Bilinear Generalized Approximate Message Passing—Part I: Derivation

Jason T. Parker, *Member, IEEE*, Philip Schniter, *Fellow, IEEE*, and Volkan Cevher, *Senior Member, IEEE*

Abstract—**In this paper, we extend the generalized approximate message passing (G-AMP) approach, originally proposed for high-dimensional generalized-linear regression in the context of compressive sensing, to the generalized-bilinear case, which enables its application to matrix completion, robust PCA, dictionary learning, and related matrix-factorization problems. Here, in Part I of a two-part paper, we derive our Bilinear G-AMP (BiG-AMP) algorithm as an approximation of the sum-product belief propagation algorithm in the high-dimensional limit, where central-limit theorem arguments and Taylor-series approximations apply, and under the assumption of statistically independent matrix entries with known priors. In addition, we propose an adaptive damping mechanism that aids convergence under finite problem sizes, an expectation-maximization (EM)-based method to automatically tune the parameters of the assumed priors, and two rank-selection strategies. In Part II of the paper, we will discuss the specializations of EM-BiG-AMP to the problems of matrix completion, robust PCA, and dictionary learning, and we will present the results of an extensive empirical study comparing EM-BiG-AMP to state-of-the-art algorithms on each problem.**

Index Terms—**Approximate message passing, belief propagation, bilinear estimation, matrix completion, dictionary learning, robust principal components analysis, matrix factorization.**

I. Introduction

In this work, we present a new algorithmic framework for the following *generalized bilinear* inference problem: estimate the matrices $\boldsymbol{A} = [a_{mn}] \in \mathbb{R}^{M\times N}$ and $\boldsymbol{X} = [x_{nl}] \in \mathbb{R}^{N\times L}$ from a matrix observation $\boldsymbol{Y} \in \mathbb{R}^{M\times L}$ that is statistically coupled to their product, $\boldsymbol{Z} \triangleq \boldsymbol{AX}$. In doing so, we treat $\boldsymbol{A}$ and $\boldsymbol{X}$ as realizations of independent random matrices $\mathsf{\boldsymbol{A}}$ and $\mathsf{\boldsymbol{X}}$ with known separable pdfs (or pmfs in the case of discrete models)

$$p_{\mathsf{A}}(\boldsymbol{A}) = \prod_m \prod_n p_{\mathsf{a}_{mn}}(a_{mn}) \tag{1}$$

$$p_{\mathsf{X}}(\boldsymbol{X}) = \prod_n \prod_l p_{\mathsf{x}_{nl}}(x_{nl}), \tag{2}$$

and we likewise assume that the likelihood function of $\boldsymbol{Z}$ is known and separable, i.e.,

$$p_{\mathsf{Y}|\mathsf{Z}}(\boldsymbol{Y}|\boldsymbol{Z}) = \prod_m \prod_l p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z_{ml}). \tag{3}$$

Recently, various special cases of this problem have gained the intense interest of the research community, e.g.,

1) *Matrix Completion:* In this problem, one observes a few (possibly noise-corrupted) entries of a low-rank matrix and the goal is to infer the missing entries. In our framework, $\boldsymbol{Z} = \boldsymbol{AX}$ would represent the complete low-rank matrix (with tall $\boldsymbol{A}$ and wide $\boldsymbol{X}$) and $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}$ the observation mechanism, which would be (partially) informative about z_{ml} at the observed entries $(m,l) \in \Omega$ and non-informative at the missing entries $(m,l) \notin \Omega$.
2) *Robust PCA:* Here, the objective is to recover a low-rank matrix (or its principal components) observed in the presence of noise and sparse outliers. In our framework, $\boldsymbol{Z} = \boldsymbol{AX}$ could again represent the low-rank matrix, and $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}$ the noise-and-outlier-corrupted observation mechanism. Alternatively, $\boldsymbol{X}$ could also capture the outliers, as described in the sequel.
3) *Dictionary Learning:* Here, the objective is to learn a dictionary $\boldsymbol{A}$ for which there exists a sparse data representation $\boldsymbol{X}$ such that $\boldsymbol{AX}$ closely matches the observed data $\boldsymbol{Y}$. In our framework, $\{p_{\mathsf{x}_{nl}}\}$ would be chosen to induce sparsity, $\boldsymbol{Z} = \boldsymbol{AX}$ would represent the noiseless observations, and $\{p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}\}$ would model the (possibly noisy) observation mechanism.

While a plethora of approaches to these problems have been proposed based on optimization techniques (e.g., [5]–[15]), greedy methods (e.g., [16]–[20]), Bayesian sampling methods (e.g., [21], [22]), variational methods (e.g., [23]–[27]), and discrete message passing (e.g., [28]), ours is based on the *Approximate Message Passing* (AMP) framework, an instance of loopy belief propagation (LBP) [29] that was recently developed to tackle *linear* [30]–[32] and *generalized linear* [33] inference problems encountered in the context of compressive sensing (CS). In the generalized-linear CS problem, one estimates $\boldsymbol{x} \in \mathbb{R}^N$ from observations $\boldsymbol{y} \in \mathbb{R}^M$ that are statistically coupled to the transform outputs $\boldsymbol{z} = \boldsymbol{Ax}$ through a separable

Manuscript received December 09, 2013; revised June 05, 2014; accepted August 19, 2014. Date of publication September 12, 2014; date of current version October 22, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ivan W. Selesnick. The work of J. T. Parker on this project has been supported by AFOSR Lab Task 11RY02COR. The work of P. Schniter on this project has been supported by NSF grants IIP-0968910, CCF-1018368, CCF-1218754, and by DARPA/ONR grant N66001-10-1-4090. The work of V. Cevher was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633. Portions of this work were presented at the Workshop on Signal Processing with Adaptive Sparse Structured Representations, Edinburgh, Scotland, 2011 [2], the Information Theory and Applications Workshop, La Jolla, CA, USA, 2012 [3], and at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2012 [4].

J. T. Parker is with the Sensors Directorate, Air Force Research Laboratory, Dayton, OH 45433 USA (e-mail: jason.parker.13@us.af.mil).

P. Schniter is with The Ohio State University, Columbus OH 43210 USA (e-mail: schniter@ece.osu.edu).

V. Cevher is with the École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: volkan.cevher@epfl.ch).

Digital Object Identifier 10.1109/TSP.2014.2357776

likelihood function $p_{\mathbf{y}|\mathbf{z}}(\boldsymbol{y}|\boldsymbol{z})$, where in this case the transform $\boldsymbol{A}$ is *fixed and known*.

In the context of CS, the AMP framework yields algorithms with remarkable properties: i) solution trajectories that, in the large-system limit (i.e., as $M, N \rightarrow \infty$ with M/N fixed, under iid sub-Gaussian $\boldsymbol{A}$) are governed by a state-evolution whose fixed points—when unique—yield the true posterior means [34], [35] and ii) a low implementation complexity (i.e., dominated by one multiplication with $\boldsymbol{A}$ and $\boldsymbol{A}^\top$ per iteration, and relatively few iterations) [32]. Thus, a natural question is whether the AMP framework can be successfully applied to the generalized *bilinear* problem described earlier.

In this manuscript, which is Part I of a two-part work, we propose an AMP-based approach to generalized bilinear inference that we henceforth refer to as *Bilinear Generalized AMP* (BiG-AMP), and we uncover special cases under which the general approach can be simplified. In addition, we propose an adaptive damping [36] mechanism, an expectation-maximization (EM)-based [37] method of tuning the parameters of $p_{\mathsf{a}_{mn}}$, $p_{\mathsf{x}_{nl}}$, and $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}$ (in case they are unknown), and methods to select the rank N (in case it is unknown). In the case that $p_{\mathsf{a}_{mn}}$, $p_{\mathsf{x}_{nl}}$, and/or $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}$ are completely unknown, they can be modeled as Gaussian-mixtures with mean/variance/weight parameters learned via EM [38]. In Part II [1], we detail the application of BiG-AMP to matrix completion, robust PCA, and dictionary learning, and present the results of an extensive numerical investigation into the performance of BiG-AMP in each application. Those empirical results demonstrate that BiG-AMP yields an excellent combination of estimation accuracy and runtime when compared to existing state-of-the-art algorithms for each application.

Although the AMP methodology is itself restricted to separable known pdfs (1)–(3), the results of Part II suggest that this limitation is not an issue for many practical problems of interest. However, in problems where the separability assumption is too constraining, it can be relaxed through the use of hidden (coupling) variables, as originally proposed in the context of "turbo-AMP" [39] and applied to BiG-AMP in [40]. Due to space limitations, however, this approach will not be discussed here. Finally, although we focus on real-valued random variables, all of the methodology described in this work can be easily extended to circularly symmetric complex-valued random variables.

We now discuss related work. One possibility of applying AMP methods to matrix completion was suggested by Montanari in [32, Sec. 9.7.3] but the approach described there differs from BiG-AMP in that it was i) constructed from a factor graph with *vector-valued* variables and ii) restricted to the (incomplete) additive white Gaussian noise (AWGN) observation model. Moreover, no concrete algorithm nor performance evaluation was reported. Since we first reported on BiG-AMP in [2], [3], Rangan and Fletcher [41] proposed an AMP-based approach for the estimation of *rank-one* matrices from AWGN-corrupted observations, and showed that it can be characterized by a state evolution in the large-system limit. More recently, Krzakala, Mézard, and Zdeborová [42] proposed an AMP-based approach to blind calibration and dictionary learning in AWGN that bears similarity to a special case of BiG-AMP, and derived a state-evolution using the cavity method. Their method, however, was not numerically successful in solving dictionary learning problems [42]. The BiG-AMP algorithm that we derive here is a generalization of those in [41], [42] in that it handles *generalized* bilinear observations rather than AWGN-corrupted ones. Moreover, our work is the first to detail adaptive damping, parameter tuning, and rank-selection mechanisms for AMP based bilinear inference, and it is the first to present an in-depth numerical investigation involving both synthetic and real-world datasets. An application/extension of the BiG-AMP algorithm described here to hyperspectral unmixing (an instance of non-negative matrix factorization) was recently proposed in [40].

The remainder of the document is organized as follows. Section II derives the BiG-AMP algorithm, and Section III presents several special-case simplifications of BiG-AMP. Section IV describes the adaptive damping mechanism, and Section V the EM-based tuning of prior parameters and selection of rank N. Finally, Section VI concludes. In Part II [1], we detail the application of BiG-AMP to matrix completion, robust PCA, and dictionary learning, and present the results of an extensive numerical investigation into the performance of BiG-AMP in each application.

Notation: Throughout, we use san-serif font (e.g., x) for random variables and serif font (e.g., x) otherwise. We use boldface capital letters (e.g., $\boldsymbol{\mathsf{X}}$ and $\boldsymbol{X}$) for matrices, boldface small letters (e.g., $\boldsymbol{\mathsf{x}}$ and $\boldsymbol{x}$) for vectors, and non-bold small letters (e.g., x and x) for scalars. We then use $p_{\mathsf{x}}(x)$ to denote the pdf of random quantity x, and $\mathcal{N}(x; \widehat{x}, \nu^x)$ to denote the Gaussian pdf for a scalar random variable with mean $\widehat{x}$ and variance ν^x. Also, we use $\mathrm{E}\{\mathsf{x}\}$ and $\mathrm{var}\{\mathsf{x}\}$ to denote mean and variance of x, respectively, and $D(p_1 \| p_2)$ for the Kullback-Leibler (KL) divergence between pdfs p_1 and p_2. For a matrix $\boldsymbol{X}$, we use $x_{nl} = [\boldsymbol{X}]_{nl}$ to denote the entry in the n^{th} row and l^{th} column, $\|\boldsymbol{X}\|_F$ to denote the Frobenius norm, and $\boldsymbol{X}^\top$ to denote transpose. Similarly, we use x_n to denote the n^{th} entry in vector $\boldsymbol{x}$ and $\|\boldsymbol{x}\|_p = (\sum_n |x_n|^p)^{1/p}$ to denote the ℓ_p norm.

II. Bilinear Generalized AMP

A. Problem Formulation

For the statistical model (1)–(3), the posterior distribution is

$$\begin{aligned} & p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}}(\boldsymbol{X}, \boldsymbol{A} \mid \boldsymbol{Y}) \\ &= p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{A}) p_{\boldsymbol{\mathsf{X}}}(\boldsymbol{X}) p_{\boldsymbol{\mathsf{A}}}(\boldsymbol{A}) / p_{\boldsymbol{\mathsf{Y}}}(\boldsymbol{Y}) \qquad (4) \\ &\propto p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{Z}}}(\boldsymbol{Y} \mid \boldsymbol{A}\boldsymbol{X}) p_{\boldsymbol{\mathsf{X}}}(\boldsymbol{X}) p_{\boldsymbol{\mathsf{A}}}(\boldsymbol{A}) \qquad (5) \\ &= \Big[\prod_m \prod_l p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}} \Big(y_{ml} \Big| \sum_k a_{mk} x_{kl} \Big) \Big] \\ &\quad \times \Big[\prod_n \prod_l p_{\mathsf{x}_{nl}}(x_{nl}) \Big] \Big[\prod_m \prod_n p_{\mathsf{a}_{mn}}(a_{mn}) \Big], \qquad (6) \end{aligned}$$

where (4) employs Bayes' rule and $\propto$ denotes equality up to a constant scale factor.

The posterior distribution can be represented with a factor graph, as depicted in Fig. 1. There, the factors of $p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}}$ from (6) are represented by "factor nodes" that appear as black boxes, and the random variables are represented by "variable nodes"

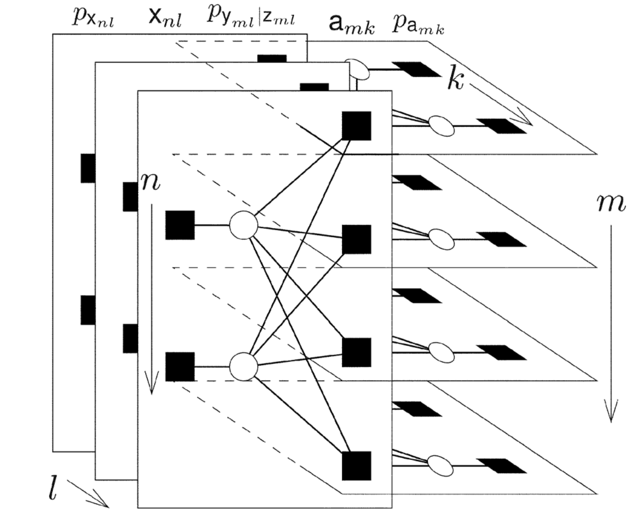

Fig. 1. The factor graph for generalized bilinear inference for (toy-sized) problem dimensions $M = 4$, $L = 3$, and $N = 2$.

that appear as white circles. Each variable node is connected to every factor node in which that variable appears. The observed data $\{y_{ml}\}$ are treated as parameters of the $p_{y_{ml}|z_{ml}}$ factor nodes in the middle of the graph, and not as random variables. The structure of Fig. 1 becomes intuitive when recalling that $\boldsymbol{Z} = \boldsymbol{AX}$ implies $z_{ml} = \sum_{n=1}^{N} a_{mn} x_{nl}$.

B. Loopy Belief Propagation

In this work, we aim to compute minimum mean-squared error (MMSE) estimates of $\boldsymbol{X}$ and $\boldsymbol{A}$, i.e., the means[1] of the marginal posteriors $p_{x_{nl}|\boldsymbol{Y}}(\cdot|\boldsymbol{Y})$ and $p_{a_{mn}|\boldsymbol{Y}}(\cdot|\boldsymbol{Y})$, for all pairs (n,l) and (m,n). Although exact computation of these quantities is generally prohibitive, they can be efficiently approximated using loopy belief propagation (LBP) [29].

In LBP, beliefs about the random variables (in the form of pdfs or log pdfs) are propagated among the nodes of the factor graph until they converge. The standard way to compute these beliefs, known as the *sum-product algorithm* (SPA) [43], [44], stipulates that the belief emitted by a variable node along a given edge of the graph is computed as the product of the incoming beliefs from all other edges, whereas the belief emitted by a factor node along a given edge is computed as the integral of the product of the factor associated with that node and the incoming beliefs on all other edges. The product of all beliefs impinging on a given variable node yields the posterior pdf for that variable. In cases where the factor graph has no loops, exact marginal posteriors result from two (i.e., forward and backward) passes of the SPA [43], [44]. For loopy factor graphs, exact inference is in general NP hard [45] and so LBP does not guarantee correct posteriors. That said, LBP has shown state-of-the-art performance in many applications, such as inference on Markov random fields [46], turbo decoding [47], LDPC decoding [48], multiuser detection [49], and compressive sensing [30], [31], [33]–[35].

In high-dimensional inference problems, exact implementation of the SPA is impractical, motivating approximations of the SPA. A notable example is the *generalized approximate message passing* (GAMP) algorithm, developed in [33] to solve the generalized CS problem, which exploits the "blessings of dimensionality" that arise when $\boldsymbol{A}$ is a sufficiently large and dense and which was rigorously analyzed in [35]. In the sequel, we derive an algorithm for the generalized bilinear inference BiG-AMP algorithm that employs GAMP-like approximations to the SPA on the factor graph in Fig. 1. As we shall see, the approximations are primarily based on central-limit-theorem (CLT) and Taylor-series arguments.

TABLE I
SPA MESSAGE DEFINITIONS AT ITERATION $t \in \mathbb{Z}$

$\Delta^{x}_{m\to nl}(t,.)$	SPA message from node $p_{y_{ml}|z_{ml}}$ to node x_{nl}
$\Delta^{x}_{m\leftarrow nl}(t,.)$	SPA message from node x_{nl} to node $p_{y_{ml}|z_{ml}}$
$\Delta^{a}_{l\to mn}(t,.)$	SPA message from node $p_{y_{ml}|z_{ml}}$ to node a_{mn}
$\Delta^{a}_{l\leftarrow mn}(t,.)$	SPA message from node a_{mn} to node $p_{y_{ml}|z_{ml}}$
$\Delta^{x}_{nl}(t,.)$	SPA-approximated log posterior pdf of x_{nl}
$\Delta^{a}_{mn}(t,.)$	SPA-approximated log posterior pdf of a_{mn}

C. Sum-Product Algorithm

In our formulation of the SPA, messages take the form of log-pdfs with arbitrary constant offsets. For example, the iteration-t (where $t \in \mathbb{Z}$) message $\Delta^{x}_{m\to nl}(t,.)$ can be converted to the pdf $\frac{1}{C}\exp(\Delta^{x}_{m\to nl}(t,.))$, where the choice of scale factor $C = \int_{x_{nl}} \exp(\Delta^{x}_{m\to nl}(t,x_{nl}))$ ensures that the pdf integrates to one. Four types of message will be used, as specified in Table I. We also find it convenient to express the (iteration-t SPA-approximated) posterior pdfs $p_{x_{nl}|\boldsymbol{Y}}(t,.|\boldsymbol{Y})$ and $p_{a_{mn}|\boldsymbol{Y}}(t,.|\boldsymbol{Y})$ in the log domain as $\Delta^{x}_{nl}(t,.)$ and $\Delta^{a}_{mn}(t,.)$, respectively, again with arbitrary constant offsets.

Applying the SPA to the factor graph in Fig. 1, we arrive at the following update rules for the four messages in Table I.

$$\begin{aligned}
&\Delta^{x}_{m\to nl}(t,x_{nl})\\
&= \log \int_{\{a_{mk}\}_{k=1}^{N},\{x_{rl}\}_{r\neq n}} p_{y_{ml}|z_{ml}}\left(y_{ml}\,\middle|\,\sum_{k=1}^{N} a_{mk}x_{kl}\right)\\
&\quad\times \prod_{r\neq n} \exp\left(\Delta^{x}_{m\leftarrow rl}(t,x_{rl})\right) \prod_{k=1}^{N} \exp\left(\Delta^{a}_{l\leftarrow mk}(t,a_{mk})\right)\\
&\quad+ \mathit{const} \qquad (7)
\end{aligned}$$

$$\begin{aligned}
&\Delta^{x}_{m\leftarrow nl}(t+1,x_{nl})\\
&= \log p_{x_{nl}}(x_{nl}) + \sum_{k\neq m} \Delta^{x}_{k\to nl}(t,x_{nl}) + \mathit{const} \qquad (8)
\end{aligned}$$

$$\begin{aligned}
&\Delta^{a}_{l\to mn}(t,a_{mn})\\
&= \log \int_{\{a_{mr}\}_{r\neq n},\{x_{kl}\}_{k=1}^{N}} p_{y_{ml}|z_{ml}}\left(y_{ml}\,\middle|\,\sum_{k=1}^{N} a_{mk}x_{kl}\right)\\
&\quad\times \prod_{k=1}^{N} \exp\left(\Delta^{x}_{m\leftarrow kl}(t,x_{kl})\right) \prod_{r\neq n} \exp\left(\Delta^{a}_{l\leftarrow mr}(t,a_{mr})\right)\\
&\quad+ \mathit{const} \qquad (9)
\end{aligned}$$

$$\begin{aligned}
&\Delta^{a}_{l\leftarrow mn}(t+1,a_{mn})\\
&= \log p_{a_{mn}}(a_{mn}) + \sum_{k\neq l} \Delta^{a}_{k\to mn}(t,a_{mn}) + \mathit{const}, \qquad (10)
\end{aligned}$$

[1] Another worthwhile objective could be to compute the joint MAP estimate $\arg\max_{\boldsymbol{X},\boldsymbol{A}} p_{\boldsymbol{X},\boldsymbol{A}|\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{A}|\boldsymbol{Y})$; we leave this to future work.

TABLE II
BIG-AMP VARIABLE SCALINGS IN THE LARGE-SYSTEM LIMIT

$\widehat{z}_{ml}(t)$	$O(1)$	$\nu^z_{ml}(t)$	$O(1)$	$\widehat{x}_{m,nl}(t) - \widehat{x}_{nl}(t)$	$O(\frac{1}{\sqrt{N}})$
$\widehat{x}_{m,nl}(t)$	$O(1)$	$\nu^x_{m,nl}(t)$	$O(1)$	$\widehat{x}^2_{m,nl}(t) - \widehat{x}^2_{nl}(t)$	$O(\frac{1}{\sqrt{N}})$
$\widehat{x}_{nl}(t)$	$O(1)$	$\nu^x_{nl}(t)$	$O(1)$	$\nu^x_{m,nl}(t) - \nu^x_{nl}(t)$	$O(\frac{1}{\sqrt{N}})$
$\widehat{a}_{l,mn}(t)$	$O(\frac{1}{\sqrt{N}})$	$\nu^a_{l,mn}(t)$	$O(\frac{1}{N})$	$\widehat{a}_{l,mn}(t) - \widehat{a}_{mn}(t)$	$O(\frac{1}{N})$
$\widehat{a}_{mn}(t)$	$O(\frac{1}{\sqrt{N}})$	$\nu^a_{mn}(t)$	$O(\frac{1}{N})$	$\widehat{a}^2_{l,mn}(t) - \widehat{a}^2_{mn}(t)$	$O(\frac{1}{N^{3/2}})$
$\widehat{p}_{ml}(t)$	$O(1)$	$\nu^p_{ml}(t)$	$O(1)$	$\nu^a_{l,mn}(t) - \nu^a_{mn}(t)$	$O(\frac{1}{N^{3/2}})$
$\widehat{r}_{m,nl}(t)$	$O(1)$	$\nu^r_{m,nl}(t)$	$O(1)$	$\widehat{r}_{m,nl}(t) - \widehat{r}_{nl}(t)$	$O(\frac{1}{\sqrt{N}})$
$\widehat{r}_{nl}(t)$	$O(1)$	$\nu^r_{nl}(t)$	$O(1)$	$\nu^r_{m,nl}(t) - \nu^r_{nl}(t)$	$O(\frac{1}{N})$
$\widehat{q}_{l,mn}(t)$	$O(\frac{1}{\sqrt{N}})$	$\nu^q_{l,mn}(t)$	$O(\frac{1}{N})$	$\widehat{q}_{l,mn}(t) - \widehat{q}_{mn}(t)$	$O(\frac{1}{N})$
$\widehat{q}_{mn}(t)$	$O(\frac{1}{\sqrt{N}})$	$\nu^q_{mn}(t)$	$O(\frac{1}{N})$	$\nu^q_{l,mn}(t) - \nu^q_{mn}(t)$	$O(\frac{1}{N^2})$
$\widehat{s}_{ml}(t)$	$O(1)$	$\nu^s_{ml}(t)$	$O(1)$		

where const is an arbitrary constant (w.r.t x_{nl} in (7) and (8), and w.r.t a_{mn} in (9) and (10)). In the sequel, we denote the mean and variance of the pdf $\frac{1}{C}\exp(\Delta^{\mathsf{x}}_{m\leftarrow nl}(t,.))$ by $\widehat{x}_{m,nl}(t)$ and $\nu^x_{m,nl}(t)$, respectively, and we denote the mean and variance of $\frac{1}{C}\exp(\Delta^{\mathsf{a}}_{l\leftarrow mn}(t,.))$ by $\widehat{a}_{l,mn}(t)$ and $\nu^a_{l,mn}(t)$. For the log-posteriors, the SPA implies

$$\begin{aligned}&\Delta^{\mathsf{x}}_{nl}(t+1,x_{nl})\\&\quad=\log p_{\mathsf{x}_{nl}}(x_{nl})+\sum_{m=1}^{M}\Delta^{\mathsf{x}}_{m\to nl}(t,x_{nl})+\mathit{const} &(11)\\&\Delta^{\mathsf{a}}_{mn}(t+1,a_{mn})\\&\quad=\log p_{\mathsf{a}_{mn}}(a_{mn})+\sum_{l=1}^{L}\Delta^{\mathsf{a}}_{l\to mn}(t,a_{mn})+\mathit{const}, &(12)\end{aligned}$$

and we denote the mean and variance of $\frac{1}{C}\exp(\Delta^{\mathsf{x}}_{nl}(t,.))$ by $\widehat{x}_{nl}(t)$ and $\nu^x_{nl}(t)$, and the mean and variance of $\frac{1}{C}\exp(\Delta^{\mathsf{a}}_{mn}(t,.))$ by $\widehat{a}_{mn}(t)$ and $\nu^a_{mn}(t)$.

D. Approximated Factor-to-Variable Messages

We now apply AMP approximations to the SPA updates (7)–(12). As we shall see, the approximations are based primarily on central-limit-theorem (CLT) and Taylor-series arguments that become exact in the large-system limit, where $M, L, N \to \infty$ with fixed ratios M/N and L/N. (Due to the use of finite M, L, N in practice, we still regard them as approximations.) In particular, our derivation will neglect terms that vanish relative to others as $N \to \infty$, which requires that we establish certain scaling conventions. First, we assume w.l.o.g[2] that $\mathrm{E}\{\mathsf{z}^2_{ml}\}$ and $\mathrm{E}\{\mathsf{x}^2_{nl}\}$ scale as $O(1)$, i.e., that the magnitudes of these elements stay finite as $N \to \infty$, and that $\mathrm{E}\{\mathsf{a}_{mn}\} = 0$. In this case, the relationship $\mathsf{z}_{ml} = \sum_{n=1}^{N} \mathsf{a}_{mn}\mathsf{x}_{nl}$ implies that $\mathrm{E}\{\mathsf{a}^2_{mn}\}$ must scale as $O(1/N)$. These scalings are assumed to hold for random variables z_{ml}, a_{mn}, and x_{ml} distributed according to the prior pdfs, according to the pdfs corresponding to the SPA messages (7)–(10), and according to the pdfs corresponding to the SPA posterior approximations (11)–(12). These assumptions lead straightforwardly to the scalings of $\widehat{z}_{ml}(t)$, $\nu^z_{ml}(t)$, $\widehat{x}_{m,nl}(t)$, $\nu^x_{m,nl}(t)$, $\widehat{x}_{nl}(t)$, $\nu^x_{nl}(t)$, $\widehat{a}_{l,mn}(t)$, $\nu^a_{l,mn}(t)$, $\widehat{a}_{mn}(t)$, and $\nu^a_{mn}(t)$ specified in Table II. Furthermore, because $\Delta^{\mathsf{x}}_{m\to nl}(t,\cdot)$ and $\Delta^{\mathsf{x}}_{nl}(t,\cdot)$ differ by only one term out of M, it is reasonable to assume [32], [33] that the corresponding difference in means $\widehat{x}_{m,nl}(t) - \widehat{x}_{nl}(t)$ and variances $\nu^x_{m,nl}(t) - \nu^x_{nl}(t)$ are both $O(1/\sqrt{N})$, which then implies that $\widehat{x}^2_{m,nl}(t) - \widehat{x}^2_{nl}(t)$ is also $O(1/\sqrt{N})$. Similarly, because $\Delta^{\mathsf{a}}_{l\to mn}(t,\cdot)$ and $\Delta^{\mathsf{a}}_{mn}(t,\cdot)$ differ by only one term out of N, where $\widehat{a}_{l,mn}(t)$ and $\widehat{a}_{mn}(t)$ are $O(1/\sqrt{N})$, it is reasonable to assume that $\widehat{a}_{l,mn}(t) - \widehat{a}_{mn}(t)$ is $O(1/N)$ and that both $\nu^a_{l,mn}(t) - \nu^a_{mn}(t)$ and $\widehat{a}^2_{l,mn}(t) - \widehat{a}^2_{mn}(t)$ are $O(1/N^{3/2})$. The remaining entries in Table II will be explained below.

We start by approximating the message $\Delta^{\mathsf{x}}_{m\to nl}(t,.)$. Expanding (7), we find

$$\begin{aligned}&\Delta^{\mathsf{x}}_{m\to nl}(t,x_{nl})\\&=\log\int_{\{a_{mk}\}_{k=1}^{N},\{x_{rl}\}_{r\neq n}} p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}\Big(y_{ml}\Big|\overbrace{a_{mn}x_{nl}+\sum_{k=1\neq n}^{N}a_{mk}x_{kl}}^{z_{ml}}\Big)\\&\quad\times\prod_{r\neq n}\exp\left(\Delta^{\mathsf{x}}_{m\leftarrow rl}(t,x_{rl})\right)\prod_{k=1}^{N}\exp\left(\Delta^{\mathsf{a}}_{l\leftarrow mk}(t,a_{mk})\right)\\&\quad+\mathit{const}. &(13)\end{aligned}$$

For large N, the CLT motivates the treatment of z_{ml}, the random variable associated with the z_{ml} identified in (13), conditioned on $\mathsf{x}_{nl} = x_{nl}$, as Gaussian and thus completely characterized by a (conditional) mean and variance. Defining the zero-mean r.v.s $\widetilde{\mathsf{a}}_{l,mn} \triangleq \mathsf{a}_{mn} - \widehat{a}_{l,mn}(t)$ and $\widetilde{\mathsf{x}}_{m,nl} = \mathsf{x}_{nl} - \widehat{x}_{m,nl}(t)$, where $\mathsf{a}_{mn} \sim \frac{1}{C}\exp(\Delta^{\mathsf{a}}_{l\leftarrow mn}(t,\cdot))$ and $\mathsf{x}_{nl} \sim \frac{1}{C}\exp(\Delta^{\mathsf{x}}_{m\leftarrow nl}(t,\cdot))$, we can write

$$\begin{aligned}\mathsf{z}_{ml}=\big(\widehat{a}_{l,mn}(t)+\widetilde{\mathsf{a}}_{l,mn}\big)\mathsf{x}_{nl}+\sum_{k\neq n}\big(\widehat{a}_{l,mk}(t)\widehat{x}_{m,kl}(t)\\+\widehat{a}_{l,mk}(t)\widetilde{\mathsf{x}}_{m,kl}+\widetilde{\mathsf{a}}_{l,mk}\widehat{x}_{m,kl}(t)+\widetilde{\mathsf{a}}_{l,mk}\widetilde{\mathsf{x}}_{m,kl}\big) \quad (14)\end{aligned}$$

after which it is straightforward to see that

$$\mathrm{E}\{\mathsf{z}_{ml}|\mathsf{x}_{nl}=x_{nl}\}=\widehat{a}_{l,mn}(t)x_{nl}+\widehat{p}_{n,ml}(t) \quad (15)$$

$$\mathrm{var}\{\mathsf{z}_{ml}|\mathsf{x}_{nl}=x_{nl}\}=\nu^a_{l,mn}(t)x^2_{nl}+\nu^p_{n,ml}(t) \quad (16)$$

for

$$\widehat{p}_{n,ml}(t)\triangleq\sum_{k\neq n}\widehat{a}_{l,mk}(t)\widehat{x}_{m,kl}(t) \quad (17)$$

$$\begin{aligned}\nu^p_{n,ml}(t)\triangleq\sum_{k\neq n}\big(\widehat{a}^2_{l,mk}(t)\nu^x_{m,kl}(t)+\nu^a_{l,mk}(t)\widehat{x}^2_{m,kl}(t)\\+\nu^a_{l,mk}(t)\nu^x_{m,kl}(t)\big). \quad (18)\end{aligned}$$

With this conditional-Gaussian approximation, (13) becomes

$$\begin{aligned}&\Delta^{\mathsf{x}}_{m\to nl}(t,x_{nl})\approx\mathit{const}+\log\int_{z_{ml}}p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z_{ml}) &(19)\\&\quad\times\mathcal{N}\big(z_{ml};\widehat{a}_{l,mn}(t)x_{nl}+\widehat{p}_{n,ml}(t),\nu^a_{l,mn}(t)x^2_{nl}+\nu^p_{n,ml}(t)\big)\\&=H_{ml}\Big(\widehat{a}_{l,mn}(t)x_{nl}+\widehat{p}_{n,ml}(t),\\&\qquad\nu^a_{l,mn}(t)x^2_{nl}+\nu^p_{n,ml}(t);y_{ml}\Big)+\mathit{const} &(20)\end{aligned}$$

in terms of the function

$$H_{ml}\big(\widehat{q},\nu^q;y\big)\triangleq\log\int_z p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y|z)\mathcal{N}(z;\widehat{q},\nu^q). \quad (21)$$

[2] Other scalings on $\mathrm{E}\{\mathsf{z}^2_{ml}\}$, $\mathrm{E}\{\mathsf{x}^2_{nl}\}$, and $\mathrm{E}\{\mathsf{a}^2_{mn}\}$ could be used as long as they are consistent with the relationship $\mathsf{z}_{ml} = \sum_{n=1}^{N} \mathsf{a}_{mn}\mathsf{x}_{nl}$.

Unlike the original SPA message (7), the approximation (20) requires only a single integration. Still, additional simplifications are possible. First, notice that $\widehat{p}_{n,ml}(t)$ and $\nu^p_{n,ml}(t)$ differ from the corresponding n-invariant quantities

$$\widehat{p}_{ml}(t) \triangleq \sum_{k=1}^{N} \widehat{a}_{l,mk}(t)\widehat{x}_{m,kl}(t) \tag{22}$$

$$\nu^p_{ml}(t) \triangleq \sum_{k=1}^{N} \big(\widehat{a}^2_{l,mk}(t)\nu^x_{m,kl}(t) + \nu^a_{l,mk}(t)\widehat{x}^2_{m,kl}(t) + \nu^a_{l,mk}(t)\nu^x_{m,kl}(t)\big) \tag{23}$$

by one term. In the sequel, we will assume that $\widehat{p}_{ml}(t)$ and $\nu^p_{ml}(t)$ are $O(1)$ since these quantities can be recognized as the mean and variance, respectively, of an estimate of z_{ml}, which is $O(1)$. Writing the H_{ml} term in (20) using (22)–(23),

$$\begin{aligned} &H_{ml}\Big(\widehat{a}_{l,mn}(t)x_{nl} + \widehat{p}_{n,ml}(t), \nu^a_{l,mn}(t)x^2_{nl} + \nu^p_{n,ml}(t); y_{ml}\Big)\\ &= H_{ml}\Big(\widehat{a}_{l,mn}(t)\big(x_{nl} - \widehat{x}_{m,nl}(t)\big) + \widehat{p}_{ml}(t),\\ &\qquad \nu^a_{l,mn}(t)\big(x^2_{nl} - \widehat{x}^2_{m,nl}(t)\big) - \widehat{a}^2_{l,mn}(t)\nu^x_{m,nl}(t)\\ &\qquad - \nu^a_{l,mn}(t)\nu^x_{m,nl}(t) + \nu^p_{ml}(t); y_{ml}\Big) \qquad (24)\\ &= H_{ml}\Big(\widehat{a}_{l,mn}(t)\big(x_{nl} - \widehat{x}_{nl}(t)\big) + \widehat{p}_{ml}(t) + O(1/N),\\ &\qquad \nu^a_{l,mn}(t)\big(x^2_{nl} - \widehat{x}^2_{nl}(t)\big) + \nu^p_{ml}(t) + O(1/N); y_{ml}\Big) \qquad (25) \end{aligned}$$

where in (25) we used the facts that $\widehat{a}_{l,mn}(t)(\widehat{x}_{nl}(t) - \widehat{x}_{m,nl}(t))$ and $\nu^a_{l,mn}(t)(\widehat{x}^2_{m,nl}(t) - \widehat{x}^2_{nl}(t)) - \widehat{a}^2_{l,mn}(t)\nu^x_{m,nl}(t) - \nu^a_{l,mn}(t)\nu^x_{m,nl}(t)$ are both $O(1/N)$.

Rewriting (20) using a Taylor series expansion in x_{nl} about the point $\widehat{x}_{nl}(t)$, we get

$$\begin{aligned} &\Delta^x_{m\to nl}(t, x_{nl}) \approx \mathit{const}\\ &+ H_{ml}\big(\widehat{p}_{ml}(t) + O(1/N), \nu^p_{ml}(t) + O(1/N); y_{ml}\big)\\ &+ \widehat{a}_{l,mn}(t)\big(x_{nl} - \widehat{x}_{nl}(t)\big)\\ &\qquad \times H'_{ml}\big(\widehat{p}_{ml}(t) + O(1/N), \nu^p_{ml}(t) + O(1/N); y_{ml}\big)\\ &+ 2\nu^a_{l,mn}(t)\widehat{x}_{nl}(t)\big(x_{nl} - \widehat{x}_{nl}(t)\big)\\ &\qquad \times \dot{H}_{ml}\big(\widehat{p}_{ml}(t) + O(1/N), \nu^p_{ml}(t) + O(1/N); y_{ml}\big)\\ &+ \nu^a_{l,mn}(t)\big(x_{nl} - \widehat{x}_{nl}(t)\big)^2\\ &\qquad \times \dot{H}_{ml}\big(\widehat{p}_{ml}(t) + O(1/N), \nu^p_{ml}(t) + O(1/N); y_{ml}\big)\\ &+ \frac{1}{2}\widehat{a}^2_{l,mn}(t)\big(x_{nl} - \widehat{x}_{nl}(t)\big)^2\\ &\qquad \times H''_{ml}\big(\widehat{p}_{ml}(t) + O(1/N), \nu^p_{ml}(t) + O(1/N); y_{ml}\big)\\ &+ O(1/N^{3/2}), \end{aligned} \tag{26}$$

where H'_{ml} and H''_{mn} are the first two derivatives of H_{mn} w.r.t its first argument and $\dot{H}_{ml}$ is the first derivative w.r.t its second argument. Note that, in (26) and elsewhere, the higher-order terms in the Taylor's expansion are written solely in terms of their scaling dependence on N, which is what will eventually allow us to neglect these terms (in the large-system limit).

We now approximate (26) by dropping terms that vanish, relative to the second-to-last term in (26), as $N \to \infty$. Since this second-to-last term is $O(1/N)$ due to the scalings of $\widehat{a}^2_{l,mn}(t)$, $\widehat{p}_{ml}(t)$, and $\nu^p_{ml}(t)$, we drop terms that are of order $O(1/N^{3/2})$, such as the final term. We also replace $\nu^a_{l,mn}(t)$ with $\nu^a_{mn}(t)$, and $\widehat{a}^2_{l,mn}(t)$ with $\widehat{a}^2_{mn}(t)$, since in both cases the difference is $O(1/N^{3/2})$. Finally, we drop the $O(1/N)$ terms inside the H_{ml} derivatives, which can be justified by taking a Taylor series expansion of these derivatives with respect to the $O(1/N)$ perturbations and verifying that the higher-order terms in this latter expansion are $O(1/N^{3/2})$. All of these approximations are analogous to those made in previous AMP derivations, e.g., [31], [32], and [33].

Applying these approximations to (26) and absorbing x_{nl}-invariant terms into the *const* term, we obtain

$$\begin{aligned} \Delta^x_{m\to nl}(t, x_{nl}) \approx\ &\big[\widehat{s}_{ml}(t)\widehat{a}_{l,mn}(t) + \nu^s_{ml}(t)\widehat{a}^2_{mn}(t)\widehat{x}_{nl}(t)\big]\\ &\times x_{nl} - \frac{1}{2}\big[\nu^s_{ml}(t)\widehat{a}^2_{mn}(t) - \nu^a_{mn}(t)\\ &\times \big(\widehat{s}^2_{ml}(t) - \nu^s_{ml}(t)\big)\big]x^2_{nl} + \mathit{const}, \end{aligned} \tag{27}$$

where we used the relationship

$$\dot{H}_{ml}\big(\widehat{q}, \nu^q; y\big) = \frac{1}{2}\Big[\big(H'_{ml}\big(\widehat{q}, \nu^q; y\big)\big)^2 + H''_{ml}\big(\widehat{q}, \nu^q; y\big)\Big] \tag{28}$$

and defined

$$\widehat{s}_{ml}(t) \triangleq H'_{ml}\big(\widehat{p}_{ml}(t), \nu^p_{ml}(t); y_{ml}\big) \tag{29}$$

$$\nu^s_{ml}(t) \triangleq -H''_{ml}\big(\widehat{p}_{ml}(t), \nu^p_{ml}(t); y_{ml}\big). \tag{30}$$

Note that (27) is essentially a Gaussian approximation to the pdf $\frac{1}{C}\exp(\Delta^x_{m\to nl}(t, .))$.

We show in Appendix A that

$$\widehat{s}_{ml}(t) = \frac{1}{\nu^p_{ml}(t)}\big(\widehat{z}_{ml}(t) - \widehat{p}_{ml}(t)\big) \tag{31}$$

$$\nu^s_{ml}(t) = \frac{1}{\nu^p_{ml}(t)}\left(1 - \frac{\nu^z_{ml}(t)}{\nu^p_{ml}(t)}\right), \tag{32}$$

for the conditional mean and variance

$$\widehat{z}_{ml}(t) \triangleq \mathrm{E}\{\mathsf{z}_{ml}|\mathsf{p}_{ml} = \widehat{p}_{ml}(t); \nu^p_{ml}(t)\} \tag{33}$$

$$\nu^z_{ml}(t) \triangleq \mathrm{var}\{\mathsf{z}_{ml}|\mathsf{p}_{ml} = \widehat{p}_{ml}(t); \nu^p_{ml}(t)\}, \tag{34}$$

computed according to the (conditional) pdf

$$\begin{aligned} &p_{\mathsf{z}_{ml}|\mathsf{p}_{ml}}\big(z_{ml}|\widehat{p}_{ml}(t); \nu^p_{ml}(t)\big)\\ &\quad \triangleq \frac{1}{C}p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z_{ml})\mathcal{N}\big(z_{ml}; \widehat{p}_{ml}(t), \nu^p_{ml}(t)\big), \end{aligned} \tag{35}$$

where here $C = \int_z p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z)\mathcal{N}\big(z; \widehat{p}_{ml}(t), \nu^p_{ml}(t)\big)$. In fact, (35) is BiG-AMP's iteration-t approximation to the true marginal posterior $p_{\mathsf{z}_{ml}|\boldsymbol{\mathsf{Y}}}(\cdot|\boldsymbol{Y})$. We note that (35) can also be interpreted as the (exact) posterior pdf for z_{ml} given the likelihood $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|\cdot)$ from (3) and the prior $\mathsf{z}_{ml} \sim \mathcal{N}\big(\widehat{p}_{ml}(t), \nu^p_{ml}(t)\big)$ that is implicitly assumed by iteration-t BiG-AMP.

Since $\boldsymbol{Z}^\top = \boldsymbol{X}^\top \boldsymbol{A}^\top$, the derivation of the BiG-AMP approximation of $\Delta^{\mathsf{a}}_{l\to mn}(t,.)$ closely follows the derivation for $\Delta^{\mathsf{x}}_{m\to nl}(t,.)$. In particular, it starts with (similar to (13))

$$\begin{aligned}
&\Delta^{\mathsf{a}}_{l\to mn}(t,a_{mn})\\
&= \log \int_{\{a_{mk}\}_{k\neq n},\{x_{rl}\}_{r=1}^{N}} p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}\Big(y_{ml}\Big|\overbrace{a_{mn}x_{nl}+\sum_{k\neq n}a_{mk}x_{kl}}^{z_{ml}}\Big)\\
&\quad\times \prod_{r=1}^{N}\exp\Big(\Delta^{\mathsf{x}}_{m\leftarrow rl}(t,x_{rl})\Big)\prod_{k\neq n}\exp\Big(\Delta^{\mathsf{a}}_{l\leftarrow mk}(t,a_{mk})\Big)\\
&\quad+\mathit{const}, \qquad (36)
\end{aligned}$$

where again the CLT motivates the treatment of z_{ml}, conditioned on $\mathsf{a}_{mn} = a_{mn}$, as Gaussian. Eventually we arrive at the Taylor-series approximation (similar to (27))

$$\begin{aligned}
\Delta^{\mathsf{a}}_{l\to mn}(t,a_{mn}) \approx{}& \Big[\widehat{s}_{ml}(t)\widehat{x}_{m,nl}(t)+\nu^{s}_{ml}(t)\widehat{x}^{2}_{nl}(t)\widehat{a}_{mn}(t)\Big]\\
&\times a_{mn}-\frac{1}{2}\big[\nu^{s}_{ml}(t)\widehat{x}^{2}_{nl}(t)-\nu^{x}_{nl}(t)\\
&\times\big(\widehat{s}^{2}_{ml}(t)-\nu^{s}_{ml}(t)\big)\big]a^{2}_{mn}\\
&+\mathit{const}. \qquad (37)
\end{aligned}$$

E. Approximated Variable-to-Factor Messages

We now turn to approximating the messages flowing from the variable nodes to the factor nodes. Starting with (8) and plugging in (27) we obtain

$$\begin{aligned}
&\Delta^{\mathsf{x}}_{m\leftarrow nl}(t+1,x_{nl})\\
&\approx \mathit{const}+\log p_{\mathsf{x}_{nl}}(x_{nl})+\sum_{k\neq m}\Bigg(\Big[\widehat{s}_{kl}(t)\widehat{a}_{l,kn}(t)\\
&\quad+\nu^{s}_{kl}(t)\widehat{a}^{2}_{kn}(t)\widehat{x}_{nl}(t)\Big]x_{nl}-\frac{1}{2}\Big[\nu^{s}_{kl}(t)\widehat{a}^{2}_{kn}(t)\\
&\quad-\nu^{a}_{kn}(t)\big(\widehat{s}^{2}_{kl}(t)-\nu^{s}_{kl}(t)\big)\Big]x^{2}_{nl}\Bigg) \qquad (38)\\
&= \mathit{const}+\log p_{\mathsf{x}_{nl}}(x_{nl})-\frac{1}{2\nu^{r}_{m,nl}(t)}\big(x_{nl}-\widehat{r}_{m,nl}(t)\big)^{2} \qquad (39)\\
&= \mathit{const}+\log\Big(p_{\mathsf{x}_{nl}}(x_{nl})\mathcal{N}\big(x_{nl};\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t)\big)\Big) \qquad (40)
\end{aligned}$$

for

$$\nu^{r}_{m,nl}(t)\triangleq\Bigg(\sum_{k\neq m}\widehat{a}^{2}_{kn}(t)\nu^{s}_{kl}(t)-\nu^{a}_{kn}(t)\big(\widehat{s}^{2}_{kl}(t)-\nu^{s}_{kl}(t)\big)\Bigg)^{-1} \qquad (41)$$

$$\begin{aligned}
\widehat{r}_{m,nl}(t)\triangleq{}&\widehat{x}_{nl}(t)\Big(1+\nu^{r}_{m,nl}(t)\sum_{k\neq m}\nu^{a}_{kn}(t)[\widehat{s}^{2}_{kl}(t)-\nu^{s}_{kl}(t)]\Big)\\
&+\nu^{r}_{m,nl}(t)\sum_{k\neq m}\widehat{a}_{l,kn}(t)\widehat{s}_{kl}(t). \qquad (42)
\end{aligned}$$

Since $\widehat{a}^{2}_{mn}(t)$ and $\nu^{a}_{mn}(t)$ are $O(1/N)$, and recalling $\widehat{s}^{2}_{ml}(t)$ and $\nu^{s}_{ml}(t)$ are $O(1)$, we take $\nu^{r}_{m,nl}(t)$ to be $O(1)$. Meanwhile, since $\widehat{r}_{m,nl}(t)$ is an estimate of x_{nl}, we reason that it is $O(1)$.

The mean and variance of the pdf associated with the $\Delta^{\mathsf{x}}_{m\leftarrow nl}(t+1,.)$ approximation in (40) are

$$\begin{aligned}
&\widehat{x}_{m,nl}(t+1)\\
&\triangleq\underbrace{\frac{1}{C}\int_{x}x\,p_{\mathsf{x}_{nl}}(x)\mathcal{N}\big(x;\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t)\big)}_{\triangleq\, g_{\mathsf{x}_{nl}}(\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t))} \qquad (43)\\
&\nu^{x}_{m,nl}(t+1)\\
&\triangleq\underbrace{\frac{1}{C}\int_{x}\big|x-\widehat{x}_{m,nl}(t+1)\big|^{2}p_{\mathsf{x}_{nl}}(x)\mathcal{N}\big(x;\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t)\big)}_{\nu^{r}_{m,nl}(t)g'_{\mathsf{x}_{nl}}(\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t))}\\
&\qquad (44)
\end{aligned}$$

where here $C=\int_{x}p_{\mathsf{x}_{nl}}(x)\mathcal{N}\big(x;\widehat{r}_{m,nl}(t),\nu^{r}_{m,nl}(t)\big)$ and $g'_{\mathsf{x}_{nl}}$ denotes the derivative of $g_{\mathsf{x}_{nl}}$ with respect to the first argument. The fact that (43) and (44) are related through a derivative was shown in [33].

We now derive approximations of $\widehat{x}_{m,nl}(t)$ and $\nu^{x}_{m,nl}(t)$ that avoid the dependence on the destination node m. For this, we introduce m-invariant versions of $\widehat{r}_{m,nl}(t)$ and $\nu^{r}_{m,nl}(t)$:

$$\nu^{r}_{nl}(t)\triangleq\Bigg(\sum_{m=1}^{M}\widehat{a}^{2}_{mn}(t)\nu^{s}_{ml}(t)-\nu^{a}_{mn}(t)\big(\widehat{s}^{2}_{ml}(t)-\nu^{s}_{ml}(t)\big)\Bigg)^{-1} \qquad (45)$$

$$\begin{aligned}
\widehat{r}_{nl}(t)\triangleq{}&\widehat{x}_{nl}(t)\Big(1+\nu^{r}_{nl}(t)\sum_{m=1}^{M}\nu^{a}_{mn}(t)[\widehat{s}^{2}_{ml}(t)-\nu^{s}_{ml}(t)]\Big)\\
&+\nu^{r}_{nl}(t)\sum_{m=1}^{M}\widehat{a}_{l,mn}(t)\widehat{s}_{ml}(t). \qquad (46)
\end{aligned}$$

Comparing (45)–(46) with (41)–(42) and applying previously established scalings from Table II reveals that $\nu^{r}_{m,nl}(t)-\nu^{r}_{nl}(t)$ is $O(1/N)$ and that $\widehat{r}_{m,nl}(t)=\widehat{r}_{nl}(t)-\nu^{r}_{nl}(t)\widehat{a}_{mn}(t)\widehat{s}_{ml}(t)+O(1/N)$, so that (43) implies

$$\begin{aligned}
&\widehat{x}_{m,nl}(t+1)\\
&= g_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t)-\nu^{r}_{nl}(t)\widehat{a}_{mn}(t)\widehat{s}_{ml}(t)+O(1/N),\\
&\quad\nu^{r}_{nl}(t)+O(1/N)\big) \qquad (47)\\
&= g_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t)-\nu^{r}_{nl}(t)\widehat{a}_{mn}(t)\widehat{s}_{ml}(t),\nu^{r}_{nl}(t)\big)+O(1/N) \qquad (48)\\
&= g_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t),\nu^{r}_{nl}(t)\big) \qquad (49)\\
&\quad-\nu^{r}_{nl}(t)\widehat{a}_{mn}(t)\widehat{s}_{ml}(t)g'_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t),\nu^{r}_{nl}(t)\big)+O(1/N)\\
&\approx \widehat{x}_{nl}(t+1)-\widehat{a}_{mn}(t)\widehat{s}_{ml}(t)\nu^{x}_{nl}(t+1). \qquad (50)
\end{aligned}$$

Above, (48) follows from taking Taylor series expansions around each of the $O(1/N)$ perturbations in (47); (49) follows from a Taylor series expansion in the first argument of (48) about the point $\widehat{r}_{nl}(t)$; and (50) follows by neglecting the $O(1/N)$ term (which vanishes relative to the others in the large-system limit) and applying the definitions

$$\widehat{x}_{nl}(t+1)\triangleq g_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t),\nu^{r}_{nl}(t)\big) \qquad (51)$$

$$\nu^{x}_{nl}(t+1)\triangleq\nu^{r}_{nl}(t)g'_{\mathsf{x}_{nl}}\big(\widehat{r}_{nl}(t),\nu^{r}_{nl}(t)\big), \qquad (52)$$

which match (43)–(44) sans the m dependence. Note that (50) confirms that the difference $\widehat{x}_{m,nl}(t)-\widehat{x}_{nl}(t)$ is $O(1/\sqrt{N})$, as was assumed at the start of the BiG-AMP derivation. Likewise, taking Taylor series expansions of $g'_{\mathsf{x}_{nl}}$ in (44) about the point $\widehat{r}_{nl}(t)$ in the first argument and about the point $\nu^r_{nl}(t)$ in the second argument and then comparing the result with (52) confirms that $\nu^x_{m,nl}(t) - \nu^x_{nl}(t)$ is $O(1/\sqrt{N})$.

We then repeat the above procedure to derive an approximation to $\Delta^{\mathsf{a}}_{l\leftarrow mn}(t+1,.)$ analogous to (40), whose corresponding mean is then further approximated as

$$\widehat{a}_{l,mn}(t+1) \approx \widehat{a}_{mn}(t+1) - \widehat{x}_{nl}(t)\widehat{s}_{ml}(t)\nu^a_{mn}(t+1), \tag{53}$$

for

$$\widehat{a}_{mn}(t+1) \triangleq g_{\mathsf{a}_{mn}}\big(\widehat{q}_{mn}(t), \nu^q_{mn}(t)\big) \tag{54}$$

$$\nu^a_{mn}(t+1) \triangleq \nu^q_{mn}(t) g'_{\mathsf{a}_{mn}}\big(\widehat{q}_{mn}(t), \nu^q_{mn}(t)\big) \tag{55}$$

$$g_{\mathsf{a}_{mn}}(\widehat{q}, \nu^q) \triangleq \frac{\int_a a\, p_{\mathsf{a}_{mn}}(a)\mathcal{N}(a; \widehat{q}, \nu^q)}{\int_a p_{\mathsf{a}_{mn}}(a)\mathcal{N}(a; \widehat{q}, \nu^q)} \tag{56}$$

where

$$\nu^q_{mn}(t) \triangleq \left(\sum_{l=1}^L \widehat{x}^2_{nl}(t)\nu^s_{ml}(t) - \nu^x_{nl}(t)\big(\widehat{s}^2_{ml}(t) - \nu^s_{ml}(t)\big)\right)^{-1} \tag{57}$$

$$\begin{aligned}\widehat{q}_{mn}(t) \triangleq{}& \widehat{a}_{mn}(t)\Big(1 + \nu^q_{mn}(t)\sum_{l=1}^L \nu^x_{nl}(t)[\widehat{s}^2_{ml}(t) - \nu^s_{ml}(t)]\Big) \\ &+ \nu^q_{mn}(t)\sum_{l=1}^L \widehat{x}_{m,nl}(t)\widehat{s}_{ml}(t).\end{aligned} \tag{58}$$

Arguments analogous to the discussion following (42) justify the remaining scalings in Table II.

F. Closing the Loop

The penultimate step in the derivation of BiG-AMP is to approximate earlier steps that use $\widehat{a}_{l,mn}(t)$ and $\widehat{x}_{m,nl}(t)$ in place of $\widehat{a}_{mn}(t)$ and $\widehat{x}_{nl}(t)$. For this, we start by plugging (50) and (53) into (22), which yields[3]

$$\begin{aligned}&\widehat{p}_{ml}(t) \\ &= O(1/\sqrt{N}) + \overbrace{\sum_{n=1}^N \widehat{a}_{mn}(t)\widehat{x}_{nl}(t)}^{\triangleq\, \overline{p}_{ml}(t)} - \widehat{s}_{ml}(t-1) \\ &\quad\times \sum_{n=1}^N \Big(\nu^a_{mn}(t)\widehat{x}_{nl}(t)\widehat{x}_{nl}(t-1) + \widehat{a}_{mn}(t)\widehat{a}_{mn}(t-1)\nu^x_{nl}(t)\Big) \\ &\quad + \widehat{s}^2_{ml}(t-1)\sum_{n=1}^N \widehat{a}_{mn}(t-1)\nu^a_{mn}(t)\nu^x_{nl}(t)\widehat{x}_{nl}(t-1)\end{aligned} \tag{59}$$

$$\approx \overline{p}_{ml}(t) - \widehat{s}_{ml}(t-1)\underbrace{\sum_{n=1}^N \Big(\nu^a_{mn}(t)\widehat{x}^2_{nl}(t) + \widehat{a}^2_{mn}(t)\nu^x_{nl}(t)\Big),}_{\triangleq\, \overline{\nu}^p_{ml}(t)} \tag{60}$$

where, for (60), we used $\widehat{a}^2_{mn}(t)$ in place of $\widehat{a}_{mn}(t)\widehat{a}_{mn}(t-1)$, used $\widehat{x}^2_{nl}(t)$ in place of $\widehat{x}_{nl}(t)\widehat{x}_{nl}(t-1)$, and neglected terms that are $O(1/\sqrt{N})$, since they vanish relative to the remaining $O(1)$ terms in the large-system limit.

Next we plug (50), (53), $\nu^x_{m,nl}(t) = \nu^x_{nl}(t) + O(1/\sqrt{N})$, and $\nu^a_{l,mn}(t) = \nu^a_{mn}(t) + O(1/N^{3/2})$ into (23), giving

$$\begin{aligned}\nu^p_{ml}(t) ={}& \overline{\nu}^p_{ml}(t) + \sum_{n=1}^N \nu^a_{mn}(t)\nu^x_{nl}(t) \\ &- 2\widehat{s}_{ml}(t-1)\sum_{n=1}^N \Big(\nu^a_{mn}(t)\widehat{a}_{mn}(t)\widehat{x}_{nl}(t-1)\nu^x_{nl}(t) \\ &+ \nu^a_{mn}(t)\widehat{a}_{mn}(t-1)\widehat{x}_{nl}(t)\nu^x_{nl}(t)\Big) \\ &+ \widehat{s}_{ml}(t-1)^2\sum_{n=1}^N \Big((\nu^a_{mn}(t))^2\widehat{x}^2_{nl}(t-1)\nu^x_{nl}(t) \\ &+ \nu^a_{mn}(t)\widehat{a}^2_{mn}(t-1)(\nu^x_{nl}(t))^2\Big) + O(1/\sqrt{N})\end{aligned} \tag{61}$$

$$\approx \overline{\nu}^p_{ml}(t) + \sum_{n=1}^N \nu^a_{mn}(t)\nu^x_{nl}(t), \tag{62}$$

where (62) retains only the $O(1)$ terms from (61).

Similarly, we plug (53) into (46) and (50) into (58) to obtain

$$\begin{aligned}\widehat{r}_{nl}(t) \approx{}& \widehat{x}_{nl}(t)\left(1 - \frac{\sum_{m=1}^M \nu^a_{mn}(t)\nu^s_{ml}(t)}{\sum_{m=1}^M \widehat{a}^2_{mn}(t)\nu^s_{ml}(t)}\right) \\ &+ \nu^r_{nl}(t)\sum_{m=1}^M \widehat{a}_{mn}(t)\widehat{s}_{ml}(t)\end{aligned} \tag{63}$$

$$\begin{aligned}\widehat{q}_{mn}(t) \approx{}& \widehat{a}_{mn}(t)\left(1 - \frac{\sum_{l=1}^L \nu^x_{nl}(t)\nu^s_{ml}(t)}{\sum_{l=1}^L \widehat{x}^2_{nl}(t)\nu^s_{ml}(t)}\right) \\ &+ \nu^q_{mn}(t)\sum_{l=1}^L \widehat{x}_{nl}(t)\widehat{s}_{ml}(t),\end{aligned} \tag{64}$$

where the approximations involve the use of $\widehat{s}^2_{ml}(t)$ in place of $\widehat{s}_{ml}(t)\widehat{s}_{ml}(t-1)$, of $\widehat{a}_{mn}(t)$ in place of $\widehat{a}_{mn}(t-1)$, of $\widehat{x}_{nl}(t)$ in place of $\widehat{x}_{nl}(t-1)$, and the dropping of terms that vanish in the large-system limit. Finally, we make the approximations

$$\nu^r_{nl}(t) \approx \left(\sum_{m=1}^M \widehat{a}^2_{mn}(t)\nu^s_{ml}(t)\right)^{-1} \tag{65}$$

$$\nu^q_{mn}(t) \approx \left(\sum_{l=1}^L \widehat{x}^2_{nl}(t)\nu^s_{ml}(t)\right)^{-1}, \tag{66}$$

by neglecting the $\widehat{s}^2_{ml}(t) - \nu^s_{ml}(t)$ terms in (45) and (57), as explained in Appendix B.

G. Approximated Posteriors

The final step in the BiG-AMP derivation is to approximate the SPA posterior log-pdfs in (11) and (12). Plugging (27) and (37) into those expressions, we get

$$\begin{aligned}&\Delta^{\mathsf{x}}_{nl}(t+1, x_{nl}) \\ &\approx \mathit{const} + \log\Big(p_{\mathsf{x}_{nl}}(x_{nl})\mathcal{N}\big(x_{nl}; \widehat{r}_{nl}(t), \nu^r_{nl}(t)\big)\Big)\end{aligned} \tag{67}$$

$$\begin{aligned}&\Delta^{\mathsf{a}}_{mn}(t+1, a_{mn}) \\ &\approx \mathit{const} + \log\Big(p_{\mathsf{a}_{mn}}(a_{mn})\mathcal{N}\big(a_{mn}; \widehat{q}_{mn}(t), \nu^q_{mn}(t)\big)\Big)\end{aligned} \tag{68}$$

using steps similar to (40). The associated pdfs are

$$\begin{aligned}&p_{\mathsf{x}_{nl}|\mathsf{r}_{nl}}\big(x_{nl}|\widehat{r}_{nl}(t); \nu^r_{nl}(t)\big) \\ &\triangleq \frac{1}{C_x} p_{\mathsf{x}_{nl}}(x_{nl})\mathcal{N}\big(x_{nl}; \widehat{r}_{nl}(t), \nu^r_{nl}(t)\big)\end{aligned} \tag{69}$$

[3] Recall that the error of the approximation in (50) is $O(1/N)$ and the error in (53) is $O(1/N^{3/2})$.

TABLE III
THE BIG-AMP ALGORITHM

```
definitions:
    $p_{\mathsf{z}_{ml}|\mathsf{p}_{ml}}(z|\widehat{p};\nu^p) \triangleq \frac{p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z)\,\mathcal{N}(z;\widehat{p},\nu^p)}{\int_{z'} p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z')\,\mathcal{N}(z';\widehat{p},\nu^p)}$   (D1)
    $p_{\mathsf{x}_{nl}|\mathsf{r}_{nl}}(x|\widehat{r};\nu^r) \triangleq \frac{p_{\mathsf{x}_{nl}}(x)\,\mathcal{N}(x;\widehat{r},\nu^r)}{\int_{x'} p_{\mathsf{x}_{nl}}(x')\,\mathcal{N}(x';\widehat{r},\nu^r)}$   (D2)
    $p_{\mathsf{a}_{mn}|\mathsf{q}_{mn}}(a|\widehat{q};\nu^q) \triangleq \frac{p_{\mathsf{a}_{mn}}(a)\,\mathcal{N}(a;\widehat{q},\nu^q)}{\int_{a'} p_{\mathsf{a}_{mn}}(a')\,\mathcal{N}(a';\widehat{q},\nu^q)}$   (D3)
initialization:
    $\forall m,l : \widehat{s}_{ml}(0) = 0$   (I1)
    $\forall m,n,l :$ choose $\nu^x_{nl}(1), \widehat{x}_{nl}(1), \nu^a_{mn}(1), \widehat{a}_{mn}(1)$   (I2)
for $t = 1, \ldots T_{\max}$
    $\forall m,l : \overline{\nu}^p_{ml}(t) = \sum_{n=1}^N |\widehat{a}_{mn}(t)|^2 \nu^x_{nl}(t) + \nu^a_{mn}(t)|\widehat{x}_{nl}(t)|^2$   (R1)
    $\forall m,l : \overline{p}_{ml}(t) = \sum_{n=1}^N \widehat{a}_{mn}(t)\widehat{x}_{nl}(t)$   (R2)
    $\forall m,l : \nu^p_{ml}(t) = \overline{\nu}^p_{ml}(t) + \sum_{n=1}^N \nu^a_{mn}(t)\nu^x_{nl}(t)$   (R3)
    $\forall m,l : \widehat{p}_{ml}(t) = \overline{p}_{ml}(t) - \widehat{s}_{ml}(t-1)\overline{\nu}^p_{ml}(t)$   (R4)
    $\forall m,l : \nu^z_{ml}(t) = \text{var}\{\mathsf{z}_{ml} \,|\, \mathsf{p}_{ml} = \widehat{p}_{ml}(t); \nu^p_{ml}(t)\}$   (R5)
    $\forall m,l : \widehat{z}_{ml}(t) = \text{E}\{\mathsf{z}_{ml} \,|\, \mathsf{p}_{ml} = \widehat{p}_{ml}(t); \nu^p_{ml}(t)\}$   (R6)
    $\forall m,l : \nu^s_{ml}(t) = (1 - \nu^z_{ml}(t)/\nu^p_{ml}(t))/\nu^p_{ml}(t)$   (R7)
    $\forall m,l : \widehat{s}_{ml}(t) = (\widehat{z}_{ml}(t) - \widehat{p}_{ml}(t))/\nu^p_{ml}(t)$   (R8)
    $\forall n,l : \nu^r_{nl}(t) = \big(\sum_{m=1}^M |\widehat{a}_{mn}(t)|^2 \nu^s_{ml}(t)\big)^{-1}$   (R9)
    $\forall n,l : \widehat{r}_{nl}(t) = \widehat{x}_{nl}(t)(1 - \nu^r_{nl}(t)\sum_{m=1}^M \nu^a_{mn}(t)\nu^s_{ml}(t))$
                    $+\nu^r_{nl}(t)\sum_{m=1}^M \widehat{a}^*_{mn}(t)\widehat{s}_{ml}(t)$   (R10)
    $\forall m,n : \nu^q_{mn}(t) = \big(\sum_{l=1}^L |\widehat{x}_{nl}(t)|^2 \nu^s_{ml}(t)\big)^{-1}$   (R11)
    $\forall m,n : \widehat{q}_{mn}(t) = \widehat{a}_{mn}(t)(1 - \nu^q_{mn}(t)\sum_{l=1}^L \nu^x_{nl}(t)\nu^s_{ml}(t))$
                    $+\nu^q_{mn}(t)\sum_{l=1}^L \widehat{x}^*_{nl}(t)\widehat{s}_{ml}(t)$   (R12)
    $\forall n,l : \nu^x_{nl}(t+1) = \text{var}\{\mathsf{x}_{nl} \,|\, \mathsf{r}_{nl} = \widehat{r}_{nl}(t); \nu^r_{nl}(t)\}$   (R13)
    $\forall n,l : \widehat{x}_{nl}(t+1) = \text{E}\{\mathsf{x}_{nl} \,|\, \mathsf{r}_{nl} = \widehat{r}_{nl}(t); \nu^r_{nl}(t)\}$   (R14)
    $\forall m,n : \nu^a_{mn}(t+1) = \text{var}\{\mathsf{a}_{mn} \,|\, \mathsf{q}_{mn} = \widehat{q}_{mn}(t); \nu^q_{mn}(t)\}$   (R15)
    $\forall m,n : \widehat{a}_{mn}(t+1) = \text{E}\{\mathsf{a}_{mn} \,|\, \mathsf{q}_{mn} = \widehat{q}_{mn}(t); \nu^q_{mn}(t)\}$   (R16)
    if $\sum_{m,l} |\overline{p}_{ml}(t) - \overline{p}_{ml}(t-1)|^2 \leq \tau_{\text{BiG-AMP}} \sum_{m,l} |\overline{p}_{ml}(t)|^2$, stop   (R17)
end
```

$$p_{\mathsf{a}_{mn}|\mathsf{q}_{mn}}\big(a_{mn}|\widehat{q}_{mn}(t);\nu^q_{mn}(t)\big) \triangleq \frac{1}{C_a} p_{\mathsf{a}_{mn}}(a_{mn})\mathcal{N}\big(a_{mn};\widehat{q}_{mn}(t),\nu^q_{mn}(t)\big) \tag{70}$$

for $C_x \triangleq \int_x p_{\mathsf{x}_{nl}}(x)\mathcal{N}\big(x;\widehat{r}_{nl}(t),\nu^r_{nl}(t)\big)$ and $C_a \triangleq \int_a p_{\mathsf{a}_{mn}}(a)\mathcal{N}\big(a;\widehat{q}_{mn}(t),\nu^q_{mn}(t)\big)$, which are iteration-$t$ BiG-AMP's approximations to the true marginal posteriors $p_{\mathsf{x}_{nl}|\mathbf{Y}}(x_{nl}|\boldsymbol{Y})$ and $p_{\mathsf{a}_{mn}|\mathbf{Y}}(a_{mn}|\boldsymbol{Y})$, respectively.

Note that $\widehat{x}_{nl}(t+1)$ and $\nu^x_{nl}(t+1)$ from (51)–(52) are the mean and variance, respectively, of the posterior pdf in (69). Note also that (69) can be interpreted as the (exact) posterior pdf of x_{nl} given the observation $\mathsf{r}_{nl} = \widehat{r}_{nl}(t)$ under the prior model $\mathsf{x}_{nl} \sim p_{\mathsf{x}_{nl}}$ and the likelihood model $p_{\mathsf{r}_{nl}|\mathsf{x}_{nl}}(\widehat{r}_{nl}(t)|x_{nl};\nu^r_{nl}(t)) = \mathcal{N}\big(\widehat{r}_{nl}(t);x_{nl},\nu^r_{nl}(t)\big)$ implicitly assumed by iteration-t BiG-AMP. Analogous statements can be made about the posterior pdf of a_{mn} in (70).

This completes the derivation of BiG-AMP. As a final comment, we note that some of the approximations used in (60), (63), and (64) are not exact in the large-system limit. Empirical experiments, however, suggest that the effect of these approximations are very minor. Although we would have liked to first present the "exact" version of the algorithm and then simplify it to yield the BiG-AMP in Table III, this was prevented by page constraints.

Algorithm Summary

The BiG-AMP algorithm derived in Sections II-C to II-G is summarized in Table III. There, we have included a maximum number of iterations, $T_{\max}$ and a stopping condition (R17) based on the (normalized) change in the residual and a user-defined parameter $\tau_{\text{BiG-AMP}}$. We have also written the algorithm in a more general form that allows the use of complex-valued quantities [note the complex conjugates in (R10) and (R12)], in which case $\mathcal{N}$ in (D1)-(D3) would be circular complex Gaussian. For ease of interpretation, Table III does not include the important damping modifications that will be detailed in Section IV-A. Suggestions for the initializations in (I2) will be given in the sequel.

We note that BiG-AMP avoids the use of SVD or QR decompositions, lending itself to simple and potentially parallel implementations. Its complexity order is dominated[4] by ten matrix multiplications per iteration [in steps (R1)-(R3) and (R9)-(R12)], each requiring MNL multiplications, although simplifications will be discussed in Section III.

The steps in Table III can be interpreted as follows. (R1)-(R2) compute a "plug-in" estimate $\overline{\boldsymbol{P}}$ of the matrix product $\boldsymbol{Z} = \boldsymbol{AX}$ and a corresponding set of element-wise variances $\{\overline{\nu}^p_{ml}\}$. (R3)-(R4) then apply "Onsager" correction (see [32] and [33] for discussions in the contexts of AMP and GAMP, respectively) to obtain the corresponding quantities $\widehat{\boldsymbol{P}}$ and $\{\nu^p_{ml}\}$. Using these quantities, (R5)-(R6) compute the (approximate) marginal posterior means $\widehat{\boldsymbol{Z}}$ and variances $\{\nu^z_{ml}\}$ of $\boldsymbol{Z}$. Steps (R7)-(R8) then use these posterior moments to compute the scaled residual $\widehat{\boldsymbol{S}}$ and a set of inverse-residual-variances $\{\nu^s_{ml}\}$. This interpretation becomes clear in the case of AWGN observations with noise variance ν^w, where

$$p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z_{ml}) = \mathcal{N}(y_{ml};z_{ml},\nu^w) \tag{71}$$

and hence

$$\nu^s_{ml} = \frac{1}{\nu^p_{ml} + \nu^w} \text{ and } \widehat{s}_{ml} = \frac{y_{ml} - \widehat{p}_{ml}}{\nu^p_{ml} + \nu^w}. \tag{72}$$

Steps (R9)-(R10) then use the residual terms $\widehat{\boldsymbol{S}}$ and $\{\nu^s_{ml}\}$ to compute $\widehat{\boldsymbol{R}}$ and $\{\nu^r_{nl}\}$, where $\widehat{r}_{nl}$ can be interpreted as a ν^r_{nl}-variance-AWGN corrupted observation of the true x_{nl}. Similarly, (R11)-(R12) compute $\widehat{\boldsymbol{Q}}$ and $\{\nu^q_{mn}\}$, where $\widehat{q}_{mn}$ can be interpreted as a ν^q_{mn}-variance-AWGN corrupted observation of the true a_{mn}. Finally, (R13)-(R14) merge these AWGN-corrupted observations with the priors $\{p_{\mathsf{x}_{nl}}\}$ to produce the posterior means $\widehat{\boldsymbol{X}}$ and variances $\{\nu^x_{nl}\}$; (R15)-(R16) do the same for the a_{mn} quantities.

The BiG-AMP algorithm in Table III is a direct (although non-trivial) extension of the GAMP algorithm for compressive sensing [33], which estimates $\boldsymbol{X}$ assuming perfectly known $\boldsymbol{A}$, and even stronger similarities to the $\boldsymbol{A}$-uncertain GAMP from [50], which estimates $\boldsymbol{X}$ assuming knowledge of the marginal means and variances of unknown random $\boldsymbol{A}$, but which makes no attempt to estimate $\boldsymbol{A}$ itself. In Section III-B, a simplified version of BiG-AMP will be developed that is similar to the Bayesian-AMP algorithm [31] for compressive sensing.

III. BIG-AMP SIMPLIFICATIONS

We now describe simplifications of the BiG-AMP algorithm from Table III that result from additional approximations and from the use of specific priors $p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}$, $p_{\mathsf{x}_{nl}}$, and $p_{\mathsf{a}_{mn}}$ that arise in practical applications of interest.

[4] The computations in steps (R4)-(R8) are $O(ML)$, while the remainder of the algorithm is $O(MN+NL)$. Thus, as N grows, the matrix multiplies dominate the complexity.

A. Scalar Variances

The BiG-AMP algorithm in Table III stores and processes a number of element-wise variance terms whose values vary across the elements (e.g., ν^x_{nl} can vary across n and l). The use of scalar variances (i.e., uniform across m, n, l) significantly reduces the memory and complexity of the algorithm.

To derive scalar-variance BiG-AMP, we first assume $\forall n, l : \nu^x_{nl}(t) \approx \nu^x(t) \triangleq \frac{1}{NL}\sum_{n=1}^{N}\sum_{l=1}^{L}\nu^x_{nl}(t)$ and $\forall m, n : \nu^a_{mn}(t) \approx \nu^a(t) \triangleq \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\nu^a_{mn}(t)$, so from (R1)

$$\begin{aligned}\overline{\nu}^p_{ml}(t) &\approx \nu^x(t)\sum_{n=1}^{N}|\widehat{a}_{mn}(t)|^2 + \nu^a(t)\sum_{n=1}^{N}|\widehat{x}_{nl}(t)|^2 && (73)\\ &\approx \frac{\|\widehat{\boldsymbol{A}}(t)\|_F^2}{M}\nu^x(t) + \frac{\|\widehat{\boldsymbol{X}}(t)\|_F^2}{L}\nu^a(t) \triangleq \overline{\nu}^p(t). && (74)\end{aligned}$$

Note that using (74) in place of (R1) avoids two matrix multiplies. Plugging these approximations into (R3) gives

$$\nu^p_{ml}(t) \approx \overline{\nu}^p(t) + N\nu^a(t)\nu^x(t) \triangleq \nu^p(t) \tag{75}$$

which, when used in place of (R3), avoids another matrix multiply. Even with the above scalar-variance approximations, $\{\nu^s_{ml}(t)\}$ from (R5) are not guaranteed to be equal (except in special cases like AWGN $p_{y_{ml}|z_{ml}}$). Still, they can be approximated as such using $\nu^s(t) \triangleq \frac{1}{ML}\sum_{m=1}^{M}\sum_{l=1}^{L}\nu^s_{ml}(t)$, in which case

$$\nu^r_{nl}(t) \approx \frac{1}{\nu^s(t)\sum_{m=1}^{M}|\widehat{a}_{mn}(t)|^2} \approx \frac{N}{\nu^s(t)\|\widehat{\boldsymbol{A}}(t)\|_F^2} \triangleq \nu^r(t) \tag{76}$$

$$\nu^q_{mn}(t) \approx \frac{1}{\nu^s(t)\sum_{l=1}^{L}|\widehat{x}_{nl}(t)|^2} \approx \frac{N}{\nu^s(t)\|\widehat{\boldsymbol{X}}(t)\|_F^2} \triangleq \nu^q(t). \tag{77}$$

Using (76) in place of (R9) and (77) in place of (R11) avoids two matrix multiplies and $NL + MN - 2$ scalar divisions, and furthermore allows (R10) and (R12) to be implemented as

$$\widehat{r}_{nl}(t) = \widehat{x}_{nl}(t)\left(1 - \frac{MN\nu^a(t)}{\|\widehat{\boldsymbol{A}}(t)\|_F^2}\right) + \nu^r(t)\sum_{m=1}^{M}\widehat{a}_{mn}(t)\widehat{s}_{ml}(t) \tag{78}$$

$$\widehat{q}_{mn}(t) = \widehat{a}_{mn}(t)\left(1 - \frac{NL\nu^x(t)}{\|\widehat{\boldsymbol{X}}(t)\|_F^2}\right) + \nu^q(t)\sum_{l=1}^{L}\widehat{x}_{nl}(t)\widehat{s}_{ml}(t), \tag{79}$$

saving two more matrix multiplies, and leaving a total of only three matrix multiplies per iteration.

B. Possibly Incomplete AWGN Observations

We now consider a particular observation model wherein the elements of $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X}$ are AWGN-corrupted at a subset of indices $\Omega \subset (1\ldots M) \times (1\ldots L)$ and unobserved at the remaining indices, noting that the standard AWGN model (71) is the special case where $|\Omega| = ML$. This "possibly incomplete AWGN" (PIAWGN) model arises in a number of important applications, such as matrix completion and dictionary learning.

We can state the PIAWGN model probabilistically as

$$p_{y_{ml}|z_{ml}}(y_{ml}|z_{ml}) = \begin{cases}\mathcal{N}(y_{ml}; z_{ml}, \nu^w) & (m,l) \in \Omega\\ \mathbb{1}_{y_{ml}} & (m,l) \notin \Omega,\end{cases} \tag{80}$$

where ν^w is the noise variance on the non-missing observations and $\mathbb{1}_y$ denotes a point mass at $y = 0$. Thus, at the observed entries $(m,l) \in \Omega$, the quantities $\widehat{s}_{ml}$ and ν^s_{ml} calculated using the AWGN expressions (72), while at the "missing" entries $(m,l) \notin \Omega$, where y_{ml} is invariant to z_{ml}, we have $\mathrm{E}\{\mathsf{z}_{ml}|\mathsf{p}_{ml} = \widehat{p}_{ml}; \nu^p_{ml}\} = \widehat{p}_{ml}$ and $\mathrm{var}\{\mathsf{z}_{ml}|\mathsf{p}_{ml} = \widehat{p}_{ml}; \nu^p_{ml}\} = \nu^p_{ml}$, so that $\widehat{s}_{ml} = 0$ and $\nu^s_{ml} = 0$. This is expected, given that ν^s can be interpreted as an inverse residual variance and $\widehat{s}$ as a ν^s-scaled residual. In summary, the PIAWGN model yields

$$\widehat{s}_{ml}(t) = \begin{cases}\frac{y_{ml} - \widehat{p}_{ml}(t)}{\nu^p_{ml}(t) + \nu^w} & (m,l) \in \Omega\\ 0 & (m,l) \notin \Omega\end{cases} \tag{81}$$

$$\nu^s_{ml}(t) = \begin{cases}\frac{1}{\nu^p_{ml}(t) + \nu^w} & (m,l) \in \Omega\\ 0 & (m,l) \notin \Omega\end{cases}. \tag{82}$$

When the PIAWGN model is combined with the scalar-variance approximations from Section III-A, BiG-AMP simplifies considerably. To see this, we start by using $\nu^p(t)$ from (75) in place of $\nu^p_{ml}(t)$ in (81)–(82), resulting in

$$\widehat{\boldsymbol{S}}(t) = P_\Omega\left(\frac{\boldsymbol{Y} - \widehat{\boldsymbol{P}}(t)}{\nu^p(t) + \nu^w}\right) \tag{83}$$

$$\nu^s(t) = \frac{\delta}{\nu^w + \nu^p(t)}, \tag{84}$$

where $\delta \triangleq \frac{|\Omega|}{ML}$ denotes the fraction of observed entries and $P_\Omega : \mathbb{R}^{M\times L} \to \mathbb{R}^{M\times L}$ is the projection operator defined by

$$\big[P_\Omega(\boldsymbol{Z})\big]_{ml} \triangleq \begin{cases} z_{ml} & (m,l) \in \Omega\\ 0 & (m,l) \notin \Omega\end{cases}. \tag{85}$$

We can then write (R10) and (R12) as

$$\widehat{\boldsymbol{R}}(t) = \widehat{\boldsymbol{X}}(t)\left(1 - \frac{MN\nu^a(t)}{\|\widehat{\boldsymbol{A}}(t)\|_F^2}\right) + \frac{N}{\delta\|\widehat{\boldsymbol{A}}(t)\|_F^2}\widehat{\boldsymbol{A}}^\top(t)\widehat{\boldsymbol{V}}(t) \tag{86}$$

$$\widehat{\boldsymbol{Q}}(t) = \widehat{\boldsymbol{A}}(t)\left(1 - \frac{NL\nu^x(t)}{\|\widehat{\boldsymbol{X}}(t)\|_F^2}\right) + \frac{N}{\delta\|\widehat{\boldsymbol{X}}(t)\|_F^2}\widehat{\boldsymbol{V}}(t)\widehat{\boldsymbol{X}}^\top(t) \tag{87}$$

using (78)–(79) and (83)–(85) with

$$\begin{aligned}\widehat{\boldsymbol{V}}(t) &\triangleq P_\Omega\big(\boldsymbol{Y} - \widehat{\boldsymbol{P}}(t)\big) && (88)\\ &= P_\Omega\big(\boldsymbol{Y} - \overline{\boldsymbol{P}}(t)\big) + \overline{\nu}^p(t)\widehat{\boldsymbol{S}}(t-1) && (89)\\ &= P_\Omega\big(\boldsymbol{Y} - \overline{\boldsymbol{P}}(t)\big) + \frac{\overline{\nu}^p(t)}{\nu^p(t-1) + \nu^w}\widehat{\boldsymbol{V}}(t-1), && (90)\end{aligned}$$

since P_Ω is a projection operator, and using (R4) and (83).

Scalar-variance BiG-AMP under PIAWGN observations is summarized in Table IV. Note that the residual matrix $\widehat{\boldsymbol{U}}(t) \triangleq$

TABLE IV
SCALAR-VARIANCE BIG-AMP WITH PIAWGN $p_{\mathsf{Y}|\mathsf{Z}}$

```
initialization:
    $\widehat{\boldsymbol{V}}(0) = \boldsymbol{0}$                                                        (I1p)
    choose $\nu^x(1), \widehat{\boldsymbol{X}}(1), \nu^a(1), \widehat{\boldsymbol{A}}(1)$                  (I2p)
for $t = 1, \ldots T_{\max}$
    $G_a(t) = \frac{N}{\delta \| \widehat{\boldsymbol{A}}(t) \|_F^2}$                                   (R1p)
    $G_x(t) = \frac{N}{\delta \| \widehat{\boldsymbol{X}}(t) \|_F^2}$                                   (R2p)
    $\widehat{\boldsymbol{U}}(t) = P_\Omega\big(\boldsymbol{Y} - \widehat{\boldsymbol{A}}(t)\widehat{\boldsymbol{X}}(t)\big)$   (R3p)
    $\overline{\nu}^p(t) = \big(\frac{\nu^x(t)}{MG_a(t)} + \frac{\nu^a(t)}{LG_x(t)}\big)\frac{N}{\delta}$   (R4p)
    $\nu^p(t) = \overline{\nu}^p(t) + N\nu^a(t)\nu^x(t)$                                     (R5p)
    $\widehat{\boldsymbol{V}}(t) = \widehat{\boldsymbol{U}}(t) + \frac{\overline{\nu}^p(t)}{\nu^p(t-1)+\nu^w}\widehat{\boldsymbol{V}}(t-1)$   (R6p)
    $\nu^r(t) = G_a(t)\big(\nu^p(t) + \nu^w\big)$                                        (R7p)
    $\widehat{\boldsymbol{R}}(t) = (1 - M\delta\nu^a(t)G_a(t))\widehat{\boldsymbol{X}}(t) + G_a(t)\widehat{\boldsymbol{A}}^{\mathsf{H}}(t)\widehat{\boldsymbol{V}}(t)$   (R8p)
    $\nu^q(t) = G_x(t)\big(\nu^p(t) + \nu^w\big)$                                        (R9p)
    $\widehat{\boldsymbol{Q}}(t) = (1 - L\delta\nu^x(t)G_x(t))\widehat{\boldsymbol{A}}(t) + G_x(t)\widehat{\boldsymbol{V}}(t)\widehat{\boldsymbol{X}}^{\mathsf{H}}(t)$   (R10p)
    $\nu^x(t+1) = \frac{1}{NL}\sum_{n=1}^N \sum_{l=1}^L \mathrm{var}\{\mathsf{x}_{nl} \mid \boldsymbol{Y}; \widehat{r}_{nl}(t), \nu^r(t)\}$   (R11p)
    $\forall n, l : \widehat{x}_{nl}(t+1) = \mathrm{E}\{\mathsf{x}_{nl} \mid \boldsymbol{Y}; \widehat{r}_{nl}(t), \nu^r(t)\}$   (R12p)
    $\nu^a(t+1) = \frac{1}{MN}\sum_{m=1}^M \sum_{n=1}^N \mathrm{var}\{\mathsf{a}_{mn} | \boldsymbol{Y}; \widehat{q}_{mn}(t), \nu^q(t)\}$   (R13p)
    $\forall m, n : \widehat{a}_{mn}(t+1) = \mathrm{E}\{\mathsf{a}_{mn} \mid \boldsymbol{Y}; \widehat{q}_{mn}(t), \nu^q(t)\}$   (R14p)
    if $\|\widehat{\boldsymbol{U}}(t) - \widehat{\boldsymbol{U}}(t-1)\|_F^2 \le \tau_{\text{BiG-AMP}} \|\widehat{\boldsymbol{U}}(t)\|_F^2$, stop   (R15p)
end
```

TABLE V
BIG-AMP-LITE: SCALAR-VARIANCE, PIAWGN, GAUSSIAN p_{X} AND p_{A}

```
initialization:
    $\widehat{\boldsymbol{V}}(0) = \boldsymbol{0}$                                                        (I1i)
    choose $\nu^x(1), \widehat{\boldsymbol{X}}(1), \nu^a(1), \widehat{\boldsymbol{A}}(1)$                  (I2i)
for $t = 1, \ldots T_{\max}$
    $G_a(t) = \frac{N}{\delta \| \widehat{\boldsymbol{A}}(t) \|_F^2}$                                   (R1i)
    $G_x(t) = \frac{N}{\delta \| \widehat{\boldsymbol{X}}(t) \|_F^2}$                                   (R2i)
    $\widehat{\boldsymbol{U}}(t) = P_\Omega\big(\boldsymbol{Y} - \widehat{\boldsymbol{A}}(t)\widehat{\boldsymbol{X}}(t)\big)$   (R3i)
    $\overline{\nu}^p(t) = \big(\frac{\nu^x(t)}{MG_a(t)} + \frac{\nu^a(t)}{LG_x(t)}\big)\frac{N}{\delta}$   (R4i)
    $\nu^p(t) = \overline{\nu}^p(t) + N\nu^a(t)\nu^x(t)$                                     (R5i)
    $\widehat{\boldsymbol{V}}(t) = \widehat{\boldsymbol{U}}(t) + \frac{\overline{\nu}^p(t)}{\nu^p(t-1)+\nu^w}\widehat{\boldsymbol{V}}(t-1)$   (R6i)
    $\nu^r(t) = G_a(t)\big(\nu^p(t) + \nu^w\big)$                                        (R7i)
    $\nu^q(t) = G_x(t)\big(\nu^p(t) + \nu^w\big)$                                        (R8i)
    $\nu^x(t+1) = \big(\frac{1}{\nu^r(t)} + \frac{1}{\nu_0^x}\big)^{-1}$                    (R9i)
    $\widehat{\boldsymbol{X}}(t+1) = \frac{\nu^x(t+1)}{\nu^r(t)}\big((1 - M\delta\nu^a(t)G_a(t))\widehat{\boldsymbol{X}}(t)$
        $+ G_a(t)\widehat{\boldsymbol{A}}^{\mathsf{H}}(t)\widehat{\boldsymbol{V}}(t)\big)$                    (R10i)
    $\nu^a(t+1) = \big(\frac{1}{\nu^q(t)} + \frac{1}{\nu_0^a}\big)^{-1}$                    (R11i)
    $\widehat{\boldsymbol{A}}(t+1) = \frac{\nu^a(t+1)}{\nu^q(t)}\big((1 - L\delta\nu^x(t)G_x(t))\widehat{\boldsymbol{A}}(t)$
        $+ G_x(t)\widehat{\boldsymbol{V}}(t)\widehat{\boldsymbol{X}}^{\mathsf{H}}(t)\big)$                    (R12i)
    if $\|\widehat{\boldsymbol{U}}(t) - \widehat{\boldsymbol{U}}(t-1)\|_F^2 \le \tau_{\text{BiG-AMP}} \|\widehat{\boldsymbol{U}}(t)\|_F^2$, stop   (R13i)
end
```

$P_\Omega(\boldsymbol{Y} - \widehat{\boldsymbol{A}}(t)\widehat{\boldsymbol{X}}(t))$ needs to be computed and stored only at the observed entries $(m,l) \in \Omega$, leading to significant savings[5] when the observations are highly incomplete (i.e., $|\Omega| \ll ML$). The same is true for the Onsager-corrected residual, $\widehat{\boldsymbol{V}}(t)$. Thus, the algorithm in Table IV involves only three (partial) matrix multiplies [in steps (R3p), (R8p), and (R10p), respectively], each of which can be computed using only $N|\Omega|$ scalar multiplies.

We note that Krzakala, Mézard, and Zdeborová recently proposed an AMP-based approach to blind calibration and dictionary learning [42] that bears close similarity[6] to BiG-AMP under the special case of AWGN-corrupted observations (i.e., $|\Omega| = ML$) and scalar variances. Their derivation differs significantly from that in Section II due to the many simplifications offered by this special case.

C. Zero-Mean iid Gaussian Priors on $\boldsymbol{\mathsf{A}}$ and $\boldsymbol{\mathsf{X}}$

In this section we will investigate the simplifications that result in the case that both $p_{\mathsf{a}_{mn}}$ and $p_{\mathsf{x}_{nl}}$ are zero-mean iid Gaussian, i.e.,

$$p_{\mathsf{x}_{nl}}(x) = \mathcal{N}(x; 0, \nu_0^x)\ \forall n, l \tag{91}$$

$$p_{\mathsf{a}_{mn}}(a) = \mathcal{N}(a; 0, \nu_0^a)\ \forall m, n, \tag{92}$$

which, as will be discussed later, is appropriate for matrix completion. In this case, straightforward calculations reveal that $\mathrm{E}\{x_{nl}|r_{nl} = \widehat{r}_{nl}; \nu_{nl}^r\} = \widehat{r}_{nl}\nu_0^x/(\nu_{nl}^r + \nu_0^x)$ and $\mathrm{var}\{x_{nl}|r_{nl} = \widehat{r}_{nl}; \nu_{nl}^r\}) = \nu_0^x\nu_{nl}^r/(\nu_{nl}^r + \nu_0^x)$ and, similarly, that $\mathrm{E}\{a_{mn}|q_{mn} = \widehat{q}_{mn}; \nu_{mn}^q\} = \widehat{q}_{mn}\nu_0^a/(\nu_{mn}^q + \nu_0^a)$ and $\mathrm{var}\{a_{mn}|q_{mn} = \widehat{q}_{mn}, \nu_{mn}^q\} = \nu_0^a\nu_{mn}^q/(\nu_{mn}^q + \nu_0^a)$. Combining these iid Gaussian simplifications with the scalar-variance simplifications from Section III-A yields an algorithm whose computational cost is dominated by three matrix multiplies per iteration, each with a cost of MNL scalar multiplies. The precise number of multiplies it consumes depends on the assumed likelihood model that determines steps (R7g)-(R8g).

Additionally incorporating the PIAWGN observations from Section III-B reduces the cost of the three matrix multiplies to only $N|\Omega|$ scalar multiplies each, and yields the "BiG-AMP-Lite" algorithm summarized in Table V, consuming $(3N + 5)|\Omega| + 3(MN + NL) + 29$ multiplies per iteration.

IV. ADAPTIVE DAMPING

The approximations made in the BiG-AMP derivation presented in Section II were well-justified in the large system limit, i.e., the case where $M, N, L \to \infty$ with fixed $\frac{M}{N}$ and $\frac{L}{N}$. In practical applications, however, these dimensions (especially N) are finite, and hence the algorithm presented in Section II may diverge. In case of compressive sensing, the use of "damping" with GAMP yields provable convergence guarantees with arbitrary matrices [36]. Here, we propose to incorporate damping into BiG-AMP. Moreover, we propose to *adapt* the damping of these variables to ensure that a particular cost criterion decreases monotonically (or near-monotonically), as described in the sequel. The specific damping strategy that we adopt is similar to that described in [51] and coded in [52].

A. Damping

In BiG-AMP, the iteration-t damping factor $\beta(t) \in (0, 1]$ is used to slow the evolution of certain variables, namely $\overline{\nu}_{ml}^p$, ν_{ml}^p,

[5] Similar computational savings also occur with incomplete non-Gaussian observations.

[6] The approach in [42] does not compute (or use) $\nu^p(t)$ as given in lines (R4p)-(R5p) of Table IV, but rather uses an empirical average of the squared Onsager-corrected residual in place of our $\nu^p(t) + \nu^w$ throughout their algorithm.

ν^s_{ml}, $\widehat{s}_{ml}$, $\widehat{x}_{nl}$, and $\widehat{a}_{mn}$. To do this, steps (R1), (R3), (R7), and (R8) in Table III are replaced with

$$\begin{aligned}\overline{\nu}^p_{ml}(t) &= \beta(t)\Big(\sum_{n=1}^{N}|\widehat{a}_{mn}(t)|^2\nu^x_{nl}(t) + \nu^a_{mn}(t)|\widehat{x}_{nl}(t)|^2\Big)\\ &\quad + (1-\beta(t))\overline{\nu}^p_{ml}(t-1) \qquad (93)\\ \nu^p_{ml}(t) &= \beta(t)\Big(\overline{\nu}^p_{ml}(t) + \sum_{n=1}^{N}\nu^a_{mn}(t)\nu^x_{nl}(t)\Big)\\ &\quad + (1-\beta(t))\nu^p_{ml}(t-1) \qquad (94)\\ \nu^s_{ml}(t) &= \beta(t)\big(1-\nu^z_{ml}(t)/\nu^p_{ml}(t)\big)/\nu^p_{ml}(t)\\ &\quad + (1-\beta(t))\nu^s_{ml}(t-1) \qquad (95)\\ \widehat{s}_{ml}(t) &= \beta(t)\big(\widehat{z}_{ml}(t) - \widehat{p}_{ml}(t)\big)/\nu^p_{ml}(t)\\ &\quad + (1-\beta(t))\widehat{s}_{ml}(t-1), \qquad (96)\end{aligned}$$

and the following are inserted between (R8) and (R9):

$$\overline{x}_{nl}(t) = \beta(t)\widehat{x}_{nl}(t) + (1-\beta(t))\overline{x}_{nl}(t-1) \qquad (97)$$

$$\overline{a}_{mn}(t) = \beta(t)\widehat{a}_{mn}(t) + (1-\beta(t))\overline{a}_{mn}(t-1). \qquad (98)$$

The newly defined state variables $\overline{x}_{nl}(t)$ and $\overline{a}_{mn}(t)$ are then used in place of $\widehat{x}_{nl}(t)$ and $\widehat{a}_{mn}(t)$ in steps (R9)-(R12) [but not (R1)-(R2)] of Table III. A similar approach can be used for the algorithm in Table IV (with the damping applied to $\widehat{\boldsymbol{V}}(t)$ instead of $\widehat{\boldsymbol{S}}(t)$) and those in Table V. Notice that, when $\beta(t) = 1$, the damping has no effect, whereas when $\beta(t) = 0$, all quantities become frozen in t.

B. Adaptive Damping

The idea behind adaptive damping is to monitor a chosen cost criterion $J(t)$ and decrease $\beta(t)$ when the cost has not decreased sufficiently[7] relative to $\{J(\tau)\}_{\tau=t-1-T}^{t-1}$ for some "step window" $T \geq 0$. This mechanism allows the cost criterion to increase over short intervals of T iterations and in this sense is similar to the procedure used by SpaRSA [53]. We now describe how the cost criterion $J(t)$ is constructed, building on ideas in [54].

Notice that, for fixed observations $\boldsymbol{Y}$, the joint posterior pdf solves the (trivial) KL-divergence minimization problem

$$p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}} = \arg\min_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}} D(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}\|p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}}). \qquad (99)$$

The factorized form (5) of the posterior allows us to write

$$\begin{aligned}&D(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}\|p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}}) - \log p_{\boldsymbol{\mathsf{Y}}}(\boldsymbol{Y})\\ &= \int_{\boldsymbol{A},\boldsymbol{X}} b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(\boldsymbol{A},\boldsymbol{X})\log\frac{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(\boldsymbol{A},\boldsymbol{X})}{p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{Z}}}(\boldsymbol{Y}|\boldsymbol{A}\boldsymbol{X})p_{\boldsymbol{\mathsf{X}}}(\boldsymbol{X})p_{\boldsymbol{\mathsf{A}}}(\boldsymbol{A})} \qquad (100)\\ &= D(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}\|p_{\boldsymbol{\mathsf{A}}}p_{\boldsymbol{\mathsf{X}}}) - \int_{\boldsymbol{A},\boldsymbol{X}} b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(\boldsymbol{A},\boldsymbol{X})\log p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{Z}}}(\boldsymbol{Y}|\boldsymbol{A}\boldsymbol{X}) \quad (101)\end{aligned}$$

Equations (99) and (101) then imply that

$$p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}} = \arg\min_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}} J(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}) \qquad (102)$$

$$J(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}) \triangleq D(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}\|p_{\boldsymbol{\mathsf{A}}}p_{\boldsymbol{\mathsf{X}}}) - \mathrm{E}_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}}\big\{\log p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{Z}}}(\boldsymbol{Y}|\boldsymbol{\mathsf{A}}\boldsymbol{\mathsf{X}})\big\}. \qquad (103)$$

To judge whether a given time-t BiG-AMP approximation "$b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)$" of the joint posterior $p_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}|\boldsymbol{\mathsf{Y}}}$ is better than the previous approximation $b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t-1)$, one could in principle plug the posterior approximation expressions (69)– (70) into (103) and then check whether $J(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)) < J(b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t-1))$. But, since the expectation in (103) is difficult to evaluate, we approximate the cost (103) by using, in place of $\boldsymbol{\mathsf{A}}\boldsymbol{\mathsf{X}}$, an independent Gaussian matrix[8] whose component means and variances are matched to those of $\boldsymbol{\mathsf{A}}\boldsymbol{\mathsf{X}}$. Taking the joint BiG-AMP posterior approximation $b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)$ to be the product of the marginals from (69)–(70), the resulting component means and variances are

$$\begin{aligned}\mathrm{E}_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)}\{[\boldsymbol{\mathsf{A}}\boldsymbol{\mathsf{X}}]_{ml}\} &= \sum_n \mathrm{E}_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)}\{\mathsf{a}_{mn}\}\mathrm{E}_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}}(t)\{\mathsf{x}_{nl}\} \qquad (104)\\ &= \sum_n \widehat{a}_{mn}(t)\widehat{x}_{nl}(t) = \overline{p}_{ml}(t) \qquad (105)\\ \mathrm{var}_{b_{\boldsymbol{\mathsf{X}},\boldsymbol{\mathsf{A}}}(t)}\{[\boldsymbol{\mathsf{A}}\boldsymbol{\mathsf{X}}]_{ml}\} &= \sum_n \widehat{a}^2_{mn}(t)\nu^x_{nl}(t) + \nu^a_{mn}(t)\widehat{x}^2_{nl}(t)\\ &\quad + \nu^a_{mn}(t)\nu^x_{nl}(t) \qquad (106)\\ &= \nu^p_{ml}(t). \qquad (107)\end{aligned}$$

In this way, the approximate iteration-t cost becomes

$$\begin{aligned}\widehat{J}(t) &= \sum_{n,l} D\Big(p_{\mathsf{x}_{nl}|\mathsf{r}_{nl}}\big(\cdot\,\big|\widehat{r}_{nl}(t);\nu^r_{nl}(t)\big)\Big\|p_{\mathsf{x}_{nl}}(\cdot)\Big)\\ &\quad + \sum_{m,n} D\Big(p_{\mathsf{a}_{mn}|\mathsf{q}_{mn}}\big(\cdot\,\big|\widehat{q}_{mn}(t);\nu^q_{mn}(t)\big)\Big\|p_{\mathsf{a}_{mn}}(\cdot)\Big)\\ &\quad - \sum_{m,l} \mathrm{E}_{\mathsf{z}_{ml}\sim\mathcal{N}(\overline{p}_{ml}(t);\nu^p_{ml}(t))}\big\{\log p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|\mathbf{z}_{ml})\big\}.\end{aligned} \qquad (108)$$

Intuitively, the first term in (108) penalizes the deviation between the (BiG-AMP approximated) posterior and the assumed prior on $\boldsymbol{\mathsf{X}}$, the second penalizes the deviation between the (BiG-AMP approximated) posterior and the assumed prior on $\boldsymbol{\mathsf{A}}$, and the third term rewards highly likely estimates $\boldsymbol{\mathsf{Z}}$.

V. Parameter Tuning and Rank Selection

A. Parameter Tuning via Expectation Maximization

Recall that BiG-AMP requires the specification of priors $p_{\boldsymbol{\mathsf{X}}}(\boldsymbol{X}) = \prod_{n,l} p_{\mathsf{x}_{nl}}(x_{nl})$, $p_{\boldsymbol{\mathsf{A}}}(\boldsymbol{A}) = \prod_{m,n} p_{\mathsf{a}_{mn}}(a_{mn})$, and $p_{\boldsymbol{\mathsf{Y}}|\boldsymbol{\mathsf{Z}}}(\boldsymbol{Y}|\boldsymbol{Z}) = \prod_{m,l} p_{\mathsf{y}_{ml}|\mathsf{z}_{ml}}(y_{ml}|z_{ml})$. In practice, although one may know appropriate families for these distributions, the exact parameters that govern them are generally unknown. For example, one may have good reason to believe apriori that the observations are AWGN corrupted, justifying the choice

[7] The following adaptation procedure is borrowed from GAMPmatlab[52], where it has been established to work well in the context of GAMP-based compressive sensing. When the current cost $J(t)$ is not smaller than the largest cost in the most recent `stepWindow` iterations, then the "step" is deemed unsuccessful, the damping factor $\beta(t)$ is reduced by the factor `stepDec`, and the step is attempted again. These attempts continue until either the cost criterion decreases or the damping factor reaches `stepMin`, at which point the step is considered successful, or the iteration count exceeds $T_{\max}$ or the damping factor reaches `stepTol`, at which point the algorithm terminates. When a step is deemed successful, the damping factor is increased by the factor `stepInc`, up to the allowed maximum value `stepMax`.

[8] The GAMP work [54] uses a similar approximation.

$p_{y_{ml}|z_{ml}}(y_{ml}|z_{ml}) = \mathcal{N}(y_{ml}; z_{ml}, \nu^w)$, but the noise variance ν^w may be unknown. In this section, we outline a methodology that takes a given set of BiG-AMP parameterized priors $\{p_{x_{nl}}(\cdot;\boldsymbol{\theta}), p_{a_{mn}}(\cdot;\boldsymbol{\theta}), p_{y_{ml}|z_{ml}}(y_{ml}|\cdot;\boldsymbol{\theta})\}_{\forall m,n,l}$ and tunes the parameter vector $\boldsymbol{\theta}$ using an expectation-maximization (EM) [37] based approach, with the goal of maximizing the likelihood, i.e., finding $\widehat{\boldsymbol{\theta}} \triangleq \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{Y}}(\boldsymbol{Y};\boldsymbol{\theta})$. The approach presented here can be considered as a generalization of the GAMP-based work [38] to BiG-AMP.

Taking $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{Z}$ to be the hidden variables, the EM recursion can be written as [37]

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{k+1} &= \arg\max_{\boldsymbol{\theta}} \mathrm{E}\Big\{ \log p_{\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y};\boldsymbol{\theta}) \Big| \boldsymbol{Y}; \widehat{\boldsymbol{\theta}}^k \Big\} \\ &= \arg\max_{\boldsymbol{\theta}} \Big\{ \sum_{n,l} \mathrm{E}\Big\{ \log p_{x_{nl}}(x_{nl};\boldsymbol{\theta}) \Big| \boldsymbol{Y}; \widehat{\boldsymbol{\theta}}^k \Big\} \\ &\quad + \sum_{m,n} \mathrm{E}\Big\{ \log p_{a_{mn}}(a_{mn};\boldsymbol{\theta}) \Big| \boldsymbol{Y}; \widehat{\boldsymbol{\theta}}^k \Big\} \\ &\quad + \sum_{m,l} \mathrm{E}\Big\{ \log p_{y_{ml}|z_{ml}}(y_{ml}|z_{ml};\boldsymbol{\theta}) \Big| \boldsymbol{Y}; \widehat{\boldsymbol{\theta}}^k \Big\} \Big\} \quad (109)\end{aligned}$$

where for (109) we used the fact $p_{\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{X},\boldsymbol{A},\boldsymbol{Z},\boldsymbol{Y};\boldsymbol{\theta}) = p_{\boldsymbol{X}}(\boldsymbol{X};\boldsymbol{\theta}) p_{\boldsymbol{A}}(\boldsymbol{A};\boldsymbol{\theta}) p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y}|\boldsymbol{Z};\boldsymbol{\theta}) \mathbb{1}_{\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{X}}$ and the factorizability of $p_{\boldsymbol{X}}$, $p_{\boldsymbol{A}}$, and $p_{\boldsymbol{Y}|\boldsymbol{Z}}$. As can be seen from (109), knowledge of the marginal posteriors $\{p_{x_{nl}|\boldsymbol{Y}}, p_{a_{mn}|\boldsymbol{Y}}, p_{z_{ml}|\boldsymbol{Y}}\}_{\forall m,n,l}$ is sufficient to compute the EM update. Since the exact marginal posteriors are unknown, we employ BiG-AMP's approximations from (69), (70), and (35) for approximate EM. In addition, we adopt the "incremental" update strategy from [55], where the maximization over $\boldsymbol{\theta}$ is performed one element at a time while holding the others fixed.

As a concrete example, consider updating the noise variance ν^w under the PIAWGN model (80). Equation (109) suggests

$$(\nu^w)^{k+1} = \arg\max_{\nu^w} \sum_{(m,l)\in\Omega} \int_{z_{ml}} p_{z_{ml}|\boldsymbol{Y}}(z_{ml}|\boldsymbol{Y}) \times \log \mathcal{N}(y_{ml}; z_{ml}, \nu^w), \quad (110)$$

where the true marginal posterior $p_{z_{ml}|\boldsymbol{Y}}(\cdot|\boldsymbol{Y})$ is replaced with the most recent BiG-AMP approximation $p_{z_{ml}|p_{ml}}(\cdot|\widehat{p}_{ml}(T_{\max}); \nu^p_{ml}(T_{\max}), \widehat{\boldsymbol{\theta}}^k)$, where "most recent" is with respect to both EM and BiG-AMP iterations. Zeroing the derivative of the sum in (110) with respect to ν^w,

$$(\nu^w)^{k+1} = \frac{1}{|\Omega|} \sum_{(m,l)\in\Omega} \big(y_{ml} - \widehat{z}_{ml}(T_{\max})\big)^2 + \nu^z_{ml}(T_{\max}), \quad (111)$$

where $\widehat{z}_{ml}(t)$ and $\nu^z_{ml}(t)$ are the BiG-AMP approximated posterior mean and variance from (33)–(34).

The overall procedure can be summarized as follows. From a suitable initialization $\widehat{\boldsymbol{\theta}}^0$, BiG-AMP is run using the priors $\{p_{x_{nl}}(\cdot;\widehat{\boldsymbol{\theta}}^0), p_{a_{mn}}(\cdot;\widehat{\boldsymbol{\theta}}^0), p_{y_{ml}|z_{ml}}(y_{ml}|\cdot;\widehat{\boldsymbol{\theta}}^0)\}_{\forall m,n,l}$ and iterated to completion, yielding approximate marginal posteriors on $\{x_{nl}, a_{mn}, z_{ml}\}_{\forall m,n,l}$. These posteriors are used in (109) to update the parameters $\boldsymbol{\theta}$ one element at a time, yielding $\widehat{\boldsymbol{\theta}}^1$. BiG-AMP is then run using the priors $\{p_{x_{nl}}(\cdot;\widehat{\boldsymbol{\theta}}^1), p_{a_{mn}}(\cdot;\widehat{\boldsymbol{\theta}}^1), p_{y_{ml}|z_{ml}}(y_{ml}|\cdot;\widehat{\boldsymbol{\theta}}^1)\}_{\forall m,n,l}$, and so on. A detailed discussion in the context of GAMP, along with explicit update equations for the parameters of Bernoulli-Gaussian-mixture pdfs, can be found in [38].

B. Rank Selection

BiG-AMP and EM-BiG-AMP, as described up to this point, require the specification of the rank N, i.e., the number of columns in $\boldsymbol{A}$ (and rows in $\boldsymbol{X}$) in the matrix factorization $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X}$. Since, in many applications, the best choice of N is difficult to specify in advance, we now describe two procedures to estimate N from the data $\boldsymbol{Y}$, building on well-known rank-selection procedures.

1) Penalized Log-Likelihood Maximization: Consider a set of possible models $\{\mathcal{H}_N\}_{N=1}^{\overline{N}}$ for the observation $\boldsymbol{Y}$ where, under $\mathcal{H}_N$, EM-BiG-AMP estimates $\boldsymbol{\Theta}_N = \{\boldsymbol{A}_N, \boldsymbol{X}_N, \boldsymbol{\theta}\}$. Here, the subscripts on $\boldsymbol{A}_N$ and $\boldsymbol{X}_N$ indicate the restriction to N columns and rows, $\boldsymbol{\theta}$ refers to the vector of parameters defined in Section V-A, and the subscript on $\boldsymbol{\Theta}_N$ indicates the dependence of the overall number of parameters in $\boldsymbol{\Theta}_N$ with the rank N. Because the selection rule $\widehat{N} = \arg\max_N p_{\boldsymbol{Y}}(\boldsymbol{Y};\mathcal{H}_N)$ is typically intractable, several well-known rules of the form

$$\widehat{N} = \arg\max_{N=1,\ldots,\overline{N}} 2\log p_{\boldsymbol{Y}|\boldsymbol{\Theta}_N}(\boldsymbol{Y}|\widehat{\boldsymbol{\Theta}}_N) - \eta(N) \quad (112)$$

have been developed, such as the Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC) [56]. In (112), $\widehat{\boldsymbol{\Theta}}_N$ is the ML estimate of $\boldsymbol{\Theta}_N$ under $\boldsymbol{Y}$, and $\eta(\cdot)$ is a penalty function that depends on the *effective* number of scalar parameters N_{eff} estimated under model $\mathcal{H}_N$ (which depends on N) and possibly on the number of scalar parameters $|\Omega|$ that make up the observation $\boldsymbol{Y}$.

Applying this methodology to EM-BiG-AMP, where $p_{\boldsymbol{Y}|\boldsymbol{\Theta}_N}(\boldsymbol{Y}|\boldsymbol{\Theta}_N) = p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y}|\boldsymbol{A}_N\boldsymbol{X}_N;\boldsymbol{\theta})$, we obtain the rank-selection rule

$$\widehat{N} = \arg\max_{N=1,\ldots,\overline{N}} 2\log p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y}|\widehat{\boldsymbol{A}}_N\widehat{\boldsymbol{X}}_N;\widehat{\boldsymbol{\theta}}) - \eta(N). \quad (113)$$

Since N_{eff} depends on the application (e.g., matrix completion, robust PCA, dictionary learning), detailed descriptions of $\eta(\cdot)$ are postponed to [1].

To perform the maximization over N in (113), we start with a small hypothesis N_1 and run EM-BiG-AMP to completion, generating the (approximate) MMSE estimates $\widehat{\boldsymbol{A}}_{N_1}, \widehat{\boldsymbol{X}}_{N_1}$ and ML estimate $\widehat{\boldsymbol{\theta}}$, which are then used to evaluate[9] the penalized log-likelihood in (113). The N hypothesis is then increased by a fixed value (i.e., $N_2 = N_1 + \texttt{rankStep}$), initializations of $(\boldsymbol{A}_{N_2}, \boldsymbol{X}_{N_2}, \boldsymbol{\theta})$ are chosen based on the previously computed $(\widehat{\boldsymbol{A}}_{N_1}, \widehat{\boldsymbol{X}}_{N_1}, \widehat{\boldsymbol{\theta}})$, and EM-BiG-AMP is run to completion,

[9] Since we compute approximate MMSE estimates rather than ML estimates, we are in fact evaluating a lower bound on the penalized log-likelihood.

yielding estimates $(\widehat{\boldsymbol{A}}_{N_2}, \widehat{\boldsymbol{X}}_{N_2}, \widehat{\boldsymbol{\theta}})$ with which the penalized likelihood is again evaluated. This process continues until either the value of the penalized log-likelihood decreases, in which case $\widehat{N}$ is set at the previous (i.e., maximizing) hypothesis of N, or the maximum-allowed rank $\overline{N}$ is reached.

2) Rank Contraction: We now describe an alternative rank-selection procedure that is appropriate when $\boldsymbol{Z}$ has a "cliff" in its singular value profile and which is reminiscent of that used in LMaFit [13]. In this approach, EM-BiG-AMP is initially configured to use the maximum-allowed rank, i.e., $N = \overline{N}$. After the first EM iteration, the singular values $\{\sigma_n\}$ of the estimate $\widehat{\boldsymbol{X}}$ and the corresponding pairwise ratios $R_n = \sigma_n/\sigma_{n+1}$ are computed,[10] from which a candidate rank estimate $\widehat{N} = \arg\max_n R_n$ is identified, corresponding to the largest gap in successive singular values. However, this candidate is accepted only if this maximizing ratio exceeds the average ratio by the user-specified parameter τ_{MOS} (e.g., $\tau_{\mathrm{MOS}} = 5$), i.e., if

$$R_{\widehat{N}} > \frac{\tau_{\mathrm{MOS}}}{\overline{N}-2}\sum_{i \neq \widehat{N}} R_i, \tag{114}$$

and if $\widehat{N}/\overline{N}$ is sufficiently small. Increasing τ_{MOS} makes the approach less prone to selecting an erroneous rank during the first few iterations, but making the value too large prevents the algorithm from detecting small gaps between the singular values. If $\widehat{N}$ is accepted, then the matrices $\boldsymbol{A}$ and $\boldsymbol{X}$ are pruned to size $\widehat{N}$ and EM-BiG-AMP is run to convergence. If not, EM-BiG-AMP is run for one more iteration, after which a new candidate $\widehat{N}$ is identified and checked for acceptance, and so on.

In many cases, a rank candidate is accepted after a small number of iterations, and thus only a few SVDs need be computed. This procedure has the advantage of running EM-BiG-AMP to convergence only once, rather than several times under different hypothesized ranks. However, when the singular values of $\boldsymbol{Z}$ decay smoothly, this procedure can mis-estimate the rank, as discussed in [13].

VI. Conclusion

In this work, we proposed and derived BiG-AMP, an extension of the G-AMP algorithm [33] originally proposed for high-dimensional generalized-*linear* regression in the context of compressive sensing, to generalized-*bilinear* regression, with applications in matrix completion, robust PCA, dictionary learning, and related matrix-factorization problems. In addition, we proposed an adaptive damping mechanism to aid convergence under realistic problem sizes, an expectation-maximization (EM)-based method to automatically tune the parameters of the assumed priors, and two rank-selection strategies. In Part II [1] of this two-part work, we detail the application of BiG-AMP to matrix completion, robust PCA, and dictionary learning, and we present the results of an extensive numerical investigation into the performance of BiG-AMP on both synthetic and real-world datasets. The results in [1] demonstrate that BiG-AMP yields excellent reconstruction accuracy (often best in class) while maintaining competitive runtimes, and that the proposed EM and rank-selection strategies successfully avoid the need to tune algorithmic parameters.

Appendix A

Here we derive (31)–(32), which are stated without a detailed derivation in [33]. Recalling (21) and omitting the ml subscripts for brevity, it can be seen that

$$\begin{aligned}
&H'\big(\widehat{q}, \nu^q; y\big)\\
&= \frac{\partial}{\partial \widehat{q}} \log \int p_{\mathsf{y}|\mathsf{z}}(y|z) \frac{1}{\sqrt{2\pi\nu^q}} \exp\Big(-\frac{1}{2\nu^q}(z-\widehat{q})^2\Big) dz\\
&= \frac{\partial}{\partial \widehat{q}} \Big\{ \log \int \exp\Big(\log p_{\mathsf{y}|\mathsf{z}}(y|z) - \frac{z^2}{2\nu^q} + \frac{\widehat{q}z}{\nu^q}\Big) dz - \frac{\widehat{q}^2}{2\nu^q}\Big\}\\
&= -\frac{\widehat{q}}{\nu^q} + \frac{\partial}{\partial \widehat{q}} \log \int \exp\big(\phi(u) + \widehat{q}u\big) du \text{ via } u \triangleq \frac{z}{\nu^q}
\end{aligned} \tag{115}$$

for an appropriately defined function $\phi(\cdot)$. Now, defining $p_{\mathsf{u}|\mathsf{q}}(u|\widehat{q}) \triangleq Z(\widehat{q})^{-1}\exp(\phi(u) + \widehat{q}u)$ with normalization term $Z(\widehat{q}) \triangleq \int \exp\big(\phi(u) + \widehat{q}u\big) du$, simple calculus yields

$$\frac{\partial}{\partial \widehat{q}} \log Z(\widehat{q}) = \mathrm{E}\{\mathsf{u}|\mathsf{q} = \widehat{q}\} \tag{116}$$

$$\frac{\partial^2}{\partial \widehat{q}^2} \log Z(\widehat{q}) = \mathrm{var}\{\mathsf{u}|\mathsf{q} = \widehat{q}\}. \tag{117}$$

Thus, from (115) and (116) it follows that

$$\begin{aligned}
&H'\big(\widehat{q}, \nu^q; y\big)\\
&= -\frac{\widehat{q}}{\nu^q} + \int u \frac{\exp\big(\phi(u) + \widehat{q}u\big)}{Z(\widehat{q})} du\\
&= -\frac{\widehat{q}}{\nu^q} + \int \frac{z}{\nu^q} \frac{\exp\big(\log p_{\mathsf{y}|\mathsf{z}}(y|z) - \frac{z^2}{2\nu^q} + \frac{\widehat{q}z}{\nu^q}\big)}{Z(\widehat{q})} \frac{dz}{\nu^q}\\
&= -\frac{\widehat{q}}{\nu^q} + \frac{1}{\nu^q} \int z \frac{p_{\mathsf{y}|\mathsf{z}}(y|z)\mathcal{N}(z; \widehat{q}, \nu^q)}{\int p_{\mathsf{y}|\mathsf{z}}(y|\overline{z})\mathcal{N}(\overline{z}; \widehat{q}, \nu^q) d\overline{z}} dz.
\end{aligned} \tag{118}$$

Equation (31) is then established by applying definitions (33) and (35) to (118).

Similarly, from (115) and (117),

$$\begin{aligned}
&-H''\big(\widehat{q}, \nu^q; y\big)\\
&= \frac{\partial}{\partial \widehat{q}} \Big\{ \frac{\widehat{q}}{\nu^q} - \frac{\partial}{\partial \widehat{q}} \log Z(\widehat{q}) \Big\} = \frac{1}{\nu^q} - \mathrm{var}\{\mathsf{u}|\mathsf{q} = \widehat{q}\}\\
&= \frac{1}{\nu^q} - \int \big(u - \mathrm{E}\{\mathsf{u}|\mathsf{q} = \widehat{q}\}\big)^2 \frac{\exp\big(\phi(u) + \widehat{q}u\big)}{Z(\widehat{q})} du\\
&= \frac{1}{\nu^q} - \frac{1}{(\nu^q)^2} \int \big(z - \widehat{z}\big)^2 \frac{p_{\mathsf{y}|\mathsf{z}}(y|z)\mathcal{N}(z; \widehat{q}, \nu^q)}{\int p_{\mathsf{y}|\mathsf{z}}(y|\overline{z})\mathcal{N}(\overline{z}; \widehat{q}, \nu^q) d\overline{z}} dz,
\end{aligned} \tag{119}$$

[10] In some cases the singular values of $\widehat{\boldsymbol{A}}$ could be used instead.

where $\widehat{z}$ is the expectation from (33). Equation (32) is then established by applying the definitions (34) and (35) to (119).

Appendix B

Here we explain the approximations (65)–(66). The term neglected in going from (45) to (65) can be written using (31)–(32) as

$$\begin{aligned}
&\sum_{m=1}^{M} \nu^a_{mn}(t)\big(\widehat{s}^2_{ml}(t) - \nu^s_{ml}(t)\big) \\
&= \sum_{m=1}^{M} \nu^a_{mn}(t)\left[\frac{(\widehat{z}_{ml}(t) - \widehat{p}_{ml}(t))^2 + \nu^z_{ml}(t)}{\nu^p_{ml}(t)^2} - \frac{1}{\nu^p_{ml}(t)}\right] \qquad (120) \\
&= \sum_{m=1}^{M} \frac{\nu^a_{mn}(t)}{\nu^p_{ml}(t)}\left[\mathrm{E}\left\{\frac{(\mathsf{z}_{ml} - \widehat{p}_{ml}(t))^2}{\nu^p_{ml}(t)}\right\} - 1\right] \qquad (121)
\end{aligned}$$

where the expectations are taken over $\mathsf{z}_{ml} \sim p_{\mathsf{z}_{ml}|\mathsf{p}_{ml}}\big(\cdot\,|\widehat{p}_{ml}(t); \nu^p_{ml}(t)\big)$ from (35). For GAMP, [57, Sec.VI.D] clarifies that, in the large system limit, under i.i.d priors and scalar variances, the true z_m and the iterates $\widehat{p}_m(t)$ converge empirically to a pair of random variables (z, p) that satisfy $p_{\mathsf{z}|\mathsf{p}}(z|\widehat{p}(t)) = \mathcal{N}(z; \widehat{p}(t), \nu^p(t))$. This result leads us to believe that the expectation in (121) is approximately unit-valued when averaged over m, and thus (121) is approximately zero-valued. Similar reasoning applies to (66).

Acknowledgment

The authors would like to thank Sundeep Rangan, Florent Krzakala, and Lenka Zdeborová for insightful discussions on various aspects of AMP-based inference. We would also like to thank Subhojit Som, Jeremy Vila, and Justin Ziniel for helpful discussions about EM and turbo methods for AMP.

References

[1] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854–5867, 2014.

[2] P. Schniter and V. Cevher, "Approximate message passing for bilinear models," in *Proc. Workshop Signal Process. Adapt. Sparse Struct. Repr. (SPARS)*, Edinburgh, Scotland, Jun. 2011, pp. 68–68.

[3] P. Schniter, J. T. Parker, and V. Cevher, "Bilinear generalized approximate message passing (BiG-AMP) for matrix recovery problems," presented at the Inf. Theory Appl. Workshop (ITA), La Jolla, CA, USA, Feb. 2012.

[4] J. T. Parker and P. Schniter, "Bilinear generalized approximate message passing (BiG-AMP) for matrix completion," presented at the Asilomar Conf., Pacific Grove, CA, USA, Nov. 2012.

[5] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010 [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/080738970

[6] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Math. Programm.*, vol. 128, no. 1-2, pp. 321–353, 2011 [Online]. Available: http://dx.doi.org/10.1007/s10107–009-0306–5

[7] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma, "Stable principal component pursuit," in *Proc. IEEE Int. Symp. Inf. Theory Process. (ISIT)*, Jun. 2010, pp. 1518–1522.

[8] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010 [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/070697835

[9] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," arXiv preprint, arXiv:1009.5055 [Online]. Available: http://arxiv.org/abs/1009.5055

[10] M. Tao and X. Yuan, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM J. Optim.*, vol. 21, no. 1, pp. 57–81, 2011.

[11] T. Zhou and D. Tao, L. Getoor and T. Scheffer, Eds., "Godec: Randomized low-rank and sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn. (ICML-11), Ser. ICML'11, ACM*, New York, NY, USA, Jun. 2011, pp. 33–40.

[12] H. Ghasemi, M. Malek-Mohammadi, M. Babaei-Zadeh, and C. Jutten, "SRF: Matrix completion based on smoothed rank function," presented at the IEEE Int. Conf. Acoust. Speech, Signal Process., Prague, Czech Republic, May 2011.

[13] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Programm. Comput.*, vol. 4, pp. 333–361, 2012 [Online]. Available: http://dx.doi.org/10.1007/s12532–012-0044–1

[14] A. Kyrillidis and V. Cevher, "Matrix recipes for hard thresholding methods," *J. Math. Imag. Vis.*, vol. 48, pp. 235–265, 2014.

[15] G. Marjanovic and V. Solo, "On optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, 2012.

[16] J. Haldar and D. Hernando, "Rank-constrained solutions to linear matrix equations using Power Factorization," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 584–587, Jul. 2009.

[17] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Allerton Conf. Commun. Control Comput.*, Sep. 2010, pp. 704–711.

[18] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[19] W. Dai and O. Milenkovic, "SET: An algorithm for consistent matrix completion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2010, pp. 3646–3649.

[20] J. He, L. Balzano, and J. Lui, Online robust subspace tracking from partial information arXiv:1109.3827 2011.

[21] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 1257–1264, 2008.

[22] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, Dec. 2011.

[23] Y. Lim and Y. Teh, "Variational Bayesian approach to movie rating prediction," in *Proc. KDD Cup Workshop*, 2007, pp. 15–21, Citeseer.

[24] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.

[25] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc., Series B (Statist. Methodol.)*, vol. 61, no. 3, pp. 611–622, 1999.

[26] F. Léger, G. Yu, and G. Sapiro, "Efficient matrix completion with Gaussian models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 1113–1116.

[27] N. Wang, T. Yao, J. Wang, and D. Yeung, A. Fitzgibbon, Ed., "A probabilistic approach to robust matrix factorization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, vol. VII, pp. 126–139.

[28] B.-H. Kim, A. Yedla, and H. Pfister, "Imp: A message-passing algorithm for matrix completion," in *Proc. 6th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, 2010, pp. 462–466.

[29] B. J. Frey and D. J. C. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Proc. Neural Inf. Process. Syst. Conf.*, Denver, CO, USA, 1997, pp. 479–485.

[30] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[31] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Proc. Inf. Theory Workshop*, Cairo, Egypt, Jan. 2010, pp. 1–5.

[32] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing: Theory and Applicant*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[33] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, Aug. 2011, pp. 2168–2172.

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

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

[36] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of generalized approximate message passing with arbitrary matrices," presented at the IEEE Int. Symp. Inf. Theory, Honolulu, HI, USA, Jul. 2014.

[37] A. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1–17, 1977.

[38] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[39] P. Schniter, "Turbo reconstruction of structured sparse signals," in *Proc. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 2010, pp. 1–6.

[40] J. Vila, P. Schniter, and J. Meola, "Hyperspectral image unmixing via bilinear generalized approximate message passing," *Proc. SPIE*, vol. 8743, no. 87430Y, p. 9, 2013.

[41] S. Rangan and A. K. Fletcher, "Iterative estimation of constrained rank-one matrices in noise," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 1246–1250.

[42] F. Krzakala, M. Mézard, and L. Zdeborová, "Phase diagram and approximate message passing for blind calibration and dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 659–663.

[43] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA, USA: Morgan Kaufmann, 1988.

[44] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, pp. 498–519, Feb. 2001.

[45] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artif. Intell.*, vol. 42, pp. 393–405, 1990.

[46] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Intl. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, Oct. 2000.

[47] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of pearl's 'belief propagation' algorithm," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 2, pp. 140–152, Feb. 1998.

[48] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. New York, NY, USA: Cambridge Univ. Press, 2003.

[49] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[50] J. T. Parker, V. Cevher, and P. Schniter, "Compressive sensing under matrix uncertainties: An approximate message passing approach," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2011, pp. 804–808.

[51] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," presented at the Allerton Conf. Commun., Control, Comput., Monticello, IL, USA, Oct. 2012.

[52] S. Rangan *et al.*, GAMPmatlab [Online]. Available: https://sourceforge.net/projects/gampmatlab/

[53] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

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

[55] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA, USA: MIT Press, 1999, pp. 355–368.

[56] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[57] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," Aug. 2012, arXiv:1010.5141v2.

Jason T. Parker (M'06) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from The Ohio State University, Columbus, in 2004, 2006, and 2014, respectively. Since 2006, he has been a research engineer with the U.S. Air Force Research Laboratory Sensors Directorate. His research interests currently include compressive sensing, adaptive signal processing, and inverse problems, with applications to radar target detection and imaging.

Philip Schniter (F'14) received the B.S. and M.S. degrees in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1992 and 1993, respectively, and the Ph.D. degree in Electrical Engineering from Cornell University in Ithaca, NY, in 2000.

From 1993 to 1996 he was employed by Tektronix Inc. in Beaverton, OR as a systems engineer. After receiving the Ph.D. degree, he joined the Department of Electrical and Computer Engineering at The Ohio State University, Columbus, where he is currently a Professor and a member of the Information Processing Systems (IPS) Lab. In 2008–2009, he was a visiting professor at Eurecom, Sophia Antipolis, France, and Supélec, Gif-sur-Yvette, France.

In 2003, Dr. Schniter received the National Science Foundation CAREER Award. His areas of interest currently include statistical signal processing, wireless communications and networks, and machine learning.

Volkan Cevher (SM'10) received the B.S. (valedictorian) degree in electrical engineering in 1999 from Bilkent University in Ankara, Turkey, and he received the Ph.D. degree in Electrical and Computer Engineering in 2005 from the Georgia Institute of Technology in Atlanta. He held research scientist positions at the University of Maryland, College Park from 2006 to 2007 and at Rice University in Houston, Texas, from 2008 to 2009. Currently, he is an Assistant Professor at the Swiss Federal Institute of Technology Lausanne with a complimentary appointment at the Electrical and Computer Engineering Department at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory. He received a Best Paper Award at SPARS in 2009 and an ERC StG in 2011.