Z_2 + \bar{\mu}_1 X_*, \ldots, Z_{\ell+1} + \bar{\mu}_\ell X_*) \stackrel{\text{d}}{=} (X_1, \ldots, X_\ell), \\
&(\boldsymbol{m}^1, \ldots, \boldsymbol{m}^\ell, \boldsymbol{\varepsilon}) \xrightarrow{W_2} (M_1, \ldots, M_\ell, \varepsilon) \stackrel{\text{d}}{=} (G, R_1, \ldots, R_{\ell-1}, \varepsilon).
\end{aligned} \tag{103}$$

Since the functions f_ℓ and h_ℓ are Lipschitz in each argument, (103) together with Lemma E.1 implies that

$$\begin{aligned}
&\lim_{n\to\infty} \langle \partial_k f_\ell(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*, \ldots, \boldsymbol{z}^{\ell+1} + \bar{\mu}_\ell \boldsymbol{x}^*) \rangle = \mathbb{E}\{\partial_k f_\ell(X_1, \ldots, X_\ell)\}, \quad 1 \leq k \leq \ell \leq (t-1), \\
&\lim_{n\to\infty} \langle \partial_g h_\ell(\boldsymbol{m}^2, \ldots, \boldsymbol{m}^\ell, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})) \rangle = \mathbb{E}\{\partial_g h_\ell(R_1, \ldots, R_{\ell-1}), q(G, \varepsilon)\}, \\
&\lim_{n\to\infty} \langle \partial_k h_\ell(\boldsymbol{m}^2, \ldots, \boldsymbol{m}^\ell, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})) \rangle = \mathbb{E}\{\partial_k h_\ell(R_1, \ldots, R_{\ell-1}, q(G, \varepsilon))\}, \quad 1 \leq k < \ell \leq t.
\end{aligned} \tag{104}$$

Therefore, $\hat{\boldsymbol{\Psi}}_t \to \widetilde{\boldsymbol{\Psi}}_t$ and $\hat{\boldsymbol{\Phi}}_{t+1} \to \widetilde{\boldsymbol{\Phi}}_{t+1}$, where $\widetilde{\boldsymbol{\Psi}}_t, \widetilde{\boldsymbol{\Phi}}_{t+1}$ are defined in (65). Consequently, the matrix $\boldsymbol{M}^{\mathsf{b}}_{t+1}$ and the coefficients $(\mathsf{b}_{t+1,j})$, defined via (60)-(61), converge to the following limits:

$$\lim_{n\to\infty} \boldsymbol{M}^{\mathsf{b}}_{t+1} = \widetilde{\boldsymbol{M}}^{\mathsf{b}}_{t+1} \equiv \delta \sum_{j=0}^{t} \bar{\kappa}_{2(j+1)} \widetilde{\boldsymbol{\Phi}}_{t+1} (\widetilde{\boldsymbol{\Psi}}_{t+1} \widetilde{\boldsymbol{\Phi}}_{t+1})^j, \qquad \lim_{n\to\infty} \mathsf{b}_{t+1,j} = \bar{\mathsf{b}}_{t+1,j}, \quad j \in [t], \tag{105}$$

where $(\bar{\mathsf{b}}_{t+1,j})$ are computed according to (61) from the last row of $\widetilde{\boldsymbol{M}}^{\mathsf{b}}_{t+1}$. By the induction hypothesis (98), we have $\|\hat{\boldsymbol{x}}^i - \boldsymbol{v}^{i+1}\|^2/d \to 0$ for $i \le (t-1)$. Therefore the term T_2 in (100) tends to 0 as $n \to \infty$.

Using (10), the coefficients (β_{ti}) are determined by the entries of the matrices $\boldsymbol{\Psi}_{t+1}$ and $\boldsymbol{\Phi}_{t+1}$, defined in (6). The non-zero entries of these matrices are of the form

$$\begin{aligned}
&\langle \partial_k f_\ell(\boldsymbol{x}^1, \dots, \boldsymbol{x}^\ell) \rangle, \quad 1 \le k \le \ell \le t, \\
&\langle \partial_g h_\ell(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{\ell-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) \rangle, \quad \langle \partial_k h_\ell(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{\ell-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) \rangle, \quad 1 \le k < \ell \le t.
\end{aligned}$$

By the induction hypothesis (74)-(75) for $(t-1)$ and (103), we have

$$\begin{aligned}
&(\boldsymbol{x}^1, \dots, \boldsymbol{x}^{t-1}) \xrightarrow{W_2} (Z_2 + \bar{\mu}_1 X_*, \dots, Z_t + \bar{\mu}_{t-1} X_*) \overset{\mathrm{d}}{=} (X_1, \dots, X_{t-1}) \\
&(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{t-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) \xrightarrow{W_2} (M_2, \dots, M_t, q(M_1, \varepsilon)) \overset{\mathrm{d}}{=} (R_1, \dots, R_{t-1}, q(G, \varepsilon)).
\end{aligned} \tag{106}$$

Since f_ℓ and h_ℓ are Lipschitz in each argument, (106) together with Lemma E.1 implies

$$\begin{aligned}
&\lim_{n\to\infty} \langle \partial_k f_\ell(\boldsymbol{x}^1, \dots, \boldsymbol{x}^\ell) \rangle = \mathbb{E}\{\partial_k f_\ell(X_1, \dots, X_\ell)\}, \quad 1 \le k \le \ell \le (t-1), \\
&\lim_{n\to\infty} \langle \partial_g h_\ell(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{\ell-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) \rangle = \mathbb{E}\{\partial_g h_\ell(R_1, \dots, R_{\ell-1}), q(G, \varepsilon))\}, \\
&\lim_{n\to\infty} \langle \partial_k h_\ell(\boldsymbol{r}^1, \dots, \boldsymbol{r}^{\ell-1}, q(\boldsymbol{g}, \boldsymbol{\varepsilon})) \rangle = \mathbb{E}\{\partial_k h_\ell(R_1, \dots, R_{\ell-1}, q(G, \varepsilon))\}, \quad 1 \le k < \ell \le t.
\end{aligned} \tag{107}$$

Therefore, $\boldsymbol{\Psi}_t \to \bar{\boldsymbol{\Psi}}_t$ and $\boldsymbol{\Phi}_{t+1} \to \bar{\boldsymbol{\Phi}}_{t+1}$, where the entries of $\bar{\boldsymbol{\Psi}}_t, \bar{\boldsymbol{\Phi}}_{t+1}$ are defined as in (16). We note that computing $\boldsymbol{M}^{\beta}_{t+1}$ defined in (8) requires knowledge of only $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Phi}_{t+1}$ since the last row of $\boldsymbol{\Psi}_{t+1}$ is zeroed out in the multiplication with $\boldsymbol{\Phi}_{t+1}$. Therefore,

$$\lim_{n\to\infty} \boldsymbol{M}^{\beta}_{t+1} = \bar{\boldsymbol{M}}^{\beta}_{t+1} \equiv \delta \sum_{j=0}^{t} \bar{\kappa}_{2(j+1)} \bar{\boldsymbol{\Phi}}_{t+1} (\bar{\boldsymbol{\Psi}}_{t+1} \bar{\boldsymbol{\Phi}}_{t+1})^j, \qquad \lim_{n\to\infty} \beta_{t,i} = \bar{\beta}_{t,i}, \quad i \in [t-1]. \tag{108}$$

where $(\bar{\beta}_{t,i})$ are computed according to (10) from the last row of $\bar{\boldsymbol{M}}^{\beta}_{t+1}$. Since the limits in (104) and (107) are the same, using the formulas for $\bar{\boldsymbol{\Psi}}_t, \bar{\boldsymbol{\Phi}}_{t+1}$ (from (16)) and for $\widetilde{\boldsymbol{\Psi}}_t, \widetilde{\boldsymbol{\Phi}}_{t+1}$ (from (65)), we have

$$\bar{\boldsymbol{\Psi}}_t = \widetilde{\boldsymbol{\Psi}}_t, \qquad \bar{\boldsymbol{\Phi}}_{t+1} = \widetilde{\boldsymbol{\Phi}}_{t+1}, \qquad \bar{\boldsymbol{M}}^{\beta}_{t+1} = \widetilde{\boldsymbol{M}}^{\mathsf{b}}_{t+1}. \tag{109}$$

Combining (105), (108), (109) and recalling that $\bar{\mu}_t = (\boldsymbol{M}^{\beta}_{t+1})_{t+1,1}$, we obtain the claims in (101)-(102). We have therefore shown that each of the four terms in (100) tends to zero, and hence

$$\lim_{n\to\infty} \frac{\|\boldsymbol{x}^t - \boldsymbol{z}^{t+1} - \bar{\mu}_t \boldsymbol{x}^*\|^2}{d} = 0. \tag{110}$$

Moreover, since $\hat{\boldsymbol{x}}^t = f_t(\boldsymbol{x}^1, \dots, \boldsymbol{x}^t)$ and $\boldsymbol{v}^{t+1} = f_t(\boldsymbol{z}^2 + \bar{\mu}_1 \boldsymbol{x}^*, \dots, \boldsymbol{z}^{t+1} + \bar{\mu}_t \boldsymbol{x}^*)$ with f_t Lipschitz, we also have

$$\lim_{n\to\infty} \frac{\|\hat{\boldsymbol{x}}^t - \boldsymbol{v}^{t+1}\|^2}{d} = 0. \tag{111}$$

Using (110)-(111) together with the bounds in (85) then yields

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\|\boldsymbol{x}^t\|^2}{d} = \lim_{n\to\infty} \frac{\|\boldsymbol{z}^{t+1} + \bar{\mu}_t \boldsymbol{x}^*\|^2}{d} = \mathbb{E}\{(Z_{t+1} + \bar{\mu}_t X)^2\} = \mathbb{E}\{X_t^2\}, \\
&\lim_{n\to\infty} \frac{\|\hat{\boldsymbol{x}}^t\|^2}{d} = \lim_{n\to\infty} \frac{\|\boldsymbol{v}^{t+1}\|^2}{d} = \mathbb{E}\{V_{t+1}^2\} = \mathbb{E}\{\hat{X}_t^2\},
\end{aligned} \tag{112}$$

where the last equality in each line above is due to Lemma D.2. Using (110)-(112) and the induction hypothesis (98) in (77) yields the result (74).

The proof of (75) is along similar lines. From the definitions of $\boldsymbol{r}^t$ and $\boldsymbol{m}^{t+1}$ (see (5) and (31)), we have

$$\begin{aligned}\boldsymbol{r}^t - \boldsymbol{m}^{t+1} &= \boldsymbol{A}(\hat{\boldsymbol{x}}^t - \boldsymbol{v}^{t+1}) + \sum_{i=1}^{t}(\mathsf{a}_{t+1,i+1}\boldsymbol{u}^{i+1} - \alpha_{t,i}\boldsymbol{s}^i) \\ &= \boldsymbol{A}(\hat{\boldsymbol{x}}^t - \boldsymbol{v}^{t+1}) + \sum_{i=1}^{t}\mathsf{a}_{t+1,i+1}(\boldsymbol{u}^{i+1} - \boldsymbol{s}^i) + \sum_{i=1}^{t}(\mathsf{a}_{t+1,i+1} - \alpha_{t,i})\boldsymbol{s}^i.\end{aligned} \tag{113}$$

Using the Cauchy-Schwarz inequality, we obtain

$$\frac{\|\boldsymbol{r}^t - \boldsymbol{m}^{t+1}\|^2}{n} \le 2(t+2)\left[\|\boldsymbol{A}\|_{\text{op}}^2\frac{\|\hat{\boldsymbol{x}}^t - \boldsymbol{v}^{t+1}\|^2}{n} + \sum_{i=1}^{t}\mathsf{a}_{t+1,i+1}^2\frac{\|\boldsymbol{u}^{i+1} - \boldsymbol{s}^i\|^2}{n} + \sum_{i=1}^{t}(\mathsf{a}_{t+1,i+1} - \alpha_{t,i})^2\frac{\|\boldsymbol{s}^i\|^2}{n}\right]. \tag{114}$$

By assumption, $\|\boldsymbol{A}\|_{\text{op}}^2 \le C$; therefore, by (111) the first term in the brackets tends to zero. By the induction hypothesis (98), we have

$$\lim_{n\to\infty}\frac{\|\boldsymbol{u}^{i+1} - \boldsymbol{s}^i\|^2}{n} = 0, \quad \lim_{n\to\infty}\frac{\|\boldsymbol{s}^i\|^2}{n} = \mathbb{E}\{S_i^2\}, \quad 1 \le i \le t. \tag{115}$$

We can also show that

$$\lim_{n\to\infty}\mathsf{a}_{t+1,i+1} = \lim_{n\to\infty}\alpha_{t,i} = \bar{\alpha}_{t,i}, \qquad 1 \le i \le t, \tag{116}$$

where $(\bar{\alpha}_{t,1}, \dots, \bar{\alpha}_{t,t})$ is defined as in (9) using the matrix $\bar{\boldsymbol{M}}_{t+1}^{\alpha} \equiv \sum_{j=0}^{t+1}\bar{\kappa}_{2(j+1)}\bar{\boldsymbol{\Psi}}_{t+1}(\bar{\boldsymbol{\Phi}}_{t+1}\bar{\boldsymbol{\Psi}}_{t+1})^j$. The proof of (116) is omitted as it is similar to that of (102): we show that $\boldsymbol{M}_{t+1}^{\alpha} \to \bar{\boldsymbol{M}}_{t+1}^{\alpha}$ and $\boldsymbol{M}_{t+1}^{\mathsf{a}} \to \widetilde{\boldsymbol{M}}_{t+1}^{\mathsf{a}} \equiv \sum_{j=0}^{t+1}\bar{\kappa}_{2(j+1)}\widetilde{\boldsymbol{\Psi}}_{t+1}(\widetilde{\boldsymbol{\Phi}}_{t+1}\widetilde{\boldsymbol{\Psi}}_{t+1})^j$, and then that $\widetilde{\boldsymbol{M}}_{t+1}^{\mathsf{a}} = \bar{\boldsymbol{M}}_{t+1}^{\alpha}$. Using (115) and (116) in (114), we obtain

$$\lim_{n\to\infty}\frac{\|\boldsymbol{r}^t - \boldsymbol{m}^{t+1}\|^2}{n} = 0, \qquad \lim_{n\to\infty}\frac{\|\boldsymbol{s}^{t+1} - \boldsymbol{u}^{t+2}\|}{n} = 0, \tag{117}$$

where the second limit holds because

$$\boldsymbol{s}^{t+1} = h_{t+1}(\boldsymbol{r}^1, \dots, \boldsymbol{r}^t, q(\boldsymbol{g}, \boldsymbol{\varepsilon})), \qquad \boldsymbol{u}^{t+2} = h_{t+1}(\boldsymbol{m}^2, \dots, \boldsymbol{m}^{t+1}, q(\boldsymbol{m}^1, \boldsymbol{\varepsilon})),$$

with h_{t+1} Lipschitz in each argument. Using (117) together with the bounds in (96) then yields

$$\begin{aligned}\lim_{n\to\infty}\frac{\|\boldsymbol{r}^t\|^2}{n} &= \lim_{n\to\infty}\frac{\|\boldsymbol{m}^{t+1}\|^2}{n} = \mathbb{E}\{M_{t+1}^2\} = \mathbb{E}\{R_t^2\}, \\ \lim_{n\to\infty}\frac{\|\boldsymbol{s}^{t+1}\|^2}{n} &= \lim_{n\to\infty}\frac{\|\boldsymbol{u}^{t+2}\|^2}{n} = \mathbb{E}\{U_{t+2}^2\} = \mathbb{E}\{S_{t+1}^2\},\end{aligned} \tag{118}$$

where the last equality in each line above is due to Lemma D.2.

Using (117), (118), and the induction hypothesis (98) in (78) yields the result (75), completing the proof.

E. An Auxiliary Lemma

Lemma E.1. *Let $F\colon \mathbb{R}^t \to \mathbb{R}$ be a Lipschitz function, and let $\partial_k F$ denote its derivative with respect to the k-th argument, for $1 \le k \le t$. Assume that $\partial_k F$ is continuous almost everywhere in the k-th argument, for each k. Let $(V_1^{(m)}, \dots, V_t^{(m)})$ be a sequence of random vectors in $\mathbb{R}^t$ converging in distribution to the random vector $(V_1, \dots, V_t)$ as $m \to \infty$. Furthermore, assume that the distribution of $(V_1, \dots, V_t)$ is absolutely continuous with respect to the Lebesgue measure. Then,*

$$\lim_{m\to\infty}\mathbb{E}\{\partial_k F(V_1^{(m)}, \dots, V_t^{(m)})\} = \mathbb{E}\{\partial_k F(V_1, \dots, V_t)\}, \qquad k \in [t].$$

The result was proved for $t = 2$ in Lemma 6 of (Bayati & Montanari, 2011). The proof for $t > 2$ is essentially the same; see also Lemma 7.14 in (Feng et al., 2021).