Iterative Estimation of Constrained Rank-One Matrices in Noise

Sundeep Rangan Member IEEE, and Alyson K. Fletcher Member IEEE

Abstract—We consider the problem of estimating a rank-one matrix in Gaussian noise under a probabilistic model for the left and right factors of the matrix. The probabilistic model can impose constraints on the factors including sparsity and positivity that arise commonly in learning problems. We propose a simple iterative procedure that reduces the problem to a sequence of scalar estimation computations. The method is similar to approximate message passing techniques based on Gaussian approximations of loopy belief propagation that have been used recently in compressed sensing. Leveraging analysis methods by Bayati and Montanari, we show that the asymptotic behavior of the estimates from the proposed iterative procedure is described by a simple scalar equivalent model, where the distribution of the estimates is identical to certain scalar estimates of the variables in Gaussian noise. Moreover, the effective Gaussian noise level is described by a set of state evolution equations. The proposed method thus provides a computationally simple and general method for rank-one estimation problems with a precise analysis in certain high-dimensional settings.

Index Terms—approximate message passing, belief propagation, compressed sensing, dictionary learning, low-rank matrices, matrix factorization.

I. INTRODUCTION

We consider the problem of estimating vectors $u_0 \in \mathbb{R}^m$ and $v_0 \in \mathbb{R}^n$ from a matrix $A \in \mathbb{R}^{m \times n}$ of the form

$$A = u_0 v_0^T + \sqrt{m} W,$$

where $W$ represents some unknown noise and $\sqrt{m}$ is a normalization factor. The problem can be considered as a rank-one special case of finding a low-rank matrix in the presence of noise. Such low-rank estimation problems arise in a range of applications including blind channel estimation [1], antenna array processing [2], subspace system identification [3], and principal component or factor analysis [4].

When the noise term $W$ is zero, the vector pair $(u_0, v_0)$ can be recovered exactly, up to a scaling, from the maximal left and right singular vectors of $A$ [5]. However, in the presence of noise, the rank-one matrix can in general only be estimated approximately. In this case, a priori information or constraints on $(u_0, v_0)$ may improve the estimation. Such constraints arise, for example, in factor analysis in statistics, where one of the factors is often constrained to be either positive or sparse [6]. Similar sparsity constraints arise in the problem of dictionary learning [7]. In digital communications, one of the factors could come from a discrete QAM constellation.

Unfortunately, optimal estimation with constraints on $u_0$ or $v_0$ is often computationally intractable. The chief problem is the bilinear nature of the term $u_0 v_0^T$. However, the term is linear in $u_0$ and $v_0$ separately. Thus, many suboptimal estimation methods are performed in an iterative manner, alternately estimating $u_0$ and $v_0$ individually, while holding the estimate of the other factor constant.

This paper proposes a variant of such alternating optimization procedures that we call Iterative Factorization (IterFac), stated in detail in Algorithm 1 below. Through proper selection of relevant parameters, the IterFac algorithm can perform alternating minimizations to optimizations of the form

$$(\hat{u}, \hat{v}) := \arg \min_{u \in \mathbb{R}^m, v \in \mathbb{R}^n} F(u, v) = \frac{1}{2m} \| A - u v^T \|_F^2 + c_U(u) + c_V(v).$$

Here, $\|X\|_F$ is the Frobenius norm and $c_U(u)$ and $c_V(v)$ are cost or regularization functions on the left and right factors. In the case when the cost functions are separable, the IterFac algorithm reduces the vector optimization problem to a sequence of scalar optimization problems on the individual components of $u$ and $v$; it is thus computationally simple. The IterFac methodology is also general since the method can be applied to essentially arbitrary separable cost functions.

Of course, such iterative algorithms are by no means new; they underlie many existing methods, including the classic alternating power method for finding maximal singular values [5] and some alternating methods in sparse or non-negative dictionary learning [7]–[11]. More recently, an alternating method has been proposed for estimation with matrix uncertainties [12], also using an approximate message passing technique. The IterFac iterations also have some similarities to the updates in gradient descent methods in sparse dictionary learning such as in [13].

Our main contribution is to show that, under a particular setting of a damping parameter, the IterFac algorithm admits an asymptotically-exact characterization when $W$ is i.i.d. Gaussian noise and the components of the true vectors $u_0$ and $v_0$ have limiting empirical distributions. In this scenario, we show that the empirical joint distribution of the components of $u_0$ and the corresponding estimates from the IterFac algorithm are described by a simple scalar equivalent model where the IterFac component estimates are identically distributed to scalar estimates of the variables corrupted by Gaussian noise.
noise. Moreover, the effective Gaussian noise level in this model is described by a simple set of scalar state evolution (SE) equations. From the scalar equivalent model, one can compute the asymptotic value of almost any component-separable metric including mean-squared error or correlation. Thus, in addition to being computationally simple and general, the IterFac algorithm admits a precise analysis in the case of Gaussian noise. Moreover, since fixed points of the IterFac algorithm correspond, under suitable circumstances, to local minima of objectives such as \( G \), the analysis can be used to characterize the behavior of such minima—even if alternate algorithms to IterFac are used.

The main analytical tool is a recently-developed technique by Bayati and Montanari [14] used in the analysis of approximate message passing (AMP) algorithms. AMP methods are Gaussian approximations of loopy belief propagation for estimation of vectors under large random linear measurements. The work [14] applied AMP techniques to compressed sensing, and proved that, in the limit for large Gaussian mixing matrices, the behavior of AMP estimates can be described by a scalar equivalent model with effective noise levels defined by certain scalar state evolution (SE) equations. Similar SE analyses have appeared in related contexts [15]–[20]. To prove the SE equations for the IterFac algorithm, we apply a key theorem from [14] with a simple change of variables and a slight modification to account for parameter adaptation. A conference version of this paper appeared in [21]. This paper provides all the proofs along with more detailed discussions and simulations.

II. ITERATIVE RANK-ONE FACTORIZATION

For a matrix \( A \in \mathbb{R}^{m \times n} \), we consider the following iterative algorithm for estimating the rank-one factors \( u_0 \) and \( v_0 \).

**Algorithm 1 Iterative Factorization (IterFac)**

**Require:** Matrix \( A \in \mathbb{R}^{m \times n} \) and factor selection functions \( G_u(t, p, \lambda_u) \) and \( G_v(t, q, \lambda_v) \).

1: \( t \leftarrow 0 \)
2: Select initial values \( u(0), v(0) \)
3: repeat
4: \{ Update estimate of \( u \} \)
5: Select parameters \( \lambda_u(t) \) and \( \mu_u(t) \)
6: \( p(t) \leftarrow (1/m)Av(t) + \mu_u(t)u(t) \)
7: \( u(t+1) \leftarrow G_u(t, p(t), \lambda_u(t)) \)
8: \{ Update estimate of \( v \} \)
9: Select parameters \( \lambda_v(t) \) and \( \mu_v(t) \)
10: \( q(t) \leftarrow (1/m)A^Tu(t+1) + \mu_v(t)v(t) \)
11: \( v(t+1) \leftarrow G_v(t, q(t), \lambda_v(t)) \)
12: until Terminate

The output of the algorithm, \((u(t), v(t))\), \( t = 0, 1, \ldots \), is a sequence of estimates for \((u_0, v_0)\). The algorithm has several parameters including the initial conditions, the parameters in lines 5 and 9 the termination condition and, most importantly, the functions \( G_u(\cdot) \) and \( G_v(\cdot) \). In each iteration, the functions \( G_u(\cdot) \) and \( G_v(\cdot) \) are used to generate the estimates of the factors \( u(t) \) and \( v(t) \). \( G_u(\cdot) \) and \( G_v(\cdot) \) will thus be called the factor selection functions.

To understand the role of the factor selection functions, suppose that we wish to perform the optimization \( \mathbb{E} \) for some regularization functions \( c_U(\cdot) \) and \( c_V(\cdot) \). Consider the factor selections given by the minimization

\[
G_u(t, p, \lambda_u) := \arg \min_{u \in \mathbb{R}^m} \left[ -p^T u + c_U(u) + \frac{\lambda_u}{2} \| u \|^2 \right],
\]

\[
G_v(t, q, \lambda_v) := \arg \min_{v \in \mathbb{R}^n} \left[ -q^T v + c_V(v) + \frac{\lambda_v}{2} \| v \|^2 \right],
\]

where the parameters \( \lambda_u(t) \) and \( \lambda_v(t) \) are given by

\[
\lambda_u(t) := \mu_u(t) + \| v(t) \|^2/m,
\]

\[
\lambda_v(t) := \mu_v(t) + \| u(t+1) \|^2/m.
\]

**Lemma 1:** Consider the outputs of Algorithm 1 with the factor selection functions \( \mathbb{E} \), parameters \( \lambda_u(t) \) and \( \lambda_v(t) \) in \( \mathbb{E} \). Then, for any initial conditions and parameters selections \( \mu_u(t) \) and \( \mu_v(t) \),

\[
u(t+1) = \arg \min_{u} F(u, v(t)) + \frac{\mu_u(t)}{2} \| u - u(t) \|^2,
\]

\[
v(t+1) = \arg \min_{v} F(u(t+1), v) + \frac{\mu_v(t)}{2} \| v - v(t) \|^2.
\]

In addition, if \( \mu_u(t) \) and \( \mu_v(t) \geq 0 \), then the objective function is monotonically decreasing

\[
F(u(t+1), v(t+1)) \leq F(u(t), v(t)).
\]

**Proof:** See Appendix A.

To understand the lemma, first suppose \( \mu_u(t) = \mu_v(t) = 0 \). In this case, [6] show that the factors \( u(t+1) \) and \( v(t+1) \) are selected to minimize the objective function \( F(\cdot) \) in the optimization \( \mathbb{E} \). Thus, IterFac can be interpreted as minimizing the objective function, one factor at a time. Positive values of \( \mu_u(t) \) and \( \mu_v(t) \) add additional terms in the minimizations in [6] that attempt to keep the new columns \( u(t+1) \) and \( v(t+1) \) close to their previous values, \( u(t) \) and \( v(t) \), and can thus be interpreted as a damping in the iterative minimization. We thus call \( \mu_u(t) \) and \( \mu_v(t) \) the damping factors.

Equation [7] states that for any positive damping, the objective function monotonically decreases. Under mild additional assumptions, such as \( F(u, v) \) being bounded below, one can show that the estimates \( (u(t), v(t)) \) converge to a local minimum of the objective.

A. Separable Functions

Our interest in the IterFac algorithm will mostly be in the case when the cost functions \( c_U(u) \) and \( c_V(v) \) are separable in that they are of the form

\[
c_U(u) = \sum_{i=1}^{m} c_{U i}(u_i), \quad c_V(v) = \sum_{j=1}^{n} c_{V j}(v_j),
\]
where \( c_{ui}(u_i) \) and \( c_{vij}(v_j) \) are cost functions on the individual components. In this case, the factor selection functions can be implemented by componentwise minimizations

\[
G_u(t, p, \lambda_u)_{i} := \arg \min_{u_i \in \mathbb{R}} \left\{ -p_i u_i + c_{ui}(u_i) + \frac{\lambda_u}{2} u_i^2 \right\} \quad (9a)
\]

\[
G_v(t, q, \lambda_v)_{j} := \arg \min_{v_j \in \mathbb{R}} \left\{ -q_j v_j + c_{vij}(v_j) + \frac{\lambda_v}{2} v_j^2 \right\} \quad (9b)
\]

We will see that a similar separability property holds for MMSE estimation under the assumption of independent priors (see Section IV-B).

In the separable case, the IterFac algorithm reduces the vector-valued bilinear optimization to a sequence of scalar minimizations and matrix transforms. The complexity is low: Assuming the complexity of each scalar minimization does not grow with the dimension, the complexity of all the scalar minimizations together will be \( O(m + n) \) per iteration. The complexity of the matrix multiplications will be \( O(mn) \), so the total cost per iteration is \( O(mn) \). In addition, we will see from the state evolution analysis in Section III-B that the per-component performance, the number of iterations will not grow with dimension. Hence the total cost of the IterFac algorithm is \( O(mn) \) — the same as a matrix multiplication. Thus, the IterFac algorithm offers the possibility, under separability assumptions on the cost functions, of computationally fast implementations for a large class of problems.

III. ASYMPTOTIC ANALYSIS UNDER GAUSSIAN NOISE

A. Model and Assumptions

We analyze the algorithm under the following assumptions. 

**Assumption 1:** Consider a sequence of random realizations of the estimation problem in Section II indexed by the dimension \( n \). The matrix \( \mathbf{A} \) and the parameters in Algorithm 1 satisfy the following:

(a) For each \( n \), the output dimension \( m = m(n) \) is deterministic and scales linearly with the input dimension in that

\[
\lim_{n \to \infty} n/m(n) = \beta \quad (10)
\]

for some \( \beta > 0 \).

(b) The matrix \( \mathbf{A} \) has the form \( \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix} \) where \( \mathbf{U}_0 \in \mathbb{R}^m \) and \( \mathbf{V}_0 \in \mathbb{R}^n \) represent “true” left and right factors of a rank one term, and \( \mathbf{W} \in \mathbb{R}^{m \times n} \) is an i.i.d. Gaussian matrix with components \( W_{ij} \sim N(0, \tau_w) \) for some \( \tau_w > 0 \).

(c) The factor selection functions \( G_u(t, p, \lambda_u) \) and \( G_v(t, q, \lambda_v) \) in lines 7 and 11 are componentwise separable in that for all component indices \( i \) and \( j \),

\[
G_u(t, p, \lambda_u)_{i} = G_u(t, p_i, \lambda_u), \quad (11a) \]

\[
G_v(t, q, \lambda_v)_{j} = G_v(t, q_j, \lambda_v), \quad (11b) \]

for some scalar functions \( G_u(t, p, \lambda_u) \) and \( G_v(t, q, \lambda_v) \). The scalar functions must be differentiable in \( p \) and \( q \). Moreover, for every \( t \), the functions \( G_u(t, p, \lambda_u) \) and \( \partial G_u(t, p, \lambda_u)/\partial p \) must be Lipschitz continuous in \( p \) with a Lipschitz constant that is continuous in \( \lambda_u \), and continuous in \( \lambda_v \), uniformly over \( p \). Similarly, for every \( t \), the functions \( G_v(t, q, \lambda_v) \) and \( \partial G_v(t, q, \lambda_v)/\partial q \) must be Lipschitz continuous in \( q \) with a Lipschitz constant that is continuous in \( \lambda_v \), and continuous in \( \lambda_u \), uniformly over \( q \).

(d) The damping factors are selected by the rules

\[
\mu_u(t+1) = - \frac{\tau_u}{m} \sum_{j=1}^{n} \frac{\partial}{\partial q_j} G_v(t, q_j(t), \lambda_v(t)) \quad (12a)
\]

\[
\mu_v(t) = - \frac{\tau_v}{m} \sum_{i=1}^{m} \frac{\partial}{\partial p_i} G_u(t, p_i(t), \lambda_u(t)) \quad (12b)
\]

with the initial damping factor \( \mu_u(0) = 0 \).

(e) The parameters \( \lambda_u(t) \) and \( \lambda_v(t) \) are computed via

\[
\lambda_u(t) = \frac{1}{m} \sum_{j=1}^{n} \phi_{\lambda_u}(t, u_{0j}, v_j(t)) \quad (13a)
\]

\[
\lambda_v(t) = \frac{1}{m} \sum_{i=1}^{m} \phi_{\lambda_v}(t, u_{0i}, u_{ij}(t+1)) \quad (13b)
\]

for (possibly vector-valued) functions \( \phi_{\lambda_u} \) and \( \phi_{\lambda_v} \) that are pseudo-Lipschitz continuous of order \( p \) for some \( p \geq 2 \).

(f) For each \( n \) and iteration number \( t \), define the sets

\[
\theta_u(t) = \left\{ (u_{0i}, u_i(t)), i = 1, \ldots, m \right\}, \quad (14a)
\]

\[
\theta_v(t) = \left\{ (v_{0j}, v_j(t)), j = 1, \ldots, n \right\}. \quad (14b)
\]

There the sets for \( t = 0 \) empirically converge with bounded moments of order \( 2p - 2 \) to the limits

\[
\lim_{n \to \infty} \theta_u(0) \overset{d}{=} (U_0, U(0)), \quad (15a)
\]

\[
\lim_{n \to \infty} \theta_v(0) \overset{d}{=} (V_0, V(0)), \quad (15b)
\]

for some random variable pairs \((U_0, U(0))\) and \((V_0, V(0))\). See Appendix B for a precise definition of the empirical convergence used here.

The assumptions need some explanations. Assumptions 1(a) and (b) simply state that we are considering an asymptotic analysis for certain large matrices \( \mathbf{A} \) consisting of a random rank one matrix plus Gaussian noise. The analysis of Algorithm 1 for higher ranks is still not known, but we provide some possible ideas later. Assumption 1(c) is a mild condition on the factor selection functions. For example, as discussed above, the separability condition would occur if we use the factor selection functions \( \hat{G} \) with element separable cost functions. Assumption (d) provides a precise prescription for the damping factors. Interestingly, these factors may be negative.

Assumption (e) allows for the parameters \( \lambda_u(t) \) and \( \lambda_v(t) \) in the factor selection functions to be data dependent, provided that they can each be determined via empirical averages of some function of the most recent data. Note that the functions in 5 are of the form 13, and therefore the factor selection functions \( \hat{G} \) would satisfy these assumptions. As discussed in Section VII, it is possible that these parameters can also be used to estimate unknown parameters in the distributions of the components of \( u_0 \) and \( v_0 \), if they are not known.
To understand Assumption 1(f), note that the set \( \theta_u(t) \) is the set of components of the “true” left factor \( u_0 \) and the corresponding components in the estimate \( u(t) \) after \( t \) iterations of Algorithm 1. Similarly, \( \theta_v(t) \) represents the components of the “true” right factor \( v_0 \) and its estimate \( v(t) \). Following Bayati and Montanari’s analysis in [14] we will treat the variables as deterministic. However, the condition (15) requires that empirical distribution of the initial conditions asymptotically behave like some random vector pairs \( (U_0, U(0)) \) and \( (V_0, V(0)) \).

### B. Main Result

Our main result, Theorem 1 below, will provide a characterization of the asymptotic empirical distribution for the sets \( \theta_u(t) \) and \( \theta_v(t) \) in [14]. Specifically, we will show that, for all \( t \), the sets have empirical limits of the form

\[
\frac{\lim_{n \to \infty} \theta_u(t)}{\lim_{n \to \infty} \theta_v(t)} = (U_0, U(t)),
\]

(16a)

\[
\frac{\lim_{n \to \infty} \theta_v(t)}{\lim_{n \to \infty} \theta_v(t)} = (V_0, V(t)),
\]

(16b)

for some random variable pairs \( (U_0, U(t)) \) and \( (V_0, V(t)) \). The distributions of the random variables \( U_0, U(t) \) and \( V_0, V(t) \) can be described recursively in \( t \) as follows: For \( t = 0 \), \( (V_0, V(0)) \) is the random variable pair in (15b) in the initial condition Assumption 1(t). Then, for \( t \geq 0 \), \( (U_0, U(t+1)) \) is given by

\[
U(t+1) = G_u(t, P(t), \bar{v}_u(t)),
\]

(17a)

\[
P(t) = \beta u_1(t) U_0 + Z_u(t),
\]

(17b)

\[
Z_u(t) \sim \mathcal{N}(0, \beta \bar{v}_u(t)),
\]

(17c)

where \( U_0 \) is the random variable in (15b) that describes the empirical limit of the components of the true vector \( u_0 \); \( G_u(\cdot) \) is the scalar function in Assumption 1(d); and \( Z_u(t) \) is Gaussian noise independent of \( U_0 \). Thus, \( U(t+1) \) is distributed identically to the output of the scalar function \( G_u(t, P(t)) \), where \( P(t) \) is a scaled and Gaussian noise-corrupted version of the true variable \( U_0 \). Similarly, \( V(t+1) \) is the output of the factor selection function \( G_v(t, \cdot) \) applied to a scaled and Gaussian noise-corrupted version of the true variable \( V_0 \). Following [22], we can thus call the result a single-letter characterization of the algorithm.

From this single-letter characterization, one can exactly compute a large class of performance metrics of the algorithm. Specifically, the empirical convergence of \( \theta_u(t) \) shows that for any pseudo-Lipschitz function \( \phi(u_0, u) \) of order \( p \), the following limit exists almost surely:

\[
\lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m} \phi(u_{i0}, u_{i1}) = \mathbb{E}[\phi(U_0, U(t))],
\]

(21)

where the expectation on the right-hand side is over the variables \( (U_0, U(t)) \) with \( U_0 \) identical to the variable in the limit in (15b) and \( U(t) \) given by (17a). This expectation can thus be explicitly evaluated by a simple two-dimensional integral and consequently any component-separable performance metric based on a suitably continuous loss function \( \phi(u_0, u) \) can be exactly computed.

For example, if we take \( \phi(u_0, u) = (u - u_0)^2 \) we can compute the asymptotic mean squared error of the estimate,

\[
\lim_{n \to \infty} \frac{1}{m} \|u(t) - u_0\|^2 = \lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m} (u_{i} - u_{i0})^2
\]

\[
= \mathbb{E}[(U_0 - U(t))^2].
\]

(22)

Also, for each \( t \), define the empirical second-order statistics

\[
\alpha_{u0}(t) = \frac{1}{m} \|u(t)|^2, \quad \alpha_{u1}(t) = \frac{1}{m} u(t)^T u_0,
\]

(23a)

\[
\alpha_{v0}(t) = \frac{1}{n} \|v(t)|^2, \quad \alpha_{v1}(t) = \frac{1}{n} v(t)^T v_0.
\]

(23b)

Since \( \|u(t)|^2 = \sum_i u_i(t)^2 \), it follows that \( \alpha_{u0}(t) \to \mathbb{E}(U(t))^2 \) almost surely as \( n \to \infty \). In this way, we obtain

### C. Scalar Equivalent Model

The main contribution of Theorem 1 is that it provides a simple scalar equivalent model for the asymptotic behavior of the algorithm. The sets \( \theta_u(t) = \{(u_{i0}, u_{i1})\} \) and \( \theta_v(t) = \{(v_{i0}, v_{i1})\} \) in [14] are the components of true vectors \( u_0 \) and \( v_0 \) and their estimates \( u(t) \) and \( v(t) \). The theorem shows that empirical distribution of these components are asymptotically equivalent to simple random variable pairs \( (U_0, U(t)) \) and \( (V_0, V(t)) \) given by (17a) and (18a). In this scalar system, the variable \( U(t+1) \) is the output of the factor selection function \( G_u(t, \cdot) \) applied to a scaled and Gaussian noise-corrupted version of the true variable \( U_0 \). Similarly, \( V(t+1) \) is the output of the factor selection function \( G_v(t, \cdot) \) applied to a scaled and Gaussian noise-corrupted version of the true variable \( V_0 \). Following [22], we can thus call the result a single-letter characterization of the algorithm.


that the following limits hold almost surely:
\[
\lim_{n \to \infty} \alpha_{u0}(t) = \mathbf{\alpha}_{u0}(t), \quad \lim_{n \to \infty} \alpha_{u1}(t) = \mathbf{\alpha}_{u1}(t), \tag{24a}
\]
\[
\lim_{n \to \infty} \alpha_{v0}(t) = \mathbf{\alpha}_{v0}(t), \quad \lim_{n \to \infty} \alpha_{v1}(t) = \mathbf{\alpha}_{v1}(t). \tag{24b}
\]
We will also use definitions
\[
\tau_u := \mathbb{E}[U_0^2], \quad \tau_v := \mathbb{E}[V_0^2]. \tag{25}
\]
From the second order statistics, we can compute the asymptotic correlation coefficient between \(u_0\) and its estimate \(\hat{u}\) given by
\[
\rho_u(t) := \lim_{n \to \infty} \frac{|u(t)^T u_0|^2}{\|u(t)\|^2 \|u_0\|^2} = \lim_{n \to \infty} \frac{(\mathbb{E}(u(t)u_0)/m)^2}{(\mathbb{E}(u(t)^2)/m)(\mathbb{E}(u_0^2)/m)} = \frac{\mathbb{E}(U(t)U_0)^2}{\mathbb{E}U(t)^2\mathbb{E}U_0^2} \frac{\sigma_u^2(t)}{\tau_u u_0(t) \tau_u}. \tag{26}
\]
Similarly, the asymptotic correlation coefficient between \(v_0\) and \(v\) has a simple expression
\[
\rho_v(t) := \lim_{n \to \infty} \frac{|v(t)^T v_0|^2}{\|v(t)\|^2 \|v_0\|^2} = \frac{\sigma_v^2(t)}{\tau_v v_0(t) \tau_v}. \tag{27}
\]
The correlation coefficient is useful, since we know that, without additional constraints, the terms \(u_0\) and \(v_0\) can only be estimated up to a scalar. The correlation coefficient is scale invariant.

IV. EXAMPLES

A. Linear Selection Functions

As a first simple application of the SE analysis, suppose we use linear selection functions of the form
\[
G_u(t,p,\lambda_u) = \lambda_u p, \quad G_v(t,q,\lambda_v) = \lambda_v q, \tag{28}
\]
where the parameters \(\lambda_u\) and \(\lambda_v\) allow for normalization or other scalings of the outputs. Linear selection functions of the form (28) arise when one selects \(G_u(\cdot)\) and \(G_v(\cdot)\) from (4) with zero cost functions, \(c_u(\cdot) = c_v(\cdot) = 0\). With zero cost functions, the correct solution to the optimization (4) is for \((u,v)\) to be the (appropriately scaled) left and right maximal singular vectors of \(A\). We will thus call the estimates \((\hat{u}(t),\hat{v}(t))\) of Algorithm [1] and linear selection functions (28) the estimated maximal singular vectors.

**Theorem 2:** Consider the state evolution equations (17), (18), and (19) with the linear selection functions (28). Then: (a) The asymptotic correlation coefficients (26) and (27) satisfy the following recursive rules:
\[
\rho_u(t+1) = \frac{\beta \tau_u \tau_v \rho_u(t)}{\beta \tau_u \tau_v \rho_u(t) + \tau_w}, \tag{29a}
\]
\[
\rho_v(t) = \frac{\tau_v \rho_v(t)}{\tau_u \tau_v \rho_u(t) + \tau_w}. \tag{29b}
\]
(b) For any positive initial condition, \(\rho_u(0) > 0\), the asymptotic correlation coefficients converge to the limits
\[
\lim_{t \to \infty} \rho_u(t) = \rho_u^* := \frac{\beta \tau_u \tau_v \rho_u(t) + \tau_w}{\tau_u \tau_v \rho_u(t) + \tau_w}, \tag{30a}
\]
\[
\lim_{t \to \infty} \rho_v(t) = \rho_v^* := \frac{\beta \tau_u \tau_v \rho_v(t) + \tau_w}{\tau_u \tau_v \rho_v(t) + \tau_w}. \tag{30b}
\]
where \([x]_+ = \max\{0,x\}\).

**Proof:** See Appendix E.

The theorem provides a set of recursive equations for the asymptotic correlation coefficients \(\rho_u(t)\) and \(\rho_v(t)\) along with simple expressions for the limiting values as \(t \to \infty\). We thus obtain exactly how correlated the estimated maximal singular vectors of a matrix \(A\) are to the rank one factors \((u_0,v_0)\). The proof of the theorem also provides expressions for the second-order statistics in (19) to be used in the scalar equivalent model.

The fixed point expressions (30) agree with the more general results in (23) that derive the correlations for ranks greater than one and low-rank recovery with missing entries. Similar results can also be found in (24). An interesting consequence of the expressions in (30) is that unless \(\sqrt{\beta \tau_u \tau_v} > \tau_w\), (31)
the asymptotic correlation coefficients are exactly zero. The ratio \(\tau_u \tau_v / \tau_w\) can be interpreted as a scaled SNR.

B. Minimum Mean-Squared Error Estimation

Next suppose that the priors on \(U_0\) and \(V_0\) are known. In this case, given (17a) and (18a), a natural choice for the factor selection functions are
\[
G_u(t,p) = \mathbb{E}[U_0 | P(t)], \quad G_v(t,q) = \mathbb{E}[V_0 | Q(t)], \tag{32}
\]
which are the MMSE estimates of the variables. In this example, there are no parameters \(\lambda_u\) or \(\lambda_v\). We can use the initial condition \(\nu_j(0) = \mathbb{E}[V_0]\) for all \(j\), so that the initial variable in (15b) is \(V(0) = \mathbb{E}[V_0]\). To analyze the algorithms define
\[
\mathcal{E}_u(\eta_u) := \text{var}(U_0 | Y = \sqrt{\eta_u} U_0 + D), \tag{33a}
\]
\[
\mathcal{E}_v(\eta_u) := \text{var}(V_0 | Y = \sqrt{\eta_u} V_0 + D), \tag{33b}
\]
where \(D \sim N(0,1)\) is independent of \(U_0\) and \(V_0\). That is, \(\mathcal{E}_u(\eta_u)\) and \(\mathcal{E}_v(\eta_u)\) are the mean-squared errors of estimating \(U_0\) and \(V_0\) from observations \(Y\) with SNRs of \(\eta_u\) and \(\eta_v\). The functions \(\mathcal{E}_u(\cdot)\) and \(\mathcal{E}_v(\cdot)\) arise in a range of estimation problems and the analytic and functional properties of these functions can be found in (25), (26).

**Theorem 3:** Consider the solutions to the SE equations (17), (18), and (19) under the MMSE selection functions (32) and initial condition \(V(0) = \mathbb{E}[V_0]\). Then:
(a) For all \(t\), the asymptotic correlation coefficients (26) and (27) satisfy the recursive relationships
\[
\rho_u(t+1) = 1 - \frac{1}{\tau_u} \mathcal{E}_u(\beta \tau_v \rho_u(t) / \tau_w), \tag{34a}
\]
\[
\rho_v(t) = 1 - \frac{1}{\tau_v} \mathcal{E}_v(\tau_u \rho_u(t) / \tau_w). \tag{34b}
\]
with the initial condition \( \rho_v(0) = \langle EV_0 \rangle^2 / \tau_v \).

(b) If, in addition, \( \mathcal{E}_u(\eta_u) \) and \( \mathcal{E}_v(\eta_v) \) are continuous, then for any positive initial condition, \( \rho_u(0) > 0 \), as \( t \to \infty \), the asymptotic correlation coefficients \( (\rho_u(t), \rho_v(t)) \) increase monotonically to fixed points \( (\rho_u^*, \rho_v^*) \) of (34) with \( \rho_u^* > 0 \).

Proof: See Appendix \[ \square \]

Again, we see that we can obtain simple, explicit recursions for the asymptotic correlations. Moreover, the asymptotic correlations provably converge to fixed points of the SE equations. The proof of the theorem also provides expressions for the second-order statistics in \[ \square \] to be used in the scalar equivalent model.

C. Zero Initial Conditions

The limiting condition in part (b) of Theorem \[ \square \] requires that \( \rho_v(0) > 0 \), which occurs when \( \mathbb{E}[V_0] \neq 0 \). Suppose, on the other hand, that \( \mathbb{E}[U_0] = \mathbb{E}[V_0] = 0 \). Then, the initial condition will be \( V(0) = \mathbb{E}[V_0] = 0 \). Under this initial condition, a simple set of calculations show that the SE equations (34) will generate a sequence with \( \rho_v(t) = \rho_u(t) = 0 \) for all \( t \). Thus, the IterFac algorithm will produce no useful estimates.

Of course, with zero mean random variables, a more sensible initial condition is to take \( \mathbf{v}(0) \) to be some non-zero random vector, as is commonly done in power algorithm recursions for computing maximal singular vectors. To understand the behavior of the algorithm under this random initial condition, let

\[
\rho_v(t, n) := \frac{\langle \mathbf{v}(t)^T \mathbf{v}_0 \rangle^2}{\|\mathbf{v}_0\|^2 \|\mathbf{v}(t)\|^2},
\]

(35)

where we have explicitly denoted the dependence on the problem dimension \( n \). From (27), we have that \( \lim_{n \to \infty} \rho_v(t, n) = \rho_v(t) \) for all \( t \). Also, with a random initial condition \( \mathbf{v}(0) \) independent of \( \mathbf{v}_0 \), it can be checked that \( \rho_v(0, n) = O(1/n) \) so that

\[
\rho_v(0) = \lim_{n \to \infty} \rho_v(0, n) = 0.
\]

Hence, from the SE equations (34), \( \rho_v(t) = \rho_u(t) = 0 \) for all \( t \). That is,

\[
\lim_{t \to \infty} \lim_{n \to \infty} \rho(t, n) = 0.
\]

(36)

This limit suggests that, even with random initial condition, the IterFac algorithm will not produce a useful estimate.

However, it is still possible that the limit in the opposite order of (36) may be non-zero:

\[
\lim_{n \to \infty} \lim_{t \to \infty} \rho(t, n) > 0.
\]

(37)

That is, for each \( n \), it may be possible to obtain a non-zero correlation, but the number of iterations for convergence increases with \( n \) since the algorithm starts from a decreasingly small initial correlation. Unfortunately, our SE analysis cannot make predictions on limits in the order of (37).

We can however analyze the following limit:

Lemma 2: Consider the MMSE SE equations (34) with random variables \( U_0 \) and \( V_0 \) such that \( \mathbb{E}[V_0] = \mathbb{E}[U_0] = 0 \). For each \( \epsilon > 0 \), let \( \rho_v^\epsilon(t) \) be the solution to the SE equations with an initial condition \( \rho_v(0) = \epsilon \). Then,

(a) If \( \sqrt{3 \tau_u \tau_v} > \tau_w \),

\[
\lim_{t \to \infty} \lim_{n \to \infty} \rho_v^\epsilon(t) > 0.
\]

(b) Conversely, if \( \sqrt{3 \tau_u \tau_v} < \tau_w \),

\[
\lim_{t \to \infty} \lim_{n \to \infty} \rho_v^\epsilon(t) = 0.
\]

Proof: See Appendix \[ \square \]

The result of the lemma is somewhat disappointing. The lemma shows that \( \sqrt{3 \tau_u \tau_v} > \tau_w \) is essentially necessary and sufficient for the IterFac algorithm with MMSE estimates to be able to overcome arbitrarily small initial conditions and obtain an estimate with a non-zero correlation to the true vector. Unfortunately, this is the identical to the condition (31) for the linear estimator to obtain a non-zero correlation. Thus, the IterFac algorithm with MMSE estimates performs no better than simple linear estimation in the initial iterations when the priors have zero means. Since linear estimation is equivalent to finding maximal singular vectors without any particular constraints, we could interpret Lemma 2 as saying that the IterFac algorithm under MMSE estimation cannot exploit structure in the components in the initial iterations. As a result, in low SNRs it may be necessary to use other algorithms as an initial condition for IterFac – such procedures, however, require further study.

V. NUMERICAL SIMULATION

To validate the SE analysis, we consider a simple case where the left factor \( \mathbf{u}_0 \in \mathbb{R}^m \) is i.i.d. Gaussian, zero mean and \( \mathbf{v}_0 \in \mathbb{R}^n \) has Bernoulli-Exponential components:

\[
\mathbf{v}_0 \sim \begin{cases} 
0 & \text{with prob } 1 - \lambda, \\
\text{Exp}(1) & \text{with prob } \lambda,
\end{cases}
\]

(40)

which provides a simple model for a sparse, positive vector. The parameter \( \lambda \) is the fraction of nonzero components and is set in this simulation to \( \lambda = 0.1 \). Note that these components have a non-zero mean so the difficulties of Section IV-C are avoided. The dimensions are \( (m, n) = (1000, 500) \), and the noise level \( \tau_w \) is set according to the scaled SNR defined as

\[
\text{SNR} = 10 \log_{10} \left( \tau_u \tau_v / \tau_w \right).
\]

(41)

Estimating the vector \( \mathbf{v}_0 \) in this set-up is related to finding sparse principal vectors of the matrix \( \mathbf{A}^T \mathbf{A} \) for which there are large number of excellent methods including \[ \square \] to name a few. These algorithms include methods based on thresholding, \( \ell_1 \)-regularization and semidefinite programming. A comparison of the IterFac against these methods would be an interesting avenue of future research. Here, we simply wish to verify the SE predictions of the IterFac method.

The results of the simulation are shown in Fig. 1, which shows the simulated and SE-predicted performance of the IterFac algorithm with both the linear and MMSE selection functions for the priors on \( \mathbf{u} \) and \( \mathbf{v} \). The algorithm is run for \( t = 10 \) iterations and the plot shows the median of the final correlation coefficient \( \rho_v(t) \) over 50 Monte Carlo trials at each value of SNR. It can be seen that the performance of the IterFac algorithm for both the linear and MMSE estimates are in excellent agreement with the SE predictions. The correlation
benefit of exploiting the prior on find the maximal singular vectors. The figure also shows the IterFac algorithm is essentially an iterative method to Section IV-B (IF-mmse). The simulated values are compared against the SE selection functions in Section IV-A (IF-lin) and MMSE selection functions in

Also shown is the simple estimate from the maximal singular vectors of 10 iterations. The simulated values are compared against the SE predictions. 

Fig. 2. Per iteration performance of the IterFac algorithm for both the linear selection functions in Section IV-A (labeled iter-lin) and MMSE selection functions in Section IV-B (labeled iter-mmse). Plotted are the correlation values after 10 iterations. The simulated values are compared against the SE predictions. Also shown is the simple estimate from the maximal singular vectors of A.

coefficient of the linear estimator also matches the correlation of the estimates produced from the maximal singular vectors of A. This is not surprising since, with linear selection functions, the IterFac algorithm is essentially an iterative method to find the maximal singular vectors. The figure also shows the benefit of exploiting the prior on v0, which is evident from the superior performance of the MMSE estimate over the linear reconstruction.

Fig. 2 shows the correlation coefficient as a function of the iteration number for the MMSE and linear methods for two values of the SNR. Again, we see that the SE predictions are closely matched to the median simulated values. In addition, we see that we get good convergence within 4 to 8 iterations. Based on the SE predictions, this number will not scale with dimension and hence the simulation suggests that only a small number of iterations will be necessary for even very large problems. All code for the simulation can be found in the public GAMP sourceforge project.

VI. LIMITATIONS AND EXTENSIONS

While our initial analysis of the IterFac algorithm is encouraging, there are several issues that can be explored for future work:

a) Extensions to higher rank: The current algorithm only works on rank one matrices, while many problems require estimation of low-rank matrices. One simple way to extend the current method to higher ranks is to perform the estimation one column at a time for each matrix factor. In each column update, the matrix A would be replaced by the residual from subtracting the current values of the other columns. Lemma 1 can be easily generalized to show monotonicity in this case. However, the state evolution analysis will be more difficult to generalize since it would require that the residual error be approximately Gaussian.

b) Unknown priors: The MMSE estimator in Section IV-B requires exact knowledge of the priors on \( U_0 \) and \( V_0 \) as well as the Gaussian noise level \( \sigma_n \). In many problems in statistics, these are not known. There are two possible solutions that may be investigated in the future. One method is to parameterize the distributions of \( U_0 \) and \( V_0 \) and estimate these parameters in the MMSE selection functions \( [32] \) — potentially through an EM type procedure as in \( [33] \). This EM type approach with hidden hyperparameters has been recently successfully used in a related approximate message passing method in \( [34] \). The analysis of the such parameter learning could possibly be accounted for through the adaptation parameters \( \lambda_0(t) \) and \( \lambda_0(t) \). A second approach is to assume that the distributions of \( U_0 \) and \( V_0 \) belong to a known family of distributions and then find a min-max solution. Such a min-max technique was proposed for AMP recovery of sparse vectors in \( [35] \). See also \( [36] \).

c) Optimality: While the current paper characterizes the performance of the IterFac algorithm, it remains open how far that performance is to optimal estimation such as the joint MMSE estimates of \( u_0 \) and \( v_0 \). We conjecture that the asymptotic correlation of the optimal estimates of \( u_0 \) and \( v_0 \) may also be fixed points to \( (34) \). If this conjecture were true than the uniqueness of the fixed points would provide a testable condition for the optimality of IterFac. The reason to suspect this optimality property is by comparison to several works that have derived similar scalar equivalent models and state evolution equations for regression problems with large random measurement matrices. In these works, belief propagation analysis of large, random sparse matrices \( [15] \), \( [17] - [19] \), as well as replica symmetric analysis of dense i.i.d. matrices \( [37] - [39] \) suggest that the performance of optimal estimators can be described by the same fixed points as the limiting points of approximate message passing (AMP). Thus,

```text
Fig. 1. Simulation of the IterFac algorithm for both the linear selection functions in Section IV-A (labeled iter-lin) and MMSE selection functions in Section IV-B (labeled iter-mmse). Plotted are the correlation values after 10 iterations. The simulated values are compared against the SE predictions.

Fig. 2. Per iteration performance of the IterFac algorithm for both the linear selection functions in Section IV-A (IF-lin) and MMSE selection functions in Section IV-B (IF-mmse). The simulated values are compared against the SE predictions.
```
AMP is potentially optimal in these scenarios when the fixed points are unique.

**Conclusions and Future Work**

We have presented a computationally-efficient method for estimating rank-one matrices in noise. The estimation problem is reduced to a sequence of scalar AWGN estimation problems which can be performed easily for a large class of priors or regularization functions on the coefficients. In the case of Gaussian noise, the asymptotic performance of the algorithm is exactly characterized by a set of scalar state evolution equations which appear to match the performance at moderate dimensions well. Thus, the methodology is computationally simple, general and admits a precise analysis in certain asymptotic, random settings. Future work include extensions to higher rank matrices and handling of the cases where the priors are not known.

**Appendix A**

**Proof of Lemma 1**

To prove (6a), first observe that

\[
F(u, v(t)) + \frac{\mu_v(t)}{2}\|u - u(t)\|^2 \leq F(u(t+1), v(t)) + \frac{\mu_u(t)}{2}\|u(t+1) - u(t)\|^2.
\]

where (a) follows from the assumption that the damping factor \(\mu_u(t)\) is non-negative and (b) follows since \(u(t+1)\) is the solution to the minimization (6a). Similarly, one can show that, since \(\mu_v(t) \geq 0\),

\[
F(u(t+1), v(t+1)) \leq F(u(t+1), v(t)).
\]

Together (43) and (44) prove (7).

**Appendix B**

**Empirical Convergence of Random Variables**

Bayati and Montanari’s analysis in [14] employs certain deterministic models on the vectors and then proves convergence properties of related empirical distributions. To apply the same analysis here, we need to review some of their definitions. We say a function \(\phi : \mathbb{R}^r \to \mathbb{R}^s\) is pseudo-Lipschitz of order \(p > 1\), if there exists an \(L > 0\) such for any \(x, y \in \mathbb{R}^r\),

\[
\|\phi(x) - \phi(y)\| \leq L(1 + \|x\|^{p-1} + \|y\|^{p-1})\|x - y\|.
\]

Now suppose that for each \(n = 1, 2, \ldots\) we have a set of vectors

\[
\theta(n) = \{v_i(n), i = 1, \ldots, \ell(n)\},
\]

where the elements are vectors \(v_i(n) \in \mathbb{R}^s\), and the size of the set is given by \(\ell(n)\). We say that the set of components of \(\theta(n)\) empirically converges with bounded moments of order \(p\) as \(n \to \infty\) to a random vector \(V\) on \(\mathbb{R}^s\) if: For all pseudo-Lipschitz continuous functions, \(\phi\), of order \(p\),

\[
\lim_{n \to \infty} \frac{1}{\ell(n)} \sum_{i=1}^{\ell(n)} \phi(v_i(n)) = \mathbb{E}[\phi(V)] < \infty.
\]

When the nature of convergence is clear, we may write (with some abuse of notation)

\[
\lim_{n \to \infty} \theta(n) \equiv V.
\]

**Appendix C**

**Bayati–Montanari Recursions with Adaptation**

Our main result will need an adaptive version of the recursion theorem of Bayati and Montanari [14]. Let \(H_u(t, d, u_0, \nu_u)\) and \(H_v(t, b, v_0, \nu_v)\) be two functions defined on arguments \(t = 0, 1, 2, \ldots\) and \(d, b, u_0\) and \(v_0\) as well as vectors \(\nu_u\) and \(\nu_v\). Given a matrix \(S \in \mathbb{R}^{m \times n}\) and vectors \(u_0\) and \(v_0\), generate a sequence of vectors \(b(t)\) and \(d(t)\) by the iterations

\[
\begin{align*}
b(t) &= S v(t) + \xi_u(t) u(t), \\
d(t) &= S^T u(t+1) + \xi_v(t) v(t)
\end{align*}
\]

where

\[
\begin{align*}
u_i(t+1) &= H_u(t, b_i(t), u_0i, \nu_u(t)), \\
v_j(t+1) &= H_v(t, d_j(t), v_0j, \nu_v(t)),
\end{align*}
\]

and \(\xi_u(t)\) and \(\xi_v(t)\) are scalar step sizes given by

\[
\begin{align*}
\xi_u(t) &= -\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial b_i} H_u(t, b_i(t), u_0i, \nu_u(t)), \\
\xi_v(t+1) &= -\frac{1}{m} \sum_{j=1}^{m} \frac{\partial}{\partial d_j} H_v(t, d_j(t), v_0j, \nu_v(t)).
\end{align*}
\]

The recursions (46) to (48) are identical to the recursions analyzed in [14], except for the introduction of the parameters \(\nu_u(t)\) and \(\nu_v(t)\). We will call these parameters adaptation parameters since they enable the functions \(H_u(\cdot)\) and \(H_v(\cdot)\) to have some data dependence, not explicitly considered in [14]. Similar to the selection of the parameters \(\lambda_u(t)\) and \(\lambda_v(t)\)
in (13), we assume that, in each iteration $t$, the adaptation parameters are selected by functions of the form,

\[
\nu_u(t) = \frac{1}{n} \sum_{j=1}^{n} \phi_u(t, v_0j, v_j(t)), \quad (49a)
\]

\[
\nu_v(t) = \frac{1}{m} \sum_{i=1}^{m} \phi_v(t, u_{0i}, u_i(t+1)) \quad (49b)
\]

where $\phi_u(\cdot)$ and $\phi_v(\cdot)$ are (possibly vector-valued) pseudo-Lipschitz continuous of order $p$ for some $p > 1$. Thus, the values of $\nu_u(t)$ and $\nu_v(t)$ depend on the outputs $v(t)$ and $u(t+1)$. Note that in equations (47) to (49), $d_i$, $u_{0i}$, $b_j$ and $v_0j$ are the components of the vectors $d$, $u_0$, $b$ and $v_0$, respectively. The algorithm is initialized with $t = 0$, $\xi_u(0) = 0$ and some vector $v(0)$.

Now, similar to Section III, consider a sequence of random realizations of the parameters indexed by the input dimension $n$. For each $n$, we assume that the output dimension $m = m(n)$ is deterministic and scales linearly as in (10) for some $\beta \geq 0$. Assume that the transform matrix $S$ has i.i.d. Gaussian components $s_{ij} \sim \mathcal{N}(0, 1/m)$. Also, assume that the empirical limits in (15) hold with bounded moments of order $2p-2$ for some limiting random variables $(U_0, U(t))$ and $(V_0, V(t))$. We will also assume the following continuity assumptions on $H_u(\cdot)$ and $H_v(\cdot)$:

Assumption 2: The function $H_u(t, b, u_0, \nu_u)$ satisfies the following continuity conditions:

(a) For every $\nu_u$ and $t$, $H_u(t, b, u_0, \nu_u)$ and its derivative $\nabla H_u(t, b, u_0, \nu_u)/\partial b$ are Lipschitz continuous in $b$ and $u_0$ for some Lipschitz constant that is continuous in $\nu_u$;

(b) For every $\nu_u$ and $t$, $H_u(t, b, u_0, \nu_u)$ and $\nabla H_u(t, b, u_0, \nu_u)/\partial b$ are continuous at $\nu_u$ uniformly over $(b, u_0)$.

The function $H_v(t, d, v_0, \nu_v)$ satisfies the analogous continuity assumptions in $d$, $v_0$ and $\nu_v$.

Under these assumption, we will show, as in Section III that for any fixed iteration $t$, the sets $\theta_u(t)$ and $\theta_v(t)$ in (14) converge empirically to the limits (16) for some random variable pairs $(U_0, U(t))$ and $(V_0, V(t))$. The random variable $U_0$ is identical to the variable in the limit (15a) and, for $t \geq 0$, $U(t)$ is given by

\[
U(t+1) = H_u(t, B(t), U(t), \nabla \theta_u(t)), \quad (50)
\]

\[
B(t) \sim \mathcal{N}(0, \tau^b(t)), \quad (51)
\]

for some deterministic constants $\nabla \theta_u(t)$ and $\tau^b(t)$ that will be defined below. Similarly, the random variable $V_0$ is identical to the variable in the limit (15b) and, for $t \geq 0$, $V(t)$ is given by

\[
V(t+1) = H_v(t, D(t), V_0, \nabla \theta_v(t)), \quad (52)
\]

\[
D(t) \sim \mathcal{N}(0, \tau^d(t)),
\]

for some constants $\nabla \theta_v(t)$ and $\tau^d(t)$, also defined below.

The constants $\tau^b(t)$, $\tau^d(t)$, $\nabla \theta_u(t)$ and $\nabla \theta_v(t)$ can be computed recursively through the following state evolution equations

\[
\tau^d(t) = \mathbb{E} [U(t+1)^2], \quad (53a)
\]

\[
\tau^b(t) = \beta \mathbb{E} [V(t)^2], \quad (53b)
\]

\[
\nabla \theta_u(t) = \mathbb{E} [\phi_u(t, V_0, V(t))], \quad (53c)
\]

\[
\nabla \theta_v(t) = \mathbb{E} [\phi_v(t, U_0, U(t+1))], \quad (53d)
\]

where the expectations are over the random variables $U(t)$ and $V(t)$ above and initialized with

\[
\tau^b(0) := \beta \mathbb{E} [V(0)^2]. \quad (54)
\]

With these definitions, we can now state the adaptive version of the result from Bayati and Montanari [14]

Theorem 4: Consider the recursion in (46) to (49). Then, under the above assumptions, for any fixed iteration number $t$, the sets $\theta_u(t)$ and $\theta_v(t)$ in (14) converge empirically to the limits (16) with bounded moments of order $p$ to the random variable pairs $(U_0, U(t))$ and $(V_0, V(t))$ described above.

Proof: We use an asterisk superscript to denote the outputs of the non-adaptive version of the recursions (46) to (49). That is, quantities such as $u^*(t), v^*(t), b^*(t), d^*(t), \ldots$, will represent the outputs generated by recursions (46) to (49) with the same initial conditions $(v(0), v(0))$ and $\xi_u(0) = \xi_v(0) = 0$, but in (47) and (48), $\nu_u(t)$ and $\nu_v(t)$ are replaced by their deterministic limits $\nabla \theta_u(t)$ and $\nabla \theta_v(t)$. Therefore, the

\[
u_u^*(t+1) = H_u(t, b^*_u(t), u_{0i}, \nabla \theta_u(t)), \quad (55a)
\]

\[
u_v^*(t+1) = H_v(t, d^*_v(t), v_{0j}, \nabla \theta_v(t+1)), \quad (55b)
\]

and

\[
\xi_u^*(t) = -\frac{1}{m} \sum_{i=1}^{m} \nabla \theta_u(t, b^*_u(t), u_{0i}, \nabla \theta_u(t)), \quad (56a)
\]

\[
\xi_v^*(t+1) = -\frac{1}{m} \sum_{j=1}^{m} \nabla \theta_v(t, d^*_v(t), v_{0j}, \nabla \theta_v(t)). \quad (56b)
\]

Now, Bayati and Montanari’s result in [14] shows that this non-adaptive algorithm satisfies the required properties. That is, the following limits hold with bounded moments of order $p$,

\[
\lim_{n \to \infty} \{u_{0i}(t), i = 1, \ldots, m\} = (U_0, U(t)) \quad (57a)
\]

\[
\lim_{n \to \infty} \{v_{0j}(v^*_j(t), j = 1, \ldots, n\} = (V_0, V(t)). \quad (57b)
\]

So, the limits (16) will be shown if we can prove the following limits hold almost surely for all $t$

\[
\lim_{n \to \infty} \frac{1}{m} \|u(t) - u^*(t)\|_p^p = 0 \quad (58a)
\]

\[
\lim_{n \to \infty} \frac{1}{m} \|v(t) - v^*(t)\|_p^p = 0, \quad (58b)
\]

where $\| \cdot \|_p$ is the $p$-norm. In the course of proving (58), we
will also show the following limits hold almost surely,

\[
\lim_{n \to \infty} \frac{1}{m} \| b(t) - b^*(t) \|_p = 0 \quad (59a)
\]

\[
\lim_{n \to \infty} \frac{1}{m} \| d(t) - d^*(t) \|_p = 0 \quad (59b)
\]

\[
\lim_{n \to \infty} \frac{1}{m} \| [\xi_u(t)] - [\xi_u^*(t)] \|_p = 0 \quad (59c)
\]

\[
\lim_{n \to \infty} \nu_u(t) = 0 \quad (59e)
\]

\[
\lim_{n \to \infty} \nu_v(t) = 0 \quad (59f)
\]

The proof of the limits (59a) and (59b) can be demonstrated via induction on \( t \) with the following straightforward (but somewhat tedious) continuity argument:

To begin the induction argument, first note that the non-adaptive algorithm has the same initial condition as the adaptive algorithm. That is, \( v^*(0) = v(0) \) and \( \xi_u(0) = \xi_u^*(0) = 0 \). Also, since \( \xi_u(0) = \xi_u^*(0) = 0 \), from (60a), the initial value of \( u(t) \) does not matter. So, without loss of generality, we can assume that the initial condition satisfies \( u(0) = u^*(0) \). Thus, the limits (58a), (58b) and (59e) hold for \( t = 0 \).

We now proceed by induction. Suppose that the limits (58a), (58b) and (59e) hold almost surely for some \( t \geq 0 \). Since \( S \) has i.i.d. components with zero mean and variance \( 1/m \), a standard result from random matrix theory (e.g. [40]) shows that its \( p \)-norm operator norm is bounded. That is, there exists a constant \( C_S > 0 \) such that

\[
\limsup_{n \to \infty} \| S \|_p \leq C_S, \quad \limsup_{n \to \infty} \| S^T \|_p \leq C_S. \quad (60)
\]

Substituting the bound (60) into (46a), we obtain

\[
\| b(t) - b^*(t) \|_p \leq \| S \|_p \| v(t) - v^*(t) \|_p + \| [\xi_u(t)] - [\xi_u^*(t)] \|_p \| u(t) - u^*(t) \|_p \leq C_S \| v(t) - v^*(t) \|_p + \| [\xi_u(t)] - [\xi_u^*(t)] \|_p \| u(t) - u^*(t) \|_p \quad (61)
\]

Now, since \( p \geq 1 \), we have that for any positive numbers \( a \) and \( b \),

\[
(a + b)^p \leq 2^{p-1}(a^p + b^p). \quad (62)
\]

Applying (62) into (61), and the fact that \( \lim_{n \to \infty} m/n = \beta \), we obtain that

\[
\frac{1}{m} \| b(t) - b^*(t) \|_p \leq C_S \frac{1}{n} \| v(t) - v^*(t) \|_p + \frac{\| [\xi_u(t)] - [\xi_u^*(t)] \|_p}{m} \| u(t) - u^*(t) \|_p + \frac{\| [\xi_u(t)] - [\xi_u^*(t)] \|_p}{m} \| u^*(t) \|_p. \quad (63)
\]

for some other constant \( C_S > 0 \). Now, since \( u^*(t) \) is the output of the non-adaptive algorithm it satisfies the limit

\[
\lim_{n \to \infty} \frac{1}{m} \| u^*(t) \|_p = \lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m} | u_i(t) |^p = \mathbb{E}[U(t)]^p < \infty. \quad (64)
\]

Substituting the bound (64) along with induction hypotheses, (58a), (58b) and (59e) into (63) shows (59a).

Next, to prove the limit (59e), first observe that since \( \phi_u(\cdot) \) is pseudo-Lipschitz continuous of order \( p \), we have that \( \mathbb{P}_u(t) \) in (53c) can be replaced by the limit of the empirical means

\[
\mathbb{P}_u(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \phi_u(t, v_{0j}, v_j^*(t)), \quad (65)
\]

where the limit holds almost surely. Combining (65) with the (49a),

\[
\limsup_{n \to \infty} \left| \nu_u(t) - \mathbb{P}_u(t) \right| \leq \limsup_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \left| \phi_u(t, v_{0j}, v_j^*(t)) - \phi_u(t, v_{0j}, v_j^*(t)) \right| \leq 1 \quad (66)
\]

Applying the fact that \( \phi_u(\cdot) \) is pseudo-Lipschitz continuous of order \( p \) to (66), we obtain that there exists a constant \( L_v > 0 \) such that

\[
\limsup_{n \to \infty} \left| \nu_u(t) - \mathbb{P}_u(t) \right| \leq \limsup_{n \to \infty} L_v \left[ \frac{1}{n} \| v(t) \|_p \right]^{(p-1)/p} \sum_{j=1}^{n} \left| \phi_u(t, v_{0j}, v_j^*(t)) \right| \left( \frac{1}{n} \| v(t) - v^*(t) \|_p \right)^{1/p} \quad (67)
\]

where the last step is due to Hölder’s inequality with the exponents

\[
\frac{p-1}{p} + 1 = 1.
\]

Now, similar to the proof of (64), one can show that the non-adaptive output satisfies the limit

\[
\lim_{n \to \infty} \frac{1}{n} \| v^*(t) \|_p = \mathbb{E}[V(t)]^p < \infty \quad (68)
\]

Also, from the induction hypothesis (58b), it follows that the non-adaptive output must satisfy the same limit

\[
\lim_{n \to \infty} \frac{1}{n} \| v(t) \|_p = \mathbb{E}[V(t)]^p < \infty. \quad (69)
\]

Applying the bounds (68) and (69) and the limit (67) shows that (59c) holds almost surely.

Now the limits (58b) and (59c) together with the continuity conditions on \( H_u(\cdot) \) in Assumption 2 show that (58a) holds almost surely for \( t + 1 \) and (59d) holds almost surely for \( t \). Using (66b), the proof of the limit (59a) is similar to the proof of (59a). These limits in turn show that the limits (58b) and (59c) hold almost surely for \( t + 1 \). We have thus shown that if (58a), (58b) and (59c) hold almost surely for some \( t \), they hold for \( t + 1 \). Thus, by induction they hold for all \( t \). Finally, applying the limits (57), (58a) and (58b) and a continuity argument shows that the desired limits (16a) hold almost surely.
APPENDIX D
PROOF OF THEOREM 1

The theorem directly follows from the adaptive Bayati–Montanari recursion theorem, Theorem 3 above, with some change of variables. Specifically, let
\[ S = \frac{1}{\sqrt{m}}W, \tag{70} \]
where \( W \) is the Gaussian noise in the rank one model in Assumption 1(b). Since \( W \) has i.i.d. components with Gaussian distributions \( N(0, \tau_w) \), the components of \( S \) will be i.i.d. with distributions \( N(0, 1/m) \).

Now, using the rank one model for \( A \) in \( \text{(1)} \)
\[ Av(t) = u_0v_0^T v(t) + \sqrt{m}Wv(t) = n\alpha_1(t)u_0 + \sqrt{m}Wv(t), \tag{71} \]
where the last step is from the definition of \( \alpha_1(t) \) in \( \text{(23)} \). Substituting \( \text{(71)} \) into the the update rule for \( p(t) \) in line 6 of Algorithm 1 we obtain
\[ p(t) = (1/m)A(v(t) + \mu_u(t)u(t)) = (1/\sqrt{m})Wv(t) + \beta\alpha_1(t)u_0 + \mu_u(t)u(t). \tag{72} \]
Note that we have used the fact that \( \beta = n/m \). Hence, if we define
\[ b(t) = \frac{1}{\sqrt{\tau_w}}(p(t) - \beta\alpha_1(t)u_0), \tag{73} \]
then \( \text{(70)} \) and \( \text{(72)} \) show that
\[ \xi_u(t) = \mu_u(t)/\sqrt{\tau_w}. \tag{74} \]
Similarly, one can show that if we define
\[ d(t) = \frac{1}{\sqrt{\tau_w}}(q(t) - \alpha_1(t)v_0), \tag{76} \]
then
\[ d(t) = S(t)^T u(t+1) + \xi_v(t)v(t), \tag{77} \]
where
\[ \xi_v(t) = \mu_v(t)/\sqrt{\tau_v}. \tag{78} \]
Next define the adaptation functions
\[ \phi_u(t, v_0, v) := (v_0, \phi_{v_0}(t, v_0, v)) \tag{79a} \]
\[ \phi_v(t, u_0, u) := (u_0, \phi_{u_0}(t, u_0, u)) \tag{79b} \]
which are the adaptation functions in \( \text{(13)} \) with additional components for the second-order statistics \( u_0 \) and \( v_0 \). Since \( \phi_{v_0}(t, \cdot) \) and \( \phi_{u_0}(t, \cdot) \) are pseudo-Lipschitz of order \( p \), so are \( \phi_u(t, \cdot) \) and \( \phi_v(t, \cdot) \). Taking the empirical means over each of the two components of \( \phi_u(\cdot) \) and \( \phi_v(\cdot) \), and applying \( \text{(13)} \) and \( \text{(23)} \), we see that if \( \nu_u(t) \) and \( \nu_v(t) \) are defined as in \( \text{(49)} \),
\[ \nu_u(t) = \frac{1}{n} \sum_{j=1}^n \phi_u(t, v_0, v_j(t)) = (\alpha_{v_1}(t), \lambda_u(t)) \tag{80a} \]
\[ \nu_v(t) = \frac{1}{m} \sum_{i=1}^m \phi_v(t, u_0, u_i(t+1)) = (\alpha_{u_1}(t+1), \lambda_v(t)) \tag{80b} \]
Therefore, \( \nu_u(t) \) and \( \nu_v(t) \) are vectors containing the parameters \( \lambda_u(t) \) and \( \lambda_v(t) \) for the factor selection functions in lines 7 and 11 of Algorithm 1 as well as the second-order statistics \( \alpha_{v_1}(t) \) and \( \alpha_{u_1}(t) \). Now, for \( \nu_u = (\alpha_{v_1}, \lambda_u) \) and \( \nu_v = (\alpha_{u_1}, \lambda_v) \) define the scalar functions
\[ H_u(t, b, u_0, \nu_u) := G_u(t, \sqrt{\tau_w}b + \beta\alpha_{v_1}u_0, \lambda_u), \tag{81a} \]
\[ H_v(t, d, v_0, \nu_v) := G_v(t, \sqrt{\tau_v}d + \alpha_{u_1}v_0, \lambda_v). \tag{81b} \]
Since \( G_u(t, p, \lambda_u) \) and \( G_v(t, q, \lambda_v) \) satisfy the continuity conditions in Assumption 1(c), \( H_u(t, b, u_0, \lambda_u) \) and \( H_v(t, d, v_0, \lambda_v) \) satisfy Assumption 2. In addition, the componentwise separability assumption in \( \text{(11)} \) implies that the updates in lines 7 and 11 of Algorithm 1 can be rewritten as
\[ u_i(t+1) = G_u(t, p_i(t), \lambda_u(t)) \tag{82a} \]
\[ v_j(t+1) = G_v(t, d_j(t), \lambda_v(t)). \tag{82b} \]
Thus, combining \( \text{(81)} \) and \( \text{(82)} \), with the definitions of \( b(t) \) and \( d(t) \) in \( \text{(73)} \) and \( \text{(76)} \), we obtain
\[ u_i(t+1) = H_u(t, b_i(t), u_0, \nu_u(t)) \tag{83a} \]
\[ v_j(t+1) = H_v(t, d_j(t), v_0, \nu_v(t)). \tag{83b} \]
Next observe that
\[ \xi_u(t) = \mu_u(t)/\sqrt{\tau_w}. \tag{74} \]
where (a) follows from the definition of \( \xi_u(t) \) in \( \text{(73)} \); (b) is the setting for \( \mu_u(t+1) \) in \( \text{(12a)} \); and (c) follows from the definition of \( H_v(t, d, \nu_v(t)) \) in \( \text{(81b)} \). Similarly, one can show that
\[ \xi_v(t) = -\frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial b_i} H_u(t, b_i(t), u_0, \nu_u(t)). \tag{85} \]
Equations \( \text{(74)}, \text{(77)}, \text{(80)}, \text{(81)}, \text{(84)} \) and \( \text{(85)} \) exactly match the recursions in equations \( \text{(46)} \) to \( \text{(49)} \). Therefore, Theorem 4 shows that the limits \( \text{(16)} \) hold in a sense that the sets \( \theta_u(t) \) and \( \theta_v(t) \) converge empirically with bounded moments of order \( p \).

We next show that the limits \( U(t) \) and \( V(t) \) on the righthand side of \( \text{(16)} \) match the descriptions in \( \text{(17a)} \) and \( \text{(18a)} \). First, define \( \alpha_{v_1}(t) \) and \( \alpha_{u_1}(t) \) in \( \text{(19)} \) and \( \lambda_u(t) \) and \( \lambda_v(t) \) as in \( \text{(20)} \). Then, from \( \text{(79)} \), the expectations \( \mathbb{E}_u(t) \) and \( \mathbb{E}_v(t) \) in \( \text{(83a)} \) and \( \text{(83b)} \) are given by
\[ \mathbb{E}(\nu_u(t)) = (\mathbb{E}_{\alpha_{v_1}(t)}, \mathbb{E}_{\lambda_u(t)}), \quad \mathbb{E}(\nu_v(t)) = (\mathbb{E}_{\alpha_{u_1}(t+1)}, \mathbb{E}_{\lambda_v(t)}). \tag{86} \]
Using \( \text{(51)}, \text{(81a)} \) and \( \text{(82a)} \), we see that
\[ U(t+1) = H_u(t, B(t), U_0, \nu_u(t)) \tag{87a} \]
\[ = G_u(t, \beta\alpha_{v_1}(t)U_0 + \sqrt{\tau_w}B(t), \lambda_u(t)). \tag{87b} \]
where \( B(t) \sim N(0, \tau_w^2(t)) \). Therefore, if we let \( Z_u(t) = \sqrt{\tau_w}B(t) \), then \( Z_u(t) \) is zero mean Gaussian with variance
\[ \mathbb{E}[Z_u^2(t)] = \tau_w^2(t) = \beta\tau_w \mathbb{E}[V(t)^2] = \beta\tau_w \mathbb{E}_0(t), \tag{88} \]
where follows from (23b) and (b) follows from the definition of \(\overline{\alpha}_0(t)\) in (19). Substituting \(Z_0(t) = \sqrt{\tau_w}B(t)\) into (87) we obtain the model for \(U(t+1)\) in (17a). Similarly, using (32) and (81b), we can obtain the model \(V(t+1)\) in (18a). Thus, we have proven that the random variables \((U_0, U(t))\) and \((V_0, V(t))\) are described by (17a) and (18a), and this completes the proof.

**APPENDIX E**

**PROOF OF THEOREM 3**

The theorem is proven by simply evaluating the second order statistics. We begin with \(\overline{\alpha}_1(t+1)\):

\[
\overline{\alpha}_{11}(t+1) = \left(\frac{(a) \overline{\alpha}(t)E[U_0U(t+1)] + (b) \overline{\alpha}_u(t)E[U_0P(t)]}{(c) \overline{\alpha}_u(t)E[U_0(\beta \overline{\alpha}_{11}(t)U_0 + Z_u(t))]} + (d) \overline{\alpha}_u(t)\beta \tau_u \overline{\alpha}_{11}(t)\right) \quad (88)
\]

where (a) is the definition in (19); (b) follows from (17a) and (23); (c) follows from (17b); and (d) follows from the independence of \(Z_u(t)\) and \(U_0\) and the definition of \(\tau_u\) in (23). Similarly, one can show that

\[
\overline{\alpha}_{00}(t+1) = \overline{\alpha}_u(t)^2 \left[\beta \tau_u \overline{\alpha}_{11}(t) + \beta \tau_w \overline{\alpha}_{00}(t)\right] \quad (89)
\]

Substituting (88) and (89) into (23), we obtain the asymptotic correlation coefficient

\[
\rho_u(t+1) = \frac{\overline{\alpha}_{11}(t)\beta^2 \tau_u^2 \overline{\alpha}_{11}(t)}{\overline{\alpha}_u(t) \left[\beta \tau_u \overline{\alpha}_{11}(t) + \beta \tau_w \overline{\alpha}_{00}(t)\right] \tau_u}
\]

\[
= \frac{\tau_u \tau_{\alpha_{11}}(t) + \tau_w \tau_{\alpha_{00}}(t)}{\beta \tau_u \rho_u(t)}
\]

\[
= \frac{\beta \tau_u \rho_u(t)}{\tau_u \rho_u(t) + \tau_w},
\]

where the last step follows from (27). This proves (29a).

A similar set of calculations shows that

\[
\overline{\alpha}_{11}(t+1) = \overline{\alpha}_u(t) \tau_u \overline{\alpha}_{11}(t+1) \quad (90a)
\]

\[
\overline{\alpha}_{00}(t+1) = \overline{\alpha}_u(t)^2 \left[\tau_u \overline{\alpha}_{11}(t+1) + \tau_w \overline{\alpha}_{00}(t)\right] \quad (90b)
\]

Applying these equations into (26) and (27), we obtain the recursion (29b). Hence, we have proven part (a) of the theorem.

For part (b), we need the following simple lemma.

**Lemma 3**: Suppose that \(H : [0, 1] \rightarrow [0, 1]\) is continuous and monotonically increasing, and \(x(t)\) is a sequence satisfying the recursive relation

\[
x(t+1) = H(x(t)),
\]

for some initial condition \(x(0) \in [0, 1]\). Then, either \(x(t)\) monotonically increases or decreases to some \(x^* = H(x^*)\).

**Proof**: This can be proven similar to [19] lemma 7.

To apply Lemma 3 observe that the recursions (29) show that

\[
\rho_u(t+1) = \frac{\beta \tau_u \rho_u(t)}{\beta \tau_u \rho_u(t) + \tau_w}
\]

\[
= \frac{\beta^2 \tau_u^2 \rho_u(t) + \tau_w}{\beta \tau_u \rho_u(t) + \tau_w},
\]

So, if we define

\[
H(\rho_u) := \frac{\beta \tau_u \rho_u(t)}{\beta \tau_u \rho_u(t) + \tau_w},
\]

then it follows from (29) that \(\rho_u(t+1) = H(\rho_u(t))\) for all \(t\). By taking the derivative, it can be checked that \(H(\rho_u)\) is monotonically increasing. It follows from Lemma 3 that \(\rho_u(t) \to \rho_u^*\) for some fixed point \(\rho_u^* = H(\rho_u^*)\) with \(\rho_u^* \in [0, 1]\).

Now, there are only two fixed point solutions to \(\rho_u = H(\rho_u)\): \(\rho_u^* = 0\) and

\[
\rho_u^* = \frac{-\beta \tau_u \rho_u \tau_u + \tau_w}{\tau_u \rho_u + \tau_w}.
\]

When

\[
\beta_0 \tau_u^2 \rho_u \tau_u \leq \tau_w^2,
\]

then \(\rho_u^*\) in (92) is not positive, so the only fixed solution in \([0, 1]\) is \(\rho_u^* = 0\). Therefore, when (92) is not satisfied, \(\rho_u(t)\) must converge to the zero fixed point: \(\rho_u(t) \to 0\).

Now, suppose that (94) is satisfied. In this case, we claim that \(\rho_u(t) \to \rho_u^*\) where \(\rho_u^*\) is in (92). We prove this claim by contradiction and suppose, instead, that \(\rho_u(t)\) converges to the other fixed point: \(\rho_u(t) \to 0\). Since Lemma 3 shows that \(\rho_u(t)\) must be either monotonically increasing or decreasing, the only way \(\rho_u(t) \to 0\) is that \(\rho_u(t)\) monotonically decreases to zero. But, when (94) is satisfied, it can be checked that for \(\rho_u(t)\) sufficiently small and positive, \(\rho_u(t+1) > H(\rho_u(t))\). This contradicts the fact that \(\rho_u(t)\) is monotonically decreasing, and therefore, \(\rho_u(t)\) must converge to the other fixed point \(\rho_u^*\) in (93).

Hence, we have shown that when (94) is not satisfied, \(\rho_u(t) \to 0\), and when (94) is satisfied \(\rho_u(t) \to \rho_u^*\) in (93). This is equivalent to the limit in (30a). The limit (30b) is proved similarly.

**APPENDIX F**

**PROOF OF THEOREM 4**

Similar to the proof of Theorem 3 we begin by computing the second-order statistics of \((U_0, U(t))\). Since \(U(t) = E[U_0|P(t-1)]\), \(U(t)\) must be uncorrelated with the error: \(U(t)(U_0 - U(t)) = 0\). Hence,

\[
\overline{\alpha}_{00}(t) - \overline{\alpha}_{11}(t) = E[U(t)U_0 - U(t)U(t)] = 0, \quad (95)
\]

and therefore \(\overline{\alpha}_{00}(t) = \overline{\alpha}_{11}(t)\). Now, consider the measurement \(P(t)\) in (77b). The SNR in this channel is

\[
\eta_u(t) = \frac{\beta^2 \tau_u \overline{\alpha}_{11}(t)}{\beta \tau_u \overline{\alpha}_{00}(t)} = \frac{\beta \tau_u \rho_u(t)}{\tau_u},
\]

Since \(U(t+1)\) is the conditional expectation of \(U_0\) given \(P(t)\), the mean-squared error is given by \(E_u(\eta_u(t))\) defined in (33a). Therefore,

\[
E_u(\eta_u(t)) = E[U(t+1) - U_0^2]
\]

\[
= \overline{\alpha}_{00}(t+1) - 2\overline{\alpha}_{11}(t+1) + \tau_u
\]

\[
= \tau_u - \overline{\alpha}_{00}(t+1),
\]

where (a) follows from expanding the square and substituting in the definitions in (19) and (23), and (b) follows from the
fact that $\pi_{u0}(t+1) = \pi_{u1}(t+1)$ proven above. We have thus proven that

$$\pi_{u0}(t+1) = \pi_{u1}(t+1) = \tau_u - \mathcal{E}_u(\eta_u(t)).$$  \hfill (98)

Therefore, the asymptotic correlation coefficient is given by

$$\rho_v(t+1) = \frac{\sqrt{\tau_u \pi_{u0}(t+1)}}{\tau_v \pi_{v0}(t+1)} \leq 1 - \tau_v^{-1} \mathcal{E}_v(\eta_v(t)), \hfill (99)$$

where (a) follows from (26) and (b) follows from (98). Substituting in (96) into (99) proves (84a). The recursion (84b) can be proven similarly.

For the initial condition in the recursion, observe that with $V(0) = E[V_0]$, the second order statistics are given by

$$\pi_{v0}(0) = E[V(0)^2] = (E[V_0])^2$$

$$\pi_{v1}(0) = E[V_0 V(0)] = (E[V_0])^2.$$  \hfill (100)

Hence, from (27), the initial correlation coefficient is

$$\rho_v(0) = \frac{\pi_{v2}(0)}{\tau_v \pi_{v0}(0)} = \frac{(E[V_0])^2}{\tau_v},$$

which agrees with the statement in the theorem. This proves part (a).

To prove part (b), we again use Lemma 3 Define the functions

$$H_u(\rho_v) := 1 - \tau_u^{-1} \mathcal{E}_u(\rho_v / \tau_v) \hfill (100a)$$

$$H_v(\rho_v) := 1 - \tau_v^{-1} \mathcal{E}_v(\rho_v / \tau_v), \hfill (100b)$$

and their concatenation

$$H(\rho_v) = H_v(H_u(\rho_v)).$$  \hfill (101)

From (34), it follows that $\rho_v(t+1) = H(\rho_v(t))$. Now, $\mathcal{E}_u(\eta_u)$ and $\mathcal{E}_v(\eta_v)$ defined in (33) are the mean-squared errors of $U_0$ and $V_0$ under AWGN estimation measurements with SNRs $\eta_u$ and $\eta_v$. Therefore, $\mathcal{E}_u(\eta_u)$ and $\mathcal{E}_v(\eta_v)$ must be monotonically decreasing in $\eta_u$ and $\eta_v$. Therefore, $H_u(\rho_v)$ and $H_v(\rho_v)$ in (100) are monotonically increasing functions and thus so is the concatenated function $H(\rho_v)$ in (101). Also, since the assumption of part (b) is that $\mathcal{E}_u(\eta_u)$ and $\mathcal{E}_v(\eta_v)$ are continuous, $H(\rho_v)$ is also continuous. It follows from Lemma 3 that $\rho_v(t) \to \rho_v^*$ where $\rho_v^*$ is a fixed point of (34).

It remains to show $\rho_v^* > 0$. Observe that

$$\mathcal{E}_v(\eta_v) \leq \text{Var}(V_0) = E[V_0^2] - (E[V_0])^2 = \tau_v(1 - \rho_v(0)),$$

where (a) follows from the definition of $\mathcal{E}_v(\eta_v)$ in (33b) and (b) follows from the definition of $\tau_v$ in (28) and the initial condition $\rho_v(0) = (E[V_0])^2 / \tau_v$. It follows from (34b) that

$$\rho_v(t+1) = 1 - \frac{1}{\tau_v} \mathcal{E}_v(\eta_v(t)) \geq \rho_v(0).$$

Therefore, the limit point $\rho_v^*$ of $\rho_v(t)$ must satisfy $\rho_v^* \geq \rho_v(0) > 0$.

### Appendix G

**Proof of Lemma 2**

Define the functions $H_u$, $H_v$ and $H$ as in (100) and (101) from the previous proof. We know that $\rho_v(t+1) = H(\rho_v(t))$. When $E[U_0] = E[V_0] = 0$, the $\rho_v = 0$ is a fixed point of the update. We can determine the stability of this fixed point by computing the derivative of $H(\rho_v)$ at $\rho_v = 0$.

Towards this end, we first use a standard result (see, for example, [35]) that, for any prior on $U_0$ with bounded second moments, the mean-squared error in (33a) satisfies

$$\mathcal{E}_u(\eta_u) = \frac{\tau_u}{1 + \eta_u \tau_u} + O(\eta_u^2).$$  \hfill (102)

The term $\tau_u/(1 + \eta_u \tau_u)$ is the minimum mean-squared error with linear estimation of $U_0$ from an AWGN noise-corrupted measurement $Y = \sqrt{\eta_u} U_0 + D, D \sim \mathcal{N}(0, 1)$. Equation (102) arises from the fact that linear estimation is optimal in low SNRs – see (35) for details. Using (102), we can compute the derivative of $H_u$,

$$H_u'(0) = -\frac{1}{\tau_u} \frac{\partial}{\partial \rho_v} \mathcal{E}_u(\beta \tau_v \rho_v / \tau_w) \bigg|_{\rho_v=0}$$

$$= -\beta \tau_v \mathcal{E}_v(0) = \frac{\beta \tau_v \tau_w}{\tau_w} \hfill (103)$$

Similarly one can show that

$$H_v'(0) = \frac{\tau_u \tau_v}{\tau_w} \hfill (104)$$

and hence

$$H'(0) = H_v'(0)H_u'(0) = \frac{\beta \tau_u^2 \tau_v^2}{\tau_w^2} \hfill (105)$$

We now apply a standard linearization analysis of the nonlinear system $\rho_v(t+1) = H(\rho_v(t))$ around the fixed point $\rho_v^* = 0$. See, for example, [41]. If $\sqrt{\beta} \tau_u \tau_v < \tau_w$ then $H'(0) < 1$ and the fixed point is stable. Thus, for any $\rho_v(0)$ sufficiently small $\rho_v(t) \to 0$. This proves part (b) of the lemma.

On the other hand, if $\sqrt{\beta} \tau_u \tau_v > \tau_w$ then $H'(0) > 1$ and the fixed point is unstable. This will imply that for any $\rho_v(0) > 0$, $\rho_v(t)$ will diverge from zero. But, we know from Theorem 3 that $\rho_v(t)$ must converge to some fixed point of $\rho_v = H(\rho_v)$. So, the limit point must be positive. This proves part (a) of the lemma.

### Acknowledgments

The authors thank Vivek Goyal, Ulugbek Kamilov, Andrea Montanari and Phil Schniter for detailed comments on an earlier draft.

### References

[1] Z. Ding, “Matrix outer-product decomposition method for blind multiple channel identification,” *IEEE Trans. Signal Process.*, vol. 45, no. 12, pp. 3053–3061, Dec. 1997.

[2] S. A. Jafar and A. Goldsmith, “Transmitter optimization and optimality of beamforming for multiple antenna systems,” *IEEE Trans. Wireless Comm.*, vol. 3, no. 4, pp. 1165–1175, Jul. 2004.

[3] T. Katayama, *Subspace Methods for System Identification*. New York: Springer, 2010.

[4] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.
[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge Univ. Press, 1985, reprinted with corrections 1987.

[6] J. Cadima and I. T. Jolliffe, “Loadings and correlations in the interpretation of principal components,” *J. Appl. Stat.*, vol. 22, pp. 208–214, 1995.

[7] B. A. Olshausen and D. J. Field, “Natural image statistics and efficient coding,” *Network Comp. Neural Syst.*, vol. 7, no. 2, pp. 333–339, 1996.

[8] ——, “Sparse coding with an overcomplete basis set: A strategy employed by v1?” *J. Vis. Res.*, vol. 37, pp. 3311–3325, 1997.

[9] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proc. Neural Information Process. Syst.*, Vancouver, Canada, Dec. 2001.

[10] M. S. Lewicki and T. J. Sejnowski, “Learning overcomplete representations,” *Neural Compu.*, vol. 12, pp. 337–365, 2000.

[11] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[12] J. T. Parker, V. Cevher, and P. Schniter, “Compressive sensing under matrix uncertainties: An approximate message passing approach,” in *Conf. Rec. Asilomar Conf. on Signals, Syst. & Computers*, Pacific Grove, CA, Nov. 2011.

[13] J. Mairal, F. Bach, G. Sapiro, and J. Ponce, “Online dictionary learning,” *Proc. CUR. IRML*, Montréal, Canada, Jun. 2009.

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

[15] J. Boutros and G. Caire, “Iterative multiuser joint decoding: unified framework and asymptotic analysis,” *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[16] A. d’Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, “A direct formulation for sparse PCA using semidefinite programming,” *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, Mar. 2007.

[17] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *J. Comput. Graphical Stat.*, vol. 12, pp. 531–547, 2003.

[18] Z. Zhang, H. Zha, and H. Simon, “Low rank approximations with sparse factors I: Basic algorithms and error analysis,” *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 3, pp. 706–727, 2002.

[19] Y. Wu and S. Verdú, “Functional Properties of Minimum Mean-Square Error and Mutual Information,” *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1289–1301, Mar. 2012.

[20] T. T. Jolliffe, N. Trendafilov, and M. Uddin, “A modified principal component technique based on the LASSO,” *J. Comput. Graphical Stat.*, vol. 12, pp. 331–347, 2003.

[21] J. Cadima and I. T. Jolliffe, “Loadings and correlations in the interpretation of principal components,” *J. Appl. Stat.*, vol. 22, pp. 208–214, 1995.