Linear Operator Approximate Message Passing: Power Method with Partial and Stochastic Updates

Riccardo Rossetti*, Bobak Nazer†, and Galen Reeves*‡
Departments of Statistical Science* and Electrical and Computer Engineering‡, Duke University
Department of Electrical and Computer Engineering†, Boston University

***Abstract*—This paper introduces a framework for approximate message passing (AMP) in dynamic settings where the data at each iteration is passed through a linear operator. This framework is motivated in part by applications in large-scale, distributed computing where only a subset of the data is available at each iteration. An autoregressive memory term is used to mitigate information loss across iterations and a specialized algorithm, called projection AMP, is designed for the case where each linear operator is an orthogonal projection. Precise theoretical guarantees are provided for a class of Gaussian matrices and non-separable denoising functions. Specifically, it is shown that the iterates can be well-approximated in the high-dimensional limit by a Gaussian process whose second-order statistics are defined recursively via state evolution. These results are applied to the problem of estimating a rank-one spike corrupted by additive Gaussian noise using partial row updates, and the theory is validated by numerical simulations.**

I. Introduction

Approximate message passing (AMP) refers to a family of iterative algorithms that has been applied to high-dimensional inference problems including regression, matrix estimation, and channel coding; for a comprehensive reference, see the recent tutorial [1]. The basic form of an AMP algorithm can be summarized as a recursion on n-dimensional iterates,

$$x_t = Mf_t(x_{<t}) - \sum_{s<t} b_{ts} f_s(x_{<s}), \quad t = 0, 1, 2, \dots \tag{1}$$

where M is an $n \times n$ "data matrix", the $f_t : \mathbb{R}^{n\times t} \to \mathbb{R}^n$ are deterministic functions (with initialization $f_0 = f_0(\emptyset) \in \mathbb{R}^n$), the $b_{ts} \in \mathbb{R}$ are scalar "debiasing" coefficients , and the notation $x_{<t} = (x_0, \dots, x_{t-1})$ represents the collection of iterates up to time $t-1$. Depending on the application, the matrix M is obtained from the observed data via elementary preprocessing steps such as centering and symmetrization.

One of the key features of the AMP framework is that the behavior can be tracked precisely in high-dimension settings provided that the data matrix satisfies certain distributional assumptions. In these settings, the coefficients b_{ts} can be specified in a way that both accelerates the overall convergence and guarantees that the process $\{x_t\}$ can be closely approximated by a Gaussian process $\{y_t\}$ whose mean and covariance can be efficiently computed via a "state evolution" (SE) recursion.

An important consideration for very large matrix operations is that computations are distributed across multiple servers. In practice, the server response times can have a long tail [2], and waiting for the "stragglers" can significantly delay the next iteration. This delay can be mitigated via replication or coded computation [3]–[7] to ensure that the full matrix multiplication is available once enough servers respond, at the cost of additional computation per iteration to maintain a suitable erasure-correcting code. However, for high-dimensional inference tasks, it may be more efficient to proceed to the next iterate. One of the main contributions of this paper is an AMP framework that naturally captures this scenario, enabling a precise performance characterization via state evolution.

This research was supported in part by NSF Grant 1750362.

A. Overview of Main Results

This paper extends the scope of AMP to settings where the matrix may change with each iteration. Specifically, we introduce the Linear Operator AMP (OpAMP) framework in which a linear operator $\mathcal{L}_t \colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is applied to the data matrix M at each iteration, and the recursion has the form

$$x_t = \mathcal{L}_t(M) f_t(x_{<t}) - \sum_{s<t} B_{ts} f_s(x_{<s}). \tag{2}$$

Under a Gaussian assumption on the data matrix, we show how the matrices $B_{ts} \in \mathbb{R}^{n\times n}$ can be specified as a function of the $\mathcal{L}_t$ to enforce approximate Gaussianity of the iterates.

To provide a tractable model for long-term memory we also introduce an autoregressive version of (2) that has a linear dependence on the previous iterates:

$$x_t = \mathcal{L}_t(M) f_t(x_{t-1}) + \sum_{s<t} A_{ts} x_{s-1} - \sum_{s<t} B_{ts} f_s(x_{s-1}) \tag{3}$$

where the matrices A_{ts} can be designed such that x_t provides a suitable summary of the previous iterations.

Specializing to the case where each linear operator is given by $\mathcal{L}_t(M) = \Pi_t M$ for an $n \times n$ projection matrix Π_t, we define the projection AMP recursion

$$x_t = \Pi_t \Big(M f_t(x_{t-1}) - \sum_{s<t} b_{ts} f_s(x_{s-1}) \Big) + \Pi_t^{\perp} x_{t-1} \tag{4}$$

where the dependence on prior iterations is specified by the projector sequence. For the special case of commuting orthogonal projection matrices, the SE has a particularly simple recursive structure that is analysed in Section III-A.

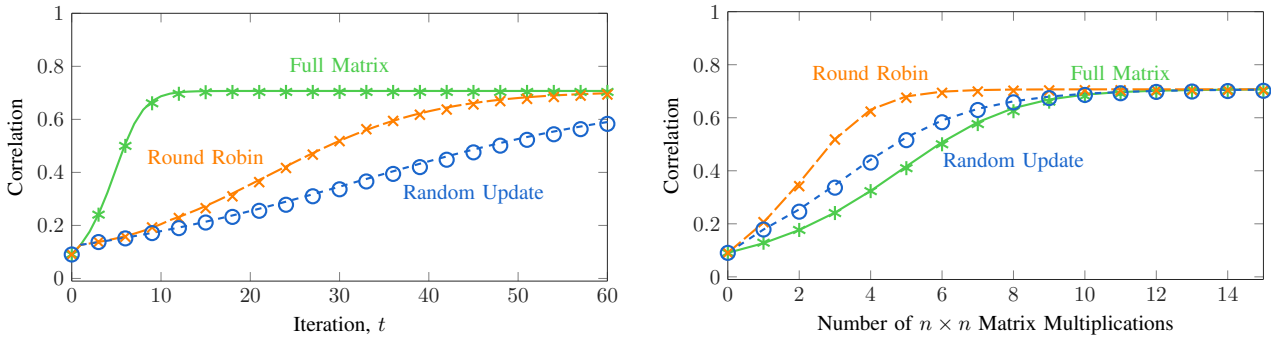

Fig. 1. Correlation $\langle \theta, \frac{x_t}{\|x_t\|} \rangle$ versus iteration count (on the left) and versus total number of $n \times n$ multiplications used (on the right). Markers denote empirical performance (averaged over 100 trials) and curves denote theoretical SE predictions.

B. Case Study: Power Iteration with Partial Updates

Consider computing the dominant eigenvalue $v_1 \in \mathbb{R}^n$ of a symmetric data matrix $M = \sum_{i=1}^n \xi_i v_i v_i^\top$ with eigenvalues $|\xi_1| \geq |\xi_2| \geq \cdots \geq |\xi_n| \in \mathbb{R}$ via power iteration:

$$x_{t+1} = M \frac{x_t}{\|x_t\|} \tag{5}$$

for some initialization $x_0 \in \mathbb{R}^n$. Assuming that the initial value x_0 has some angle $\alpha > 0$ with the leading eigenvalue v_1, classical convergence bounds [8, Theorem 8.2.1] guarantee that the rescaled iterations $x_t/\|x_t\|$ converge to v_1 geometrically fast as long as the spectral gap is greater than one. This result is general and holds for any symmetric matrix M. To present our AMP-based method, we focus on matrices M that were drawn from a rank-one spiked matrix model, $M = \lambda \theta \theta^\top + Z$ where $\lambda > 0$, $\theta \in \mathbb{R}^n$ is, in this example, presumed to be a unit vector, and Z is an $n \times n$ random matrix drawn from the Gaussian orthogonal ensemble $\mathsf{GOE}(n)$, i.e., Z is symmetric with independent $\mathsf{N}(0, 1/n)$ entries above the diagonal and independent $\mathsf{N}(0, 2/n)$ entries on the diagonal. The goal is to use the leading eigenvalue v_1 to provide an estimate of the direction of the ground-truth signal θ.

Interpreting the projection onto the unit sphere $x \mapsto x/\|x\|$ as a denoising function, a standard AMP correction term can be added to the power method to obtain the recursion

$$x_{t+1} = \frac{1}{\|x_t\|} \left(M x_t - \frac{x_{t-1}}{\|x_{t-1}\|} \right) \tag{6}$$

with the convention that $x_{-1}/\|x_{-1}\| \equiv 0$. When M is sampled from a spiked matrix model, this iteration can be tracked accurately in the large-n limit via SE. Furthermore, when $\lambda > 1$, the iterates $\{x_t\}$ can be shown to converge to the fixed point λv_1, a multiple of the leading eigenvector.

Now, consider the setting where only a fraction of the rows of M are available at each iteration t, i.e. only $\Pi_t M$ is observed for some diagonal 0–1 matrices $\{\Pi_t\}$. Based on the results in this paper, we propose the following AMP-corrected power method with partial updates:

$$x_{t+1} = \frac{1}{\|x_t\|} \Pi_t \Big(M x_t - \frac{1}{n} \sum_{s<t} \mathrm{tr}\big(\Pi_{t-1}^\perp \cdots \Pi_{s+1}^\perp \Pi_s\big) \frac{x_s}{\|x_s\|} \Big) + \Pi_t^\perp x_t \tag{7}$$

where the complementary projections are simply $\Pi_t^\perp = \mathrm{I} - \Pi_t$. This recursion is practical as it does not require knowledge of any of the model parameters and its computational cost per iteration t is dominated by the matrix multiply $\Pi_t M x_t$, which involves $O(nk_t)$ operations, where $k_t = \mathrm{rank}(\Pi_t)$. In Section III-B, we provide single-letter formulas for the SE dynamics of (7) under some structural assumptions on the signal θ and update schedule $\{\Pi_t\}$.

In Figure 1, the projection AMP algorithm in (7) is applied to a size $n = 5,000$ data matrix with $\lambda = \sqrt{2}$, and the results are averaged over 100 Monte Carlo trials, each with the same ground truth $\theta \in \{\pm \frac{1}{\sqrt{n}}\}^n$ and initialization $x_0 \in \mathbb{R}^n$ with $\langle \theta, \frac{x_0}{\|x_0\|} \rangle = 0.01$. A comparison of the empirical results and the theoretical prediction obtained from the asymptotic SE is provided for the following protocols:

- *Full matrix:* The full data matrix is applied each iteration.
- *Round robin:* The rows are partitioned into 10 equally-sized subsets. Each iteration applies the rows in a subset.
- *Random update:* Each row is updated independently with probability $1/10$.

Intuitively, the round-robin and random-update protocols converge more slowly than full-matrix AMP with respect to the raw iteration count. However, if we instead plot the performance with respect to effective number of $n \times n$ matrix multiplications, we find that the round-robin protocol is more efficient than the full-matrix protocol. (For certain parameter settings, the random-update protocols is also more efficient, but this is not always the case.)

C. Related Work

Most of the work in the AMP literature has focused on finite-memory versions of (1) where each f_t depends only on a fixed number of previous iterates [9]–[17]. The full-memory formulation in (1) has appeared in recent work as a model for generalized first-order methods [18], [19].

Our analysis builds on the theoretical framework for non-separable functions and IID Gaussian matrices introduced by Berthier et al. [20], and further developed in [21], [22] where (pseudo)-Lipschitz continuity is the only assumption placed on the function sequence and convergence is assessed in terms a sequence of suitably normalized pseudo-Lispchitz test functions. Our main theorems establish conditions under which the normalized difference $\frac{1}{\sqrt{n}}\|x_{\leq T} - y_{\leq T}\|$ over a fixed number of iterations T converges to zero in probability in the large-n limit. Concurrent work explores this form of convergence from a non-asymptotic perspective [23].

A version of the projection AMP framework of this paper was studied by Çakmak et al. [24] using non-rigorous dynamical functional theory. Projection AMP is also related to recent work that uses a full memory AMP recursion to approximate the discrete-time dynamics of gradient descent and other first-order optimizations techniques [25], [26]. These works focus on the behavior of existing optimization techniques in scaling regimes where the number of rows updated at each iteration grows sublinearly with the problem dimension. By contrast, the main focus of this paper is to design algorithms that overcome the limitations of the dynamic data as expressed in (2), e.g., by optimizing the long-term memory as function of the linear operators.

In a slightly different direction, the special case of (2) where the linear operator is fixed for all iterations has been studied by a subset of the authors in the context of the matrix tensor product model [27], [28].

Beyond the setting of IID Gaussian matrices, AMP algorithms have also been proposed and analyzed for orthogonally-invariant random matrix ensembles [29]–[36] and semirandom ensembles [37]. These results impose a separability assumption on the functions $\{f_t\}$, which precludes the general linear transformations used in the proof of our OpAMP framework. Extending the results in this paper to other matrix ensembles is an interesting direction for future research.

Recent efforts have developed distributed, accelerated, and robust variations on the power method as well as long-run convergence guarantees under suitable conditions [38]–[43]. Additionally, recent work on subspace tracking algorithms with missing data [44] has characterized the high-dimensional performance limit via differential equations [45].

II. Linear Operator AMP

This section describes the linear operator AMP framework and states our main theoretical results. Owing to space limitations, the proofs can be found in the full version [46].

To streamline the presentation, we focus on a centered version of the recursion given by

$$x_t = \mathcal{L}_t(Z) f_t(x_{<t}) - \sum_{s<t} B_{ts} f_s(x_{<s}) \tag{8}$$

where Z is a $\mathsf{GOE}(n)$ matrix. The extension to settings where the matrix has a low-rank signal component follows from standard arguments in the AMP literature; see Section III.

Each linear operator $\mathcal{L}_t$ has a (non-unique) decomposition of the form

$$\mathcal{L}_t(Z) = \sum_{k=1}^{K} L_{tk} Z R_{tk} \tag{9}$$

for $n \times n$ matrices $\{L_{tk}, R_{tk} : k = 1, \dots, K\}$. We require that the both operator norm and the rank (i.e., the smallest K such that (9) holds) are bounded uniformly w.r.t. the problem dimension n.

Our theoretical results provide a connection between the distribution of the AMP iterates and a zero-mean Gaussian process $\{y_t\}$ whose second-order statistics are described in terms of $\{\mathcal{L}_t\}$ and $\{f_t\}$ via a recursive process called *state evolution*. Starting with $\mathsf{Cov}(y_0) = \frac{1}{n}\|f_0\|^2 \mathrm{I}_n$, the covariance at time t is defined by the covariance up to $t-1$ according to

$$\mathsf{Cov}(y_s, y_t) = \sum_{l,k=1}^{K} q_{sltk} L_{sl} L_{tk}^\top, \tag{10}$$

$$q_{sltk} \coloneqq \frac{1}{n}\mathbb{E}[\langle R_{sl} f_s(y_{<s}), R_{tk} f_t(y_{<t})\rangle] \tag{11}$$

This construction holds holds for any collection of matrices $\{L_{tk}, R_{tk}\}$ satisfying the decomposition in (9).

Assumption 1. Each $f_t \colon \mathbb{R}^{n\times t} \to \mathbb{R}^n$ is L-Lipschitz continuous and satisfies $\frac{1}{\sqrt{n}}\|f_t(0)\| \leq C$ where C, L are positive constants that do not depend on n.

Assumption 2. Each $\mathcal{L}_t$ can be decomposed in the form given in (9) with $\|R_{tk}\|_{\mathrm{op}}, \|L_{tk}\|_{\mathrm{op}} \leq C'$ for all $t, k \in \mathbb{N}_0$, where C', K are are positive constants that do not depend on n.

Theorem 1. Let $\{x_t\}$ be generated by (8) and let $\{y_t\}$ be the zero-mean Gaussian process defined by the SE (10). Suppose Assumptions 1 and 2 hold, $Z \sim \mathsf{GOE}(n)$, and the matrices $\{B_{ts} : 0 \leq s < t\}$ are given by

$$B_{ts} = \sum_{k,l=1}^{K} \frac{1}{n} \operatorname{tr}(R_{tk}\mathbb{E}[\mathsf{D}_s f_t(y_{<t})] L_{sl})\, L_{tk} R_{sl} \tag{12}$$

Here, the notation D_s indicates the Jacobian matrix of $f_t(x_{<t})$ computed w.r.t. the input vector x_s. Then, for any fixed number of iterations T, there exists a sequence (in n) of couplings between $x_{\leq T}$ and $y_{\leq T}$ such that $\frac{\|x_{\leq T} - y_{\leq T}\|}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathrm{p}} 0$.

A. Autoregressive Linear Operator AMP

While the general formulation in (8) allows for arbitrary dependence on prior iterations, the question remains of how the f_t should be optimized as a function of the linear operators. Motivated by practical considerations, we introduce an autoregressive version of (8), that uses a weighted linear combination of previous updates:

$$x_t = \mathcal{L}_t(Z) f_t(x_{t-1}) + \sum_{s<t} A_{ts} x_s - \sum_{s<t} B_{ts} f_s(x_{t-1}) \tag{13}$$

Here, the $n\times n$ matrices A_{ts} describe the long-term dependence and the functions f_t are applied only to the prior iteration, reducing the complexity of both the implementation and the analysis.

To describe the SE, we define the collection of $n \times n$ matrices $\{C_{st} : 0 \le s < t\}$ according to

$$\begin{bmatrix} \mathrm{I}_n & & & \\ -A_{10} & \mathrm{I}_n & & \\ \vdots & & \ddots & \\ -A_{t0} & \cdots & -A_{t,t-1} & \mathrm{I}_n \end{bmatrix}^{-1} = \begin{bmatrix} \mathrm{I}_n & & & \\ C_{10} & \mathrm{I}_n & & \\ \vdots & & \ddots & \\ C_{t0} & \cdots & C_{t,t-1} & \mathrm{I}_n \end{bmatrix}$$

Since these matrices are block unitriangular, their inverses exist. Furthermore, using the convention that $C_{tt} = \mathrm{I}_n$, the mapping from $\{A_{ts}\}$ to $\{C_{ts}\}$ can be expressed recursively via

$$C_{ts} = \sum_{r=0}^{t-1} A_{tr} C_{rs} \tag{14}$$

The distribution of the iterates from (13) is compared with a zero-mean Gaussian process $\{y_t\}$ whose covariance is defined recursively according to

$$\mathsf{Cov}(y_s, y_t) = \sum_{s' \le s, t' \le t} \sum_{l,k=1}^{K} q_{s'lt'k} C_{ss'} L_{s'l} (L_{t'k} C_{tt'})^\top \tag{15}$$

where q_{sltk} is defined in (11).

Assumption 3. $\|A_{ts}\|_{op} \le C''$ for all s, t where C'' is a positive constant that does not depend on n.

Theorem 2. Let $\{x_t\}$ be generated by (13) and let $\{y_n\}$ be the zero-mean Gaussian process defined by the SE. Suppose Assumptions 1, 2, and 3 hold, $Z \sim \mathsf{GOE}(n)$, and the matrices $\{B_{ts} : 0 \le s < t\}$ are given by

$$B_{ts} = \sum_{k,l=1}^{K} \frac{1}{n} \operatorname{tr}\left(R_{tk} \mathbb{E}[\mathsf{D} f_t(y_{t-1})] C_{t-1,s} L_{sl}\right) L_{tk} R_{sl} \tag{16}$$

where $\{L_{tk}, R_{tk}\}$ provide the decomposition of $\mathcal{L}_t$ given in (9) and $\{C_{ts}\}$ are defined by (14). Then, for any fixed number of iterations T, there exists a sequence (in n) of couplings between $x_{\le T}$ and $y_{\le T}$ such that $\frac{\|x_{\le T} - y_{\le T}\|}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathrm{p}} 0$.

Note that in view of (14), the matrices $C_{t-1,s}$ appearing in the correction term at time t can be computed in recursively in terms of the matrices $C_{t-2,s}$ used in the previous iteration.

B. Projection AMP

The projection AMP framework is a specialization of (13) where each linear operator is given by $\mathcal{L}_t(Z) = \Pi_t Z$ for an $n \times n$ projection matrix Π_t and the autoregressive linear memory term is the complementary projection matrix $\Pi_t^\perp = \mathrm{I} - \Pi_t$ applied to the past iteration.

The projection AMP recursion can be expressed as

$$x_t = \Pi_t\Big(Z f_t(x_{t-1}) - \sum_{s<t} b_{ts} f_s(x_{s-1})\Big) + \Pi_t^\perp x_{t-1} \tag{17}$$

A useful property of this formulation is that the memory terms depend only on the projector sequence and the debiasing terms are described by scalars.

The additional structure of projection matrices also leads to simplifications for the SE. The covariance of the zero-mean Gaussian process $\{y_t\}$ is given by

$$\begin{aligned} &\mathsf{Cov}(y_s, y_t) \\ &= \sum_{s' \le s, t' \le t} \frac{1}{n} \mathbb{E}[\langle f_{s'}(y_{s'-1}), f_{t'}(y_{t'-1})\rangle] C_{ss'} \Pi_{s'} (\Pi_{t'} C_{tt'})^\top \end{aligned} \tag{18}$$

where the matrices C_{ts} are defined by

$$C_{ts} = \begin{cases} \mathrm{I}_n & s = t \\ \Pi_t^\perp \Pi_{t-1}^\perp \cdots \Pi_{s+2}^\perp \Pi_{s+1}^\perp & 0 \le s < t \end{cases} \tag{19}$$

Theorem 3. Let $\{x_t\}$ be generated by (17) and let $\{y_n\}$ be the zero-mean Gaussian process defined by the SE. Suppose Assumption 1 holds, $Z \sim \mathsf{GOE}(n)$, and the scalars $\{b_{ts} : 0 \le s < t\}$ are given by

$$b_{ts} = \frac{1}{n} \operatorname{tr}(\mathbb{E}[\mathsf{D} f_t(y_{t-1})] C_{t-1,s} \Pi_s) \tag{20}$$

Then, for any fixed number of iterations T, there exists a sequence (in n) of couplings between $x_{\le T}$ and $y_{\le T}$ such that $\frac{\|x_{\le T} - y_{\le T}\|}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathrm{p}} 0$.

The SE for projection AMP admits further simplifications for the special case of commuting orthogonal projections, i.e., each Π_s is symmetric and $\Pi_s \Pi_t = \Pi_t \Pi_s$ for all s, t. Starting with (18) and then using the fact that the Π_t and C_{ts} commute, one finds that $\mathsf{Cov}(y_t)$ satisfies the simple recursion

$$\mathsf{Cov}(y_t) = \frac{1}{n} \mathbb{E}\big[\|f_t(y_{t-1})\|^2\big] \Pi_t + \mathsf{Cov}(y_{t-1}) \Pi_t^\perp \tag{21}$$

where $\mathsf{Cov}(y_{t-1}) \Pi_t^\perp = \Pi_t^\perp \mathsf{Cov}(y_{t-1})$ is symmetric. In particular, if $\Pi_0 = \mathrm{I}_n$ and every f_t is supported on the sphere of radius $\sigma\sqrt{n}$ then it follows that $\mathsf{Cov}(y_t) = \sigma^2 \mathrm{I}_n$ for all $t \in \mathbb{N}_0$.

III. Matrix Estimation with Partial Updates

In this section, we show how our linear operator framework can be applied to settings where the entire data matrix cannot be applied at each iteration. For concreteness, we focus on the rank-one spiked matrix model

$$M = \frac{\lambda}{n} \theta \theta^\top + Z \tag{22}$$

where $\lambda > 0$ is a positive scalar, $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is the unknown signal, and $Z \sim \mathsf{GOE}(n)$ is additive noise. The goal is to recover θ from M subject to the constraints that only a subset of the rows of M can be accessed at each iteration.

The update constraints are modeled using projection AMP with projections of the form $\Pi_t = \operatorname{diag}(\delta_t)$ where $\delta_t \in \{0, 1\}^n$ is a binary vector indicating which rows can be updated in the t-th iteration. For a given sequence of "denoising" functions $\{f_t\}$, we construct a sequence of estimates $\{\hat{\theta}_t\}$ using the following version of projection AMP:

$$\hat{\theta}_t = f_t(x_{t-1}) \tag{23a}$$

$$x_t = \delta_t \circ \Big(M \hat{\theta}_t - \sum_{s<t} b_{ts} \hat{\theta}_s\Big) + (1 - \delta_t) \circ x_{t-1} \tag{23b}$$

Here, $\circ$ denotes the elementwise (Hadamard) product and 1 denotes the all ones vector. The scalar debiasing coefficients b_{ts} are defined as a function of the SE according to (28). To circumvent some cumbersome details that arise with a generic initialization, we will assume throughout this section that every row of the matrix is updated in the the first time step, i.e., $\delta_0 \equiv 1$ is the all ones vector.

State evolution. Combining Theorem 3 with recentering arguments, it can be shown that the iterates from the recursion (23) are well approximated by a Gaussian process $\{y_t\}$, whose mean and covariance are defined by a two-parameter SE:

$$q_t = \frac{1}{n}\mathbb{E}\big[\|f_t(y_{t-1})\|^2\big]; \quad r_t = \frac{1}{n}\mathbb{E}[\langle \theta, f_t(y_{t-1})\rangle] \tag{24}$$

where $q_0 = \|f_0\|^2/n$, $r_0 = \langle \theta, f_0\rangle/n$ are the overlaps arising from the initial estimate $\hat{\theta}_0 = f_0 \in \mathbb{R}^n$. Starting with $\mathbb{E}[y_0] = \lambda r_0 \theta$ and $\mathsf{Cov}(y_0) = q_0 \mathrm{I}_n$, the mean and covariance of $\{y_t\}$ are updated recursively according to

$$\mathbb{E}[y_t] = \lambda r_t \delta_t \circ \theta + (1-\delta_t) \circ \mathbb{E}[y_{t-1}]\ , \tag{25a}$$

$$\mathsf{Cov}(y_t) = q_t \operatorname{diag}(\delta_t) + \mathsf{Cov}(y_{t-1})(\mathrm{I}_n - \operatorname{diag}(\delta_t)) \tag{25b}$$

A useful property of the SE is that, for each time step t, the distribution of the i-th component of the Gaussian vector $y_t = (y_{1t}, \dots, y_{tn})$ depends only on the signal component θ_i and the last time step in whch i-th row of the matrix was updated. To see this, observe that the $\mathsf{Cov}(y_t)$ is diagonal for all iterations, and thus each y_t has independent components. For a given indicator sequence $\{\delta_t\}$, the index for the most recent update of the i-th row before time step t is encoded by the function $\tau : \mathbb{N} \times \{1, \dots, n\} \to \mathbb{N}_0$, defined by

$$\tau(t, i) \coloneqq \max\{s \in \{0, 1, \dots, t-1\} : \delta_{si} = 1\} \tag{26}$$

where δ_{si} is the i-th element of the binary vector δ_s.

From the recursive structure in (25), it follows that every component that was last updated at time step s is described by a scalar Gaussian noise model with parameters (q_s, r_s). Specifically, the variables $y_{t1}, \dots, y_{tn}$ are independent with

$$\tau(t+1, i) = s \quad \Longrightarrow \quad y_{ti} \sim \mathsf{N}(\lambda r_s \theta_i, q_s) \tag{27}$$

for all $t \in \mathbb{N}_0$. We note that the function τ provides an alternative representation the indicator sequence $\{\delta_t\}$, which can be recovered via the correspondence $\delta_{ti} = \mathbf{1}\{\tau(t+1, i) = t\}$.

Debiasing coefficients. Substituting $\operatorname{diag}(\delta_t)$ in the definition for C_{ts} (19), the debiasing terms $\{b_{ts}\}$ (20) can be compactly represented in terms of τ as

$$b_{ts} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\tau(t, i) = s\}\mathbb{E}[\mathsf{D}_{ii} f_t(y_{t-1})] \tag{28}$$

where D_{ii} denotes the partial derivative of the i-th output $f_{ti}(y_{t-1})$ with respect to the i-th input $y_{t-1,i}$.

A. Asymptotic State Evolution

In this section, we obtain a simplified characterization of the SE by focusing on the high-dimensional limit for sequence of problems, with increasing dimension n, where the signal $\theta \in \mathbb{R}^n$, initialization $\hat{\theta}_0 \in \mathbb{R}^n$, and indicator sequence $\{\delta_t\} \in \{0,1\}^{n \times \mathbb{N}_0}$ together satisfy a decoupling property.

Assumption 4. M is given by (22) and the Projection AMP recursion (23) satisfies the following conditions:

1) For each $t \in \mathbb{N}_0$, the joint empirical measure of $\{(\theta_i, \hat{\theta}_{0i}, \delta_{0i}, \dots, \delta_{ti}) : i \in [n]\}$ converges in quadratic Wasserstein distance to a limiting probability measure the form $\mu \otimes \nu_t$ where μ is a distribution on $\mathbb{R}^2$ whose marginals have unit second moments and ν_t is is the distribution of the first t entries of a binary string drawn from a probability measure ν on $\{0,1\}^{\mathbb{N}_0}$.
2) For each $t \in \mathbb{N}$, the denoiser $f_t \colon \mathbb{R}^n \to \mathbb{R}^n$ is separable and is given by $f_{ti}(x_{t-1}) = \eta_t(x_{t-1,i}; q_{\tau(t,i)}, \lambda r_{\tau(t,i)})$ where $\eta_t \colon \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous scalar denoiser that is fixed for all n and τ is the index function defined in (26).

According to our convention that every row is updated at the first time step ($t = 0$), the measure ν is supported on binary strings whose first entry is one. For each $t \in \mathbb{N}$, we define p_t to the probability mass function for the position of the last non-zero entry occurring before time t, i.e.,

$$p_t(s) \coloneqq \nu_t\big(\{\omega \in \{0,1\}^t : \omega_s = 1, \omega_{s+1} = \cdots = \omega_{t-1} = 0\}\big)$$

for all $s \in \{0, 1, \dots, t-1\}$. This probability mass function can also be defined directly via the limiting empirical measure

$$p_t(s) \coloneqq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\tau(t, i) = s\} \tag{29}$$

where Assumption 4.1 ensures that the limit exists.

B. Power Iteration

In this section, we provide asymptotic guarantees on the limiting absolute empirical correlation of the projection AMP estimates $\{\hat{\theta}_t\}$ when the denoisers f_t are chosen to be the projection onto the sphere of radius $\sqrt{n}$. Starting with $\rho_0 \coloneqq \int u\hat{u}\, d\mu(u, \hat{u})$, we define the recurrence relation

$$\rho_t = \frac{\lambda \sum_{s<t} \rho_s p_t(s)}{\sqrt{\sum_{s<t}(\lambda^2 \rho_s^2 + 1) p_t(s)}} \tag{30}$$

The following result relates the sequence $\{\rho_t\}$ and with the asymptotic correlation between the signal θ and the distributed power method estimates $\{\hat{\theta}_t\}$.

Theorem 4. Let M be a spiked matrix model (22), $(\theta, \hat{\theta}_0)$ satisfy Assumption 4 and $\|\hat{\theta}_0\| = \sqrt{n}$. Consider the estimate sequence $\{\hat{\theta}_t\}$ produced by (23) with $f_t(x) = \sqrt{n}x/\|x\|$. Then, for each $t \in \mathbb{N}$,

$$\Big|\frac{1}{n}\langle \theta, \hat{\theta}_t\rangle - \rho_t\Big| \xrightarrow[n\to\infty]{\mathrm{p}} 0 \tag{31}$$

for $\{\rho_t\}$ defined recursively as in (30).

Due to space constraints, the reader is referred to the extended version of this work [46] for a proof.

REFERENCES

[1] O. Y. Feng, R. Venkataramanan, C. Rush, and R. J. Samworth, "A unifying tutorial on approximate message passing," *Foundations and Trends® in Machine Learning*, vol. 15, no. 4, pp. 335–536, May 2022.

[2] J. Dean and L. A. Barroso, "The tail at scale," *Communications of the ACM*, vol. 56, no. 2, pp. 74–80, 2013.

[3] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2017.

[4] S. Dutta, V. Cadambe, and P. Grover, "Short-dot: Computing large linear transforms distributedly using coded short dot products," *Advances In Neural Information Processing Systems*, vol. 29, 2016.

[5] Q. Yu, M. Maddah-Ali, and S. Avestimehr, "Polynomial codes: an optimal design for high-dimensional coded matrix multiplication," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] S. Dutta, H. Jeong, Y. Yang, V. Cadambe, T. M. Low, and P. Grover, "Addressing unreliability in emerging devices and non-von neumann architectures using coded computing," *Proceedings of the IEEE*, vol. 108, no. 8, pp. 1219–1234, 2020.

[7] S. Li and S. Avestimehr, "Coded computing: Mitigating fundamental bottlenecks in large-scale distributed computing and machine learning," *Foundations and Trends in Communications and Information Theory*, vol. 17, no. 1, pp. 1–148, 2020.

[8] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[9] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

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

[11] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE International Symposium on Information Theory*, Jul. 2011, pp. 2168–2172.

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

[13] C. Rush, A. Greig, and R. Venkataramanan, "Capacity-achieving sparse superposition codes via approximate message passing decoding," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1476–1500, 2017.

[14] T. Lesieur, F. Krzakala, and L. Zdeborová, "Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 7, p. 073403, Jul. 2017.

[15] P. Pandit, M. Sahraee, S. Rangan, and A. K. Fletcher, "Asymptotics of MAP inference in deep networks," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 842–846.

[16] A. Montanari and R. Venkataramanan, "Estimation of low-rank matrices via approximate message passing," *The Annals of Statistics*, vol. 49, no. 1, pp. 321–345, Feb. 2021.

[17] J. K. Behne and G. Reeves, "Fundamental limits for rank-one matrix estimation with groupwise heteroskedasticity," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8650–8672.

[18] M. Celentano, A. Montanari, and Y. Wu, "The estimation error of general first order methods," in *Proceedings of Thirty Third Conference on Learning Theory*. PMLR, Jul. 2020, pp. 1078–1141.

[19] A. Montanari and Y. Wu, "Statistically optimal first order algorithms: A proof via orthogonalization," arXiv preprint arXiv:2201.05101, Jan. 2022.

[20] R. Berthier, A. Montanari, and P.-M. Nguyen, "State evolution for approximate message passing with non-separable functions," *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 33–79, Mar. 2020.

[21] C. Gerbelot and R. Berthier, "Graph-based approximate message passing iterations," *Information and Inference: A Journal of the IMA*, vol. 12, no. 4, pp. 2562–2628, Dec. 2023.

[22] A. Montanari and A. S. Wein, "Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation," arXiv preprint arXiv:2212.06996, Dec. 2022.

[23] G. Reeves, "Non-asymptotic bounds on approximate message passing via Gaussian coupling," *preprint*, 2024.

[24] B. Çakmak, Y. M. Lu, and M. Opper, "Analysis of random sequential message passing algorithms for approximate inference," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2022, no. 7, p. 073401, Jul. 2022.

[25] M. Celentano, C. Cheng, and A. Montanari, "The high-dimensional asymptotics of first order methods with random data," arXiv preprint arXiv:2112.07572, Dec. 2021.

[26] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova, "Rigorous dynamical mean field theory for stochastic gradient descent methods," arXiv preprint arXiv:2210.06591, Oct. 2022.

[27] G. Reeves, "Information-theoretic limits for the matrix tensor product," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 777–798, 2020.

[28] R. Rossetti and G. Reeves, "Approximate message passing for the matrix tensor product model," arXiv preprint arXiv:2306.15580, Jun. 2023.

[29] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.

[30] M. Opper, B. Çakmak, and O. Winther, "A theory of solving TAP equations for Ising models with general invariant random matrices," *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 11, p. 114002, Feb. 2016.

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

[32] K. Takeuchi, "Bayes-Optimal Convolutional AMP," *IEEE Transactions on Information Theory*, vol. 67, no. 7, pp. 4405–4428, Jul. 2021.

[33] L. Liu, S. Huang, and B. M. Kurkoski, "Memory AMP," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 8015–8039, Dec. 2022.

[34] Z. Fan, "Approximate message passing algorithms for rotationally invariant matrices," *The Annals of Statistics*, vol. 50, no. 1, pp. 197–224, Feb. 2022.

[35] X. Zhong, T. Wang, and Z. Fan, "Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization," arXiv preprint arXiv:2110.02318, Oct. 2021.

[36] J. Barbier, F. Camilli, M. Mondelli, and M. Sáenz, "Fundamental limits in structured principal component analysis and how to reach them," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, p. e2302028120, Jul. 2023.

[37] R. Dudeja, Y. M. Lu, and S. Sen, "Universality of approximate message passing with semirandom matrices," *The Annals of Probability*, vol. 51, no. 5, pp. 1616–1683, Sep. 2023.

[38] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[39] Q. Lei, K. Zhong, and I. S. Dhillon, "Coordinate-wise power method," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[40] P. Xu, B. He, C. De Sa, I. Mitliagkas, and C. Re, "Accelerated stochastic power iteration," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018.

[41] H. Raja and W. Bajwa, "Distributed stochastic algorithms for high-rate streaming principal component analysis," *Transactions on Machine Learning Research*, Oct. 2022.

[42] X. Li, S. Wang, K. Chen, and Z. Zhang, "Communication-efficient distributed SVD via local power iterations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6504–6514.

[43] Z. Xu and P. Li, "Faster noisy power method," in *International Conference on Algorithmic Learning Theory*. PMLR, 2022, pp. 1138–1164.

[44] L. Balzano, Y. Chi, and Y. M. Lu, "Streaming PCA and subspace tracking: The missing data case," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1293–1310, 2018.

[45] C. Wang, Y. C. Eldar, and Y. M. Lu, "Subspace estimation from incomplete observations: A high-dimensional analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1240–1252, Dec. 2018.

[46] R. Rossetti, B. Nazer, and G. Reeves, "Linear operator approximate message passing (OpAMP)," arXiv preprint arXiv:2405.08225, May 2024.