Fixed Points of Generalized Approximate Message Passing With Arbitrary Matrices

Sundeep Rangan, *Fellow, IEEE*, Philip Schniter, *Fellow, IEEE*, Erwin Riegler, Alyson K. Fletcher, *Member, IEEE*, and Volkan Cevher, *Senior Member, IEEE*

***Abstract*— The estimation of a random vector with independent components passed through a linear transform followed by a componentwise (possibly nonlinear) output map arises in a range of applications. Approximate message passing (AMP) methods, based on Gaussian approximations of loopy belief propagation, have recently attracted considerable attention for such problems. For large random transforms, these methods exhibit fast convergence and admit precise analytic characterizations with testable conditions for optimality, even for certain non-convex problem instances. However, the behavior of AMP under general transforms is not fully understood. In this paper, we consider the generalized AMP (GAMP) algorithm and relate the method to more common optimization techniques. This analysis enables a precise characterization of the GAMP algorithm fixed points that applies to arbitrary transforms. In particular, we show that the fixed points of the so-called max-sum GAMP algorithm for MAP estimation are critical points of a constrained maximization of the posterior density. The fixed points of the sum-product GAMP algorithm for estimation of the posterior marginals can be interpreted as critical points of a certain free energy.**

***Index Terms*— Message passing, belief propagation, variational optimization, compressed sensing, ADMM.**

I. Introduction

CONSIDER the constrained optimization problem

$$(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}) := \arg\min_{\mathbf{x},\mathbf{z}} F(\mathbf{x}, \mathbf{z}) \quad \text{s.t. } \mathbf{z} = \mathbf{A}\mathbf{x}, \tag{1}$$

Manuscript received September 1, 2015; revised April 29, 2016; accepted September 19, 2016. Date of publication October 19, 2016; date of current version November 18, 2016. The work of S. Rangan was supported by the National Science Foundation under Grant 1116589, Grant 1237821, Grant 1302336, Grant 1564142, and Grant 1547332. The work of P. Schniter was supported by the National Science Foundation under Grant CCF-1018368, Grant CCF-1218754, and Grant CCF-1527162. The work of A. K. Fletcher was supported in part by NSF under Grant CCF-1254204 and in part by the Office of Naval Research Grant N00014-15-1-677. This paper was presented at the 2013 IEEE Symposium on Information Theory [1].

S. Rangan is with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: srangan@nyu.edu).

P. Schniter is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: schniter.1@osu.edu).

E. Riegler is with the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zürich, Switzerland (e-mail: eriegler@nari.ee.ethz.ch).

A. K. Fletcher is with the Departments of Statistics, Mathematics, and Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: akfletcher@ucla.edu).

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

Communicated by A. Montanari, Associate Editor for Statistical Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2016.2619365

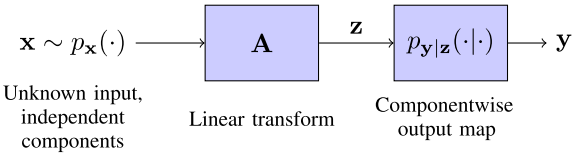

Fig. 1. System model: The GAMP method considered here can be used for approximate MAP and MMSE estimation of $\mathbf{x}$ from $\mathbf{y}$.

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, and the objective function admits a decomposition of the form

$$F(\mathbf{x}, \mathbf{z}) := f_x(\mathbf{x}) + f_z(\mathbf{z})$$
$$f_x(\mathbf{x}) = \sum_{j=1}^{n} f_{x_j}(x_j), \quad f_z(\mathbf{z}) = \sum_{i=1}^{m} f_{z_i}(z_i), \tag{2}$$

for scalar functions $f_{x_j}(\cdot)$ and $f_{z_i}(\cdot)$. One example where this optimization arises is the estimation problem in Fig. 1. Here, a random vector $\mathbf{x}$ has independent components with densities $p_{x_j}(x_j)$ and passes through a linear transform to yield an output $\mathbf{z} = \mathbf{A}\mathbf{x}$. The problem is to estimate $\mathbf{x}$ and $\mathbf{z}$ from measurements $\mathbf{y}$ generated according to a conditional density $p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})$ that is separable as a product of conditional densities $p_{y_i|z_i}(y_i|z_i)$. Under this observation model, the vectors $\mathbf{x}$ and $\mathbf{z}$ will have a posterior joint density given by

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})} e^{-F(\mathbf{x},\mathbf{z})} \mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}, \tag{3}$$

where $F(\mathbf{x}, \mathbf{z})$ is given by (2) when the scalar functions are set to the negative log prior density and likelihood:

$$f_{x_j}(x_j) = -\log p_{x_j}(x_j), \quad f_{z_i}(z_i) = -\log p_{y_i|z_i}(y_i|z_i).$$

Note that in (3), $F(\mathbf{x}, \mathbf{z})$ is implicitly a function of $\mathbf{y}$, $Z(\mathbf{y})$ is a normalization constant, and the point mass $\mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}$ imposes the linear constraint that $\mathbf{z} = \mathbf{A}\mathbf{x}$. The optimization (1) in this case produces the *maximum a posteriori* (MAP) estimate of $\mathbf{x}$ and $\mathbf{z}$. In statistics, the system in Fig. 1 is sometimes referred to as a generalized linear model [2], [3] and is used in a range of applications including regression, inverse problems, and filtering. Bayesian forms of compressed sensing can also be considered in this framework by imposing a sparse prior for the components x_j [4], [5]. In all these applications, one may instead be interested in estimating the posterior marginals $p(x_j|\mathbf{y})$ and $p(z_i|\mathbf{y})$. We relate this objective to an optimization of the form (1)-(2) in the sequel.

Most current numerical methods for solving the constrained optimization problem (1) attempt to exploit the separable structure of the objective function (2) either through generalizations of iterative shrinkage and thresholding (ISTA) algorithms [6]–[11] or the alternating direction method of multipliers (ADMM) approach [12]–[21]. There are now many of these methods, and we provide a brief review in Section II.

However, in recent years, there has been considerable interest in so-called approximate message passing (AMP) methods based on Gaussian and quadratic approximations of loopy belief propagation in graphical models [22]–[27]. The main appealing feature of the AMP algorithms is that for certain large random matrices $\mathbf{A}$, the asymptotic behavior of the algorithm can be rigorously and exactly predicted with testable conditions for optimality, even for many non-convex instances. Moreover, in the case of these large, random matrices, simulations appear to show very fast convergence of AMP methods when compared against state-of-the-art conventional optimization techniques.

Despite recent extensions to larger classes of random matrices [28], [29], the behavior of AMP methods under general $\mathbf{A}$ is not fully understood. Indeed, for general $\mathbf{A}$, it is well-known that AMP methods may diverge [30], [31]. While AMP has been successfully applied in a range of applications [32]–[36], the methods often require tuning to stabilize the algorithms. Various general procedures to stabilize AMP have also been proposed [30], [37]–[39].

To better understand these convergence issues, the broad purpose of this paper is to show that certain forms of AMP algorithms can be seen as variants of more conventional optimization methods. This analysis will enable a precise characterization of the fixed points of the AMP methods that applies to arbitrary $\mathbf{A}$, and a potential framework to understand the convergence.

Our study focuses on a Generalized AMP (GAMP) method proposed in [27] and rigorously analyzed in [40]. We consider this algorithm since many other variants of AMP are special cases of this general procedure. The GAMP method has two common versions: max-sum GAMP for the MAP estimation of the vectors $\mathbf{x}$ and $\mathbf{z}$ for the problem in Fig. 1, and sum-product GAMP for approximate inference of the posterior marginals.

For both versions of GAMP, the algorithms produce estimates $\mathbf{x}$ and $\mathbf{z}$ along with certain "quadratic" terms. Our first main result (Theorem 1) shows that the fixed points $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}})$ of max-sum GAMP are critical points of the optimization (1). In addition, the quadratic terms can be considered as diagonal approximations of the inverse Hessian of the objective function. For sum-product GAMP, we show (Theorem 2) that the algorithm's fixed points are stationary points of a certain energy function.

A conference version of this paper appeared in [1]. This paper includes all the proofs and more extensive discussion regarding relations between GAMP and classic optimization and free energy minimization techniques. In addition, since the publication of the conference version of this paper in [1], several other works such as [30], [39], [41], and [42] have built on the ideas and these are also discussed.

Algorithm 1 Generalized Approximate Message Passing (GAMP)

Require: Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, functions $f_x(\mathbf{x})$, $f_z(\mathbf{z}) \in \mathbb{R}$, and algorithm choice `MaxSum` or `SumProduct`.

$t \leftarrow 0$
Initialize $\mathbf{x}^t \in \mathbb{R}^n$, $\boldsymbol{\tau}_x^t \in \mathbb{R}^n_+$
$\mathbf{s}^{t-1} \leftarrow \mathbf{0} \in \mathbb{R}^m$
$\mathbf{S} \leftarrow \mathbf{A}.\mathbf{A}$ (componentwise square)
repeat
  {Output node update}
  $\boldsymbol{\tau}_p^t \leftarrow \mathbf{S}\boldsymbol{\tau}_x^t$
  $\mathbf{p}^t \leftarrow \mathbf{A}\mathbf{x}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}_p^t$
  if `MaxSum` **then**
    $\mathbf{z}^t \leftarrow \text{prox}_{\boldsymbol{\tau}_p^t f_z}(\mathbf{p}^t)$
    $\boldsymbol{\tau}_z^t \leftarrow \boldsymbol{\tau}_p^t.\,\text{prox}'_{\boldsymbol{\tau}_p^t f_z}(\mathbf{p}^t)$
  else if `SumProduct` **then**
    $\mathbf{z}^t \leftarrow \mathbb{E}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
    $\boldsymbol{\tau}_z^t \leftarrow \text{var}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
  end if
  $\mathbf{s}^t \leftarrow (\mathbf{z}^t - \mathbf{p}^t)./\boldsymbol{\tau}_p^t$
  $\boldsymbol{\tau}_s^t \leftarrow (\mathbf{1} - \boldsymbol{\tau}_z^t./\boldsymbol{\tau}_p^t)./\boldsymbol{\tau}_p^t$

  {Input node update}
  $\boldsymbol{\tau}_r^t \leftarrow \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s^t)$
  $\mathbf{r}^t \leftarrow \mathbf{x}^t + \boldsymbol{\tau}_r^t.\mathbf{A}^T\mathbf{s}^t$
  if `MaxSum` **then**
    $\mathbf{x}^{t+1} \leftarrow \text{prox}_{\boldsymbol{\tau}_r^t f_x}(\mathbf{r}^t)$
    $\boldsymbol{\tau}_x^{t+1} \leftarrow \boldsymbol{\tau}_r^t.\,\text{prox}'_{\boldsymbol{\tau}_r^t f_x}(\mathbf{r}^t)$
  else if `SumProduct` **then**
    $\mathbf{x}^{t+1} \leftarrow \mathbb{E}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
    $\boldsymbol{\tau}_x^{t+1} \leftarrow \text{var}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
  end if
until Terminated

II. Review of GAMP and Related Methods

A. Generalized Approximate Message Passing

Graphical-model methods [43] are a natural approach to the optimization problem (1) given the separable structure of the objective function (2). However, traditional graphical model techniques such as loopy belief propagation (loopy BP) are computationally attractive only when the constraint matrix $\mathbf{A}$ is sparse. Approximate message passing (AMP) refers to a class of Gaussian and quadratic approximations of loopy BP that can be applied to dense $\mathbf{A}$. AMP approximations of loopy BP originated in CDMA multiuser detection problems [44]–[46] and have received considerable recent attention in the context of compressed sensing [22]–[27], [47]. The Gaussian approximations used in AMP are also closely related to expectation propagation techniques [48], [49].

In this work, we study the so-called Generalized AMP (GAMP) algorithm [27] rigorously analyzed in [40]. The procedure, shown in Algorithm 1, produces a sequence of estimates $(\mathbf{x}^t, \mathbf{z}^t)$ of $(\mathbf{x}, \mathbf{z})$ along with the *quadratic terms* $\boldsymbol{\tau}_x^t, \boldsymbol{\tau}_r^t \in \mathbb{R}^n_+$ and $\boldsymbol{\tau}_z^t, \boldsymbol{\tau}_p^t, \boldsymbol{\tau}_s^t \in \mathbb{R}^m$, where $t \in \mathbb{Z}_+$ represents

Algorithm 2 Iterative Shrinkage and Thresholding Algorithm (ISTA)

Require: Matrix $\mathbf{A}$, scalar $c \geq 0$, functions $f_x(\cdot)$, $f_z(\cdot)$.
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$.
3: **repeat**
4: $\quad \mathbf{z}^t \leftarrow \mathbf{A}\mathbf{x}^t$
5: $\quad \mathbf{q}^t \leftarrow \nabla f_z(\mathbf{z}^t)$
6: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} f_x(\mathbf{x}) + (\mathbf{q}^t)^T\mathbf{A}\mathbf{x} + (c/2)\|\mathbf{x} - \mathbf{x}^t\|^2$
7: **until** Terminated

the iteration number. Here and in the sequel, we use "." to denote componentwise vector multiplication and "./" to denote componentwise vector division.

We focus on two variants of the GAMP algorithm: *max-sum GAMP* and *sum-product GAMP*.

1) Max-Sum GAMP: In the max-sum version of the algorithm, the outputs $(\mathbf{x}^t, \mathbf{z}^t)$ represent estimates of the solution to the optimization problem (1), or equivalently the MAP estimates for the posterior (3). Since the objective function has the separable form (2), each iteration of the algorithm involves four componentwise update steps: the proximal updates shown in lines 10 and 23, where

$$\text{prox}_f(v) := \arg\min_{u\in\mathbb{R}} f(u) + \frac{1}{2}(u - v)^2, \tag{4}$$

and lines 11 and 24, involving the derivative of the proximal operator from (4).

In particular, lines 10 and 11 are to be interpreted as

$$z_i^t = \text{prox}_{\tau_{p_i}^t f_{z_i}}(p_i^t), \quad i = 1, \ldots, m, \tag{5}$$

$$\tau_{z_i}^t = \tau_{p_i}^t \text{prox}'_{\tau_{p_i}^t f_{z_i}}(p_i^t), \quad i = 1, \ldots, m, \tag{6}$$

$$= \tau_{p_i}^t \left(1 + \tau_{p_i}^t \frac{\partial^2 f_{z_i}(z_i^t)}{\partial z_i^2}\right)^{-1}, \quad i = 1, \ldots, m, \tag{7}$$

with similar interpretations for lines 23 and 24. Thus, max-sum GAMP reduces the vector-valued optimization (1) to a sequence of scalar optimizations.

When discussing max-sum GAMP, we will assume that both f_x and f_z are twice differentiable and convex, so that the outputs of the proximal operator and its derivative exist and are unique. We make these assumptions for the sake of clarity, but note that—in practice—GAMP is often used with non-differentiable functions. A common example is when $f_x(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ for $\lambda > 0$, in which case

$$\text{prox}_{\tau_{r_j}^t f_{x_j}}(r_j^t) = \text{sgn}(r_j^t)\max\{|r_j^t| - \lambda\tau_{r_j}^t, 0\} \tag{8}$$

and

$$\text{prox}'_{\tau_{r_j}^t f_{x_j}}(r_j^t) = \begin{cases} 1, & |r_j^t| > \lambda\tau_{r_j}^t; \\ 0, & |r_j^t| < \lambda\tau_{r_j}^t. \end{cases} \tag{9}$$

Although $\text{prox}'_{\tau_{r_j}^t f_{x_j}}(r_j^t)$ is undefined when $r_j^t = \lambda\tau_{r_j}^t$, its value can be set to either 0 or 1 with minimal effect, because the event $r_j^t = \lambda\tau_{r_j}^t$ almost never occurs (due, e.g., to the presence of noise in r_j^t). The rigorous GAMP analysis [40] assumes only that the prox functions in lines 10 and 23 are Lipschitz continuous (and hence differentiable almost everywhere).

2) Sum-Product GAMP: The purpose of the sum-product GAMP algorithm is to provide estimates of the posterior marginals

$$p(x_j|\mathbf{y}), \quad p(z_i|\mathbf{y}), \tag{10}$$

from the joint density (3). Exact computation of these marginal densities is, in general, computationally intractable. Sum-product GAMP instead provides estimates of these densities. Specifically, at each iteration t, it forms the estimated densities, called *beliefs*, given by:

$$b_{x_j}^t(x_j) = p(x_j|r_j^t, \tau_{r_j}^t), \quad b_{z_i}^t(z_i) = p(z_i|p_i^t, \tau_{p_i}^t), \tag{11}$$

where we use the notation

$$p(x_j|r_j, \tau_{r_j}) \propto \exp\left[-f_{x_j}(x_j) - \tfrac{1}{2\tau_{r_j}}(x_j - r_j)^2\right], \tag{12a}$$

$$p(z_i|p_i, \tau_{p_i}) \propto \exp\left[-f_{z_i}(z_i) - \tfrac{1}{2\tau_{p_i}}(z_i - p_i)^2\right]. \tag{12b}$$

As we will discuss in Section IV, these belief estimates can be "derived" as estimates of the minima of a certain large system limit of the Bethe Free Energy.

Now, the products of the densities in (12) are given by

$$p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r) = \prod_{j=1}^n p(x_j|r_j, \tau_{r_j}) \propto \exp\left[-f_x(\mathbf{x}) - \frac{1}{2}\|\mathbf{x} - \mathbf{r}\|_{\boldsymbol{\tau}_r}^2\right], \tag{13a}$$

$$p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p) = \prod_{i=1}^m p(z_i|z_i, \tau_{p_i}) \propto \exp\left[-f_z(\mathbf{z}) - \frac{1}{2}\|\mathbf{z} - \mathbf{p}\|_{\boldsymbol{\tau}_p}^2\right], \tag{13b}$$

where, for any vectors $\mathbf{v} \in \mathbb{R}^r$ and $\boldsymbol{\tau} \in \mathbb{R}^r$ with $\boldsymbol{\tau} > 0$, we use the notation

$$\|\mathbf{v}\|_{\boldsymbol{\tau}}^2 := \sum_{i=1}^r \frac{|v_i|^2}{\tau_i}.$$

In the sum-product version of GAMP, the expectations and variances in lines 13, 14, 26 and 27 of Algorithm 1 are to be taken with respect to the probability density functions in (13). Thus, $\mathbf{x}^t$ and $\boldsymbol{\tau}_x^t$ are the estimates of the posterior means and variances of the components of $\mathbf{x}$ and $\mathbf{z}^t$ and $\boldsymbol{\tau}_z^t$ are the estimates of the posterior means and variances of the components of $\mathbf{z}$.

Since the densities (13) are separable, the expectations and variances can be computed via scalar integrals. Thus, the sum-product GAMP algorithm reduces the vector-valued to marginalization problem to a sequence of scalar estimation problems.

B. Iterative Shrinkage and Thresholding Algorithm

The goal in the paper is to relate the GAMP method to more conventional optimization techniques. One of the more common of such approaches is a generalization of the

Algorithm 3 Alternating Direction Method of Multipliers (ADMM)

Require: $\mathbf{A}$, α, functions $f_x(\cdot)$, $f_z(\cdot)$, $Q_x(\cdot)$, $Q_z(\cdot)$
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$, $\mathbf{z}^t$, $\mathbf{s}^t$
3: **repeat**
4: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}^t, \mathbf{s}^t) + Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha)$
5: $\quad \mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z}} L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{s}^t) + Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha)$
6: $\quad \mathbf{s}^{t+1} \leftarrow \mathbf{s}^t + \alpha(\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1})$
7: **until** Terminated

Iterative Shrinkage and Thresholding Algorithm (ISTA) shown in Algorithm 2 [6]–[9], where ∇f denotes the gradient of f.

The algorithm is built on the idea that, at each iteration t, the second cost term in the minimization $\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + f_z(\mathbf{A}\mathbf{x})$ specified by (1) is replaced by a quadratic majorizing cost $g_z(\mathbf{x}) \geq f_z(\mathbf{A}\mathbf{x})$ that coincides at the point $\mathbf{x} = \mathbf{x}^t$ (i.e., $g_z(\mathbf{x}^t) = f_z(\mathbf{A}\mathbf{x}^t)$). The function $g_z(\mathbf{x})$ defined implicitly in line 6 achieves this majorization via appropriate choice of $c > 0$. This approach is motivated by the fact that, if $f_x(\mathbf{x})$ and $f_z(\mathbf{z})$ are both separable, as in (2), then both the gradient in line 5 and minimization in line 6 can be performed componentwise. Moreover, when $f_x(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, as in the LASSO problem [50], the minimization in line 6 can be computed directly via the shrinkage and thresholding operation (8)—hence the name of the algorithm. The convergence of the ISTA method tends to be slow, but a number of enhanced methods have been successful and widely-used [8]–[11].

C. Alternating Direction Method of Multipliers

A second common class of methods is built around the Alternating Direction Method of Multipliers (ADMM) [12] approach shown in Algorithm 3. The Lagrangian for the optimization problem (1) is given by

$$L(\mathbf{x}, \mathbf{z}, \mathbf{s}) := F(\mathbf{x}, \mathbf{z}) + \mathbf{s}^T(\mathbf{z} - \mathbf{A}\mathbf{x}), \tag{14}$$

where $\mathbf{s}$ are the dual parameters. ADMM attempts to produce a sequence of estimates $(\mathbf{x}^t, \mathbf{z}^t, \mathbf{s}^t)$ that converge to a saddle point of the Lagrangian (14). The parameters of the algorithm are a step-size $\alpha > 0$ and the penalty terms $Q_z(\cdot)$ and $Q_x(\cdot)$, which classical ADMM would choose as

$$Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha) = \frac{\alpha}{2}\|\mathbf{z}^t - \mathbf{A}\mathbf{x}\|^2 \tag{15a}$$

$$Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha) = \frac{\alpha}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^{t+1}\|^2. \tag{15b}$$

When the objective function admits a separable form (2) and one uses the auxiliary function $Q_z(\cdot)$ in (15b), the $\mathbf{z}$-minimization in line 5 separates into m scalar optimizations. However, due to the quadratic term $\|\mathbf{A}\mathbf{x}\|^2$ in (15a), the $\mathbf{x}$-minimization in line 4 does not separate for general $\mathbf{A}$. To circumvent this problem, one might consider a separable inexact $\mathbf{x}$-minimization, since many inexact variants of ADMM are known to converge [13]. For example, $Q_x(\cdot)$ might be chosen to yield separability while majorizing the original cost in line 4, as was done for ISTA's line 6, i.e.,

$$Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha) = \frac{\alpha}{2}\|\mathbf{z}^t - \mathbf{A}\mathbf{x}\|^2 + \frac{1}{2}(\mathbf{x} - \mathbf{x}^t)^T(c\mathbf{I} - \alpha\mathbf{A}^T\mathbf{A})(\mathbf{x} - \mathbf{x}^t) \tag{16}$$

with $c \geq \alpha\|\mathbf{A}\|^2$, after which ADMM's line 4 would become

$$\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + \frac{c}{2}\left\|\mathbf{x} - \mathbf{x}^t + \frac{\alpha}{c}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^t - \mathbf{z}^t - \frac{1}{\alpha}\mathbf{s}^t\right)\right\|^2. \tag{17}$$

This approach is known as "linearized ADMM" [51], or as "split inexact Uzawa" [15] in the optimization literature, and it has close connections to other well-known techniques like Douglas–Rachford splitting [13], split Bregman [14], proximal forward-backward splitting [16], and various primal-dual algorithms [17]–[21]. Many other choices of penalty $Q_x(\cdot)$ have also been considered in the literature (see, e.g., the overview in [19]).

Other variants of ADMM are also possible [12]. For example, the step-size α might vary with the iteration t, or the penalty terms might have the form $(\mathbf{z} - \mathbf{A}\mathbf{x})^T\mathbf{P}(\mathbf{z} - \mathbf{A}\mathbf{x})$ for positive semidefinite $\mathbf{P}$. As we will see, these generalizations provide a connection to GAMP.

III. Fixed-Points of Max-Sum GAMP

Our first result connects the max-sum GAMP algorithm to inexact ADMM. Given points $(\mathbf{x}, \mathbf{z})$, define the matrices

$$\mathbf{Q}_x := \left(\mathrm{Diag}(\mathbf{d}_x) + \mathbf{A}^T\,\mathrm{Diag}(\mathbf{d}_z)\mathbf{A}\right)^{-1} \tag{18a}$$

$$\mathbf{Q}_z := \left(\mathrm{Diag}(\mathbf{d}_z)^{-1} + \mathbf{A}\,\mathrm{Diag}(\mathbf{d}_x)^{-1}\mathbf{A}^T\right)^{-1} \tag{18b}$$

where $\mathrm{Diag}(\mathbf{d})$ denotes the diagonal matrix with diagonal entries equal to those in the vector $\mathbf{d}$, and where $\mathbf{d}_x$ and $\mathbf{d}_z$ contain the componentwise second derivatives, i.e., the diagonals of the Hessian matrices

$$\mathbf{d}_x := \mathrm{diag}\left[\mathcal{H} f_x(\mathbf{x})\right], \quad \mathbf{d}_z := \mathrm{diag}\left[\mathcal{H} f_z(\mathbf{z})\right]. \tag{19}$$

Note that when f_x and f_z are strictly convex, the elements in $\mathbf{d}_x$ and $\mathbf{d}_z$ are positive. Observe that the matrix $\mathbf{Q}_x$ in (18a) is the inverse Hessian of the objective function $F(\mathbf{x}, \mathbf{z})$ constrained to $\mathbf{z} = \mathbf{A}\mathbf{x}$. That is,

$$\mathbf{Q}_x = \left[\mathcal{H}_{\mathbf{x}} F(\mathbf{x}, \mathbf{A}\mathbf{x})\right]^{-1}.$$

Theorem 1: The outputs of the max-sum GAMP version of Algorithm 1 satisfy the recursions

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}}\left[L(\mathbf{x}, \mathbf{z}^t, \mathbf{s}^t) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^t\|^2_{\boldsymbol{\tau}_r^t}\right] \tag{20a}$$

$$\mathbf{z}^{t+1} = \arg\min_{\mathbf{z}}\left[L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{s}^t) + \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^{t+1}\|^2_{\boldsymbol{\tau}_p^{t+1}}\right] \tag{20b}$$

$$\mathbf{s}^{t+1} = \mathbf{s}^t + (\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1})./\boldsymbol{\tau}_p^{t+1} \tag{20c}$$

where $L(\mathbf{x}, \mathbf{z}, \mathbf{s})$ is the Lagrangian defined in (14).

Now suppose that $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_s)$ is a fixed point of the algorithm (where the "hats" on $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{z}}$ are used to distinguish them from free variables). Then, this fixed point is a critical point of the constrained optimization (1) *in that $\widehat{\mathbf{z}} = \mathbf{A}\widehat{\mathbf{x}}$ and*

$$\nabla_{\mathbf{x}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}, \quad \nabla_{\mathbf{z}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}. \tag{21}$$

Moreover, the quadratic terms $\boldsymbol{\tau}_x, \boldsymbol{\tau}_s$ *are the* approximate diagonals *(as defined in Appendix A) of* $\mathbf{Q}_x$ *and* $\mathbf{Q}_z$ *in* (18) *at* $(\mathbf{x}, \mathbf{z}) = (\widehat{\mathbf{x}}, \widehat{\mathbf{z}})$.

Proof: See Appendix B. ∎

The first part of the theorem, equations (20), shows that max-sum GAMP can be interpreted as the ADMM Algorithm 3 with adaptive vector-valued step-sizes $\boldsymbol{\tau}_r^t$ and $\boldsymbol{\tau}_p^t$ and a particular choice of penalty $Q_x(\cdot)$. To more precisely connect GAMP and existing algorithms, it helps to express GAMP's $\mathbf{x}$-update (20a) as the $\theta = 0$ case of

$$\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^t + \boldsymbol{\tau}_r^t . \mathbf{A}^T\left(\theta(\mathbf{s}^{t-1} - \mathbf{s}^t) - \mathbf{s}^t\right)\right\|_{\boldsymbol{\tau}_r^t}^2, \tag{22}$$

and recognize that the ISTA-inspired inexact ADMM $\mathbf{x}$-update (17) coincides with the $\theta = 1$ case under step-sizes $\alpha = 1/\boldsymbol{\tau}_p^t$ and $c = 1/\boldsymbol{\tau}_r^t$. The convergence of this algorithm for particular $\theta \in [0, 1]$ was studied in [19]–[21] under convex functions $f_x(\cdot)$ and $f_z(\cdot)$ and non-adaptive step-sizes. Unfortunately, these convergence results do not directly apply to the adaptive vector-valued step-sizes of GAMP.

The second part of the theorem, equation (21), shows that if the algorithm converges then its fixed points will be critical points of the constrained optimization (1). This part of the theorem can be considered as a generalization of [52, Proposition 7.1], which considers quadratic f_z, and of [47, Proposition 5.1], which considers quadratic f_z and $f_x(\mathbf{x}) = \|\mathbf{x}\|_1$.

The third part of Theorem 1 then shows that the quadratic term $\boldsymbol{\tau}_x$ can be interpreted as an "approximate diagonal" of the inverse Hessian under the large random matrix model described in Appendix A.

Finally, it is useful to compare the fixed-points of GAMP with those of standard BP. A classic result of [53] shows that any fixed point for standard max-sum loopy BP is locally optimal in the sense that one cannot improve the objective function by perturbing the solution on any set of components whose variables belong to a subgraph that contains at most one cycle. In particular, if the overall graph is acyclic, any fixed-point of standard max-sum loopy BP is globally optimal. Also, for any graph, the objective function cannot be reduced by changing any individual component. The local optimality for GAMP provided by Theorem 1 is weaker than that for max-sum loopy BP in that GAMP's fixed-points only satisfy first-order conditions for saddle points of the Lagrangian. This implies that, even an individual component may only be locally optimal.

IV. Fixed-Points of Sum-Product GAMP

A. Bethe Free Energy

A classic result in graphical models is that the fixed points of loopy BP can be interpreted as critical points in the constrained minimization of a energy function known as the Bethe Free energy (BFE) [54], [55]. In this section, we will show that sum-product GAMP has a similar energy function interpretation.

Specifically, consider a set of scalar densities

$$b_{x_j}(x_j), \quad b_{z_i}(z_i), \quad q_{z_i}(z_i), \tag{23}$$

where the densities $q_{z_i}(z_i)$ are Gaussian. Given any such set, define the product densities

$$b_x(\mathbf{x}) = \prod_{j=1}^{n} b_{x_j}(x_j), \quad b_z(\mathbf{z}) = \prod_{i=1}^{m} b_{z_i}(z_i) \tag{24a}$$

$$q_z(\mathbf{z}) = \prod_{i=1}^{m} q_{z_i}(z_i), \tag{24b}$$

and the energy function

$$\begin{aligned} J_{\mathrm{SP}}(b_x, b_z, q_z) := \; & D(b_x \| e^{-f_x}) + D(b_z \| e^{-f_z}) \\ & + D(b_z \| q_z) + H(b_z), \end{aligned} \tag{25}$$

where $H(b_z)$ is the differential entropy. With these definitions, consider the constrained minimization

$$\begin{aligned} \min_{b_x, b_z, q_z} \; & J_{\mathrm{SP}}(b_x, b_z, q_z) \\ \text{s.t.} \quad & \mathbb{E}(\mathbf{z}|b_z) = \mathbb{E}(\mathbf{z}|q_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x) \\ & \boldsymbol{\tau}_p = \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x), \quad \mathbf{S} = \mathbf{A}.\mathbf{A} \\ & q_z(\mathbf{z}) \sim \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_p, \mathrm{Diag}(\boldsymbol{\tau}_p)\right), \end{aligned} \tag{26}$$

Here and below, we use $\mathbb{E}(\mathbf{x}|b_x)$ to denote the expected value of $\mathbf{x} \sim b_x$, and similar for $\mathbb{E}(\mathbf{z}|b_z)$. Also, we use $\mathrm{var}(\mathbf{x}|b_x)$ to denote the vector whose jth component is the variance of $x_j \sim b_{x_j}$, and similar for $\mathrm{var}(\mathbf{z}|b_z)$. We stress that $\mathrm{var}(\mathbf{x}|b_x)$ is a vector, not a covariance matrix. Note also that the last constraint in (26) simply states that q_z must be Gaussian with independent components.

Note that since

$$D(b_z \| q_z) + H(b_z) = -\mathbb{E}\left[\log q_z(\mathbf{z}) \mid b_z\right],$$

the objective function (25) is separately convex in (b_x, b_z) and q_z. However, it is not, in general, jointly convex in all three densities. Also, the final two constraints in the optimization (26), on the variances and Gaussianity of q_z, are also not convex.

Our main result, Theorem 2 below, shows that sum-product GAMP can be interpreted as a method to approximately minimize this non-convex energy function. This result was first stated in the conference version of this paper [1]. Since the publication of that paper, it was stated in [41] that, in the case of additive white Gaussian noise (AWGN) output channels, the constrained optimization (26) can be interpreted as an approximation of the Bethe Free energy optimization that is valid when (a) the matrix $\mathbf{A}$ has i.i.d. zero mean entries and $m, n \to \infty$, and (b) the standard marginalization constraints in the BFE optimization are replaced by matching constraints on the first and second moments. A subsequent work [42] derived a similar approximate BFE optimization for arbitrary output channels and matrix uncertainties. We will not discuss the BFE interpretation in this work; the reader is referred to [41] and [42]. However, in recognition of the relation to the Bethe free energy minimization, we will call the energy function (25) the large-system-limit Bethe Free

energy (LSL-BFE) and call the constrained minimization (26) the LSL-BFE optimization.

B. GAMP Optimization

To relate the LSL-BFE optimization (26) to sum-product GAMP, we first rewrite the optimization to remove the minima over q_z. Given a density $b_z(\mathbf{z})$, define the function

$$H_{\text{gauss}}(b_z, \boldsymbol{\tau}_p) := D(b_z \| q_z) + H(b_z), \quad q_z(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_p, \text{Diag}(\boldsymbol{\tau}_p)), \quad \boldsymbol{\mu}_p = \mathbb{E}(\mathbf{z}|b_z). \tag{27}$$

This function is simply the last two terms of $J_{\text{SP}}(b_x, b_z, q_z)$ in (25) with $q_z(\mathbf{z})$ being the Gaussian density with mean $\boldsymbol{\mu}_p = \mathbb{E}(\mathbf{z}|b_z)$ and variance $\text{var}(\mathbf{z}|q_z) = \boldsymbol{\tau}_p$. It can be calculated that

$$H_{\text{gauss}}(b_z, \boldsymbol{\tau}_p) = \frac{1}{2}\sum_{i=1}^{m}\left[\frac{\text{var}(z_i|b_{z_i})}{\tau_{p_i}} + \log(2\pi\,\tau_{p_i})\right]. \tag{28}$$

Note that from (27), $H_{\text{gauss}}(b_z, \boldsymbol{\tau}_p) \geq H(b_z)$ for all $\boldsymbol{\tau}_p$ with equality when b_z is itself Gaussian with variance $\text{var}(\mathbf{z}|b_z) = \boldsymbol{\tau}_p$. Hence, we will call $H_{\text{gauss}}(b_z, \boldsymbol{\tau}_p)$ the *Gaussian entropy upper bound* function. Using this upper bound function, we can replace the minimization over Gaussian q_z in (26) with an optimization over the vector of variances $\boldsymbol{\tau}_p$. This results in the equivalent optimization

$$\begin{aligned} \min_{b_x, b_z, \boldsymbol{\tau}_p} \quad & J_{\text{SP}}(b_x, b_z, \boldsymbol{\tau}_p) \\ \text{s.t.} \quad & \mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x) \\ & \boldsymbol{\tau}_p = \mathbf{S}\,\text{var}(\mathbf{x}|b_x), \quad \mathbf{S} = \mathbf{A}.\mathbf{A} \end{aligned} \tag{29}$$

where the objective function is

$$J_{\text{SP}}(b_x, b_z, \boldsymbol{\tau}_p) := D(b_x \| e^{-f_x}) + D(b_z \| e^{-f_z}) + H_{\text{gauss}}(b_z, \boldsymbol{\tau}_p). \tag{30}$$

With some abuse of notation, we have used $J_{\text{SP}}(\cdot)$ to denote both the LSL-BFE function in terms of q_z as in (25) and the function in terms of the variance vector $\boldsymbol{\tau}_p$ as given in (30).

Corresponding to (29), define the Lagrangian

$$L_{\text{SP}}(b_x, b_z, \boldsymbol{\tau}_p, \mathbf{s}) = J_{\text{SP}}(b_x, b_z, \boldsymbol{\tau}_p) + \mathbf{s}^T(\mathbb{E}(\mathbf{z}|b_z) - \mathbf{A}\mathbb{E}(\mathbf{x}|b_x)), \tag{31}$$

where $\mathbf{s}$ represents a vector of dual parameters. Note that this Lagrangian does *not* include the constraint $\boldsymbol{\tau}_p = \mathbf{S}\,\text{var}(\mathbf{x}|b_x)$; we will handle that separately. We can now state the main result.

Theorem 2: Consider the outputs of the sum-product GAMP version of Algorithm 1, and define the densities

$$b_x^{t+1}(\mathbf{x}) = p(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t), \quad b_z^t(\mathbf{z}) = p(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t), \tag{32}$$

where $p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r)$ and $p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p)$ are given in (13)*. Then, the GAMP algorithm input node update satisfies*

$$b_x^{t+1} = \arg\min_{b_x}\left[L_{\text{SP}}(b_x, b_z^t, \boldsymbol{\tau}_p^t, \mathbf{s}^t) + \frac{1}{2}(\boldsymbol{\tau}_s^t)^T\mathbf{S}\,\text{var}(\mathbf{x}|b_x) + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbb{E}(\mathbf{x}|b_x^t)\right\|_{\boldsymbol{\tau}_r^t}^2\right]. \tag{33}$$

where $L_{\text{SP}}(\mathbf{x}, \mathbf{z}, \mathbf{s})$ is the Lagrangian in (31)*. Similarly, the steps in the output node update for the GAMP algorithm are equivalent to:*

$$\boldsymbol{\tau}_p^t = \mathbf{S}\,\text{var}(\mathbf{x}|b_x^t), \tag{34a}$$

$$b_z^t = \arg\min_{b_z}\left[L_{\text{SP}}(b_x^t, b_z, \boldsymbol{\tau}_p^t, \mathbf{s}^{t-1}) + \frac{1}{2}\left\|\mathbb{E}(\mathbf{z}|b_z) - \mathbf{A}\mathbb{E}(\mathbf{x}|b_x^t)\right\|_{\boldsymbol{\tau}_p^t}^2\right], \tag{34b}$$

$$\mathbf{s}^t = \mathbf{s}^{t-1} + \frac{1}{\boldsymbol{\tau}_p^t}\left[\mathbb{E}(\mathbf{z}|b_z^t) - \mathbf{A}\mathbb{E}(\mathbf{x}|b_x^t)\right], \tag{34c}$$

$$\boldsymbol{\tau}_s^t = 2\nabla_{\boldsymbol{\tau}_p} L_{\text{SP}}(b_x^t, b_z^t, \boldsymbol{\tau}_p^t, \mathbf{s}^t). \tag{34d}$$

Moreover, any fixed point of the sum-product GAMP algorithm is a critical point of the constrained optimization (29)*.*

Proof: See Appendix C. ∎

Theorem 2 exposes connections between sum-product GAMP and both the ISTA and ADMM methods described earlier. The minimizations over b_x and b_z and the update of the dual parameters $\mathbf{s}^t$ in (33), (34b) and (34c) follow the format of the ADMM minimizations in Algorithm 3 for certain choices of the auxiliary functions. On the other hand, the role of $\boldsymbol{\tau}_s^t$ in (33) and (34d) follows the gradient-based method of the generalized ISTA method in Algorithm 2 for the constraint $\boldsymbol{\tau}_s = \mathbf{S}\,\text{var}(\mathbf{x}|b_x)$. So, the sum-product GAMP algorithm can be seen as a hybrid of the ISTA and ADMM methods for the optimization problem (29).

Unfortunately, this hybrid ISTA-ADMM method is non-standard and we are not aware of existing convergence theory. However, Theorem 2 at least shows that, if the sum-product GAMP algorithm converges, then its fixed points correspond to critical points of the optimization problem (29).

V. Conclusions

Although AMP methods admit precise analyses in the context of large i.i.d. transform matrices $\mathbf{A}$, their behavior for general matrices is less well-understood. This limitation is unfortunate since many transforms arising in practical problems such as imaging and regression are not well-modeled as realizations of large i.i.d. matrices. To help overcome these limitations, this paper draws connections between AMP and certain variants of standard optimization methods that employ adaptive vector-valued step-sizes. These connections enable a precise characterization of the fixed-points of both max-sum and sum-product GAMP for the case of arbitrary transform matrices $\mathbf{A}$.

However, much work remains to be done. Most importantly, while our results relate GAMP to standard optimization methods, these do not guarantee the algorithm's convergence. As mentioned in the Introduction, for general $\mathbf{A}$, it is well-known that GAMP methods may diverge [30], [31]. Several recent modifications have been proposed to improve the stability of GAMP, including damping [30], [37]. One potential line of future work is to consider alternates to GAMP that are based on direct minimization of the energy function. Some preliminary works in this regard have been presented in [38], which proposes a coordinate descent method and [39], which uses an ADMM-based method.

GAMP-based methods have also been extended in a wide variety of ways, such as combining EM with GAMP [56]–[59], turbo and hybrid GAMP methods [60], [61], applications in dictionary learning and matrix factorization [62]–[66], and applications in blind deconvolution and self-calibration [67]. Another line of work would be to understand if one can find free energy and optimization interpretations of these algorithms. For dictionary learning and matrix factorization some initial work has appeared in [42] and [65].

ACKNOWLEDGMENTS

The authors would like to thank Ulugbek Kamilov and Vivek K. Goyal for their valuable comments.

APPENDIX A
APPROXIMATE DIAGONALS

Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and positive vectors $\mathbf{d}_x \in \mathbb{R}^n$ and $\mathbf{d}_z$, consider the positive matrices (18). We analyze the asymptotic behavior of these matrices under the following assumptions:

Assumption 1: Consider a sequence of matrices $\mathbf{Q}_x$ *and* $\mathbf{Q}_z$ *of the form* (18)*, indexed by the dimension n satisfying:*

(a) *The dimension m is a deterministic function of n with* $\lim_{n\to\infty} m/n = \beta$ *for some* $\beta > 0$,

(b) *The positive vectors* $\mathbf{d}_x$ *and* $\mathbf{d}_z$ *are deterministic vectors with*

$$\limsup_{n\to\infty} \|\mathbf{d}_x\|_\infty < \infty, \quad \limsup_{n\to\infty} \|\mathbf{d}_z\|_\infty < \infty.$$

(c) *The components of* $\mathbf{A}$ *are independent, zero-mean with* $\mathrm{var}(A_{ij}) = S_{ij}$ *for some deterministic matrix* $\mathbf{S}$ *such that*

$$\limsup_n \max_{i,j} nS_{ij} < \infty.$$

Theorem 3 ([68]): Consider a sequence of matrices $\mathbf{Q}_x$ *and* $\mathbf{Q}_z$ *in Assumption 1. Then, for each n, there exists positive vectors* $\boldsymbol{\xi}_x$ *and* $\boldsymbol{\xi}_z$ *satisfying the nonlinear equations*

$$\mathbf{1}./\boldsymbol{\xi}_z = \mathbf{1}./\mathbf{d}_z + \mathbf{S}\boldsymbol{\xi}_x, \quad \mathbf{1}./\boldsymbol{\xi}_x = \mathbf{1}./\mathbf{d}_x + \mathbf{S}^T\boldsymbol{\xi}_z, \tag{35}$$

where the vector inverses are componentwise. Moreover, the vectors $\boldsymbol{\xi}_z$ *and* $\boldsymbol{\xi}_x$ *are asymptotic diagonals of* $\mathbf{Q}_x$ *and* $\mathbf{Q}_z$ *in the following sense: For any deterministic sequence of positive vectors* $\mathbf{u}_x \in \mathbb{R}^n$ *and* $\mathbf{u}_z \in \mathbb{R}^m$*, such that*

$$\limsup_{n\to\infty} \|\mathbf{u}_x\|_\infty < \infty, \quad \limsup_{n\to\infty} \|\mathbf{u}_z\|_\infty < \infty,$$

the following limits hold almost surely

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \left[u_{xj}((Q_x)_{jj} - \xi_{xj})\right] = 0$$

$$\lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^{m} \left[u_{zi}((Q_z)_{ii} - \xi_{zi})\right] = 0.$$

Proof: This result is a special case of the results in [68]. ∎

The result says that, for certain large random matrices $\mathbf{A}$, $\boldsymbol{\xi}_x$ and $\boldsymbol{\xi}_z$ are approximate diagonals of the matrices $\mathbf{Q}_x$ and $\mathbf{Q}_z$, respectively. This motivates the following definition for deterministic $\mathbf{A}$.

Definition 1: Consider matrices $\mathbf{Q}_x$ *and* $\mathbf{Q}_z$ *of the form* (18) *for some* deterministic *(i.e., non-random)* $\mathbf{A}$*,* $\mathbf{d}_x$ *and* $\mathbf{d}_z$*. Let* $\mathbf{S} = \mathbf{A}.\mathbf{A}$ *be the componentwise square of* $\mathbf{A}$*. Then, the unique positive solutions* $\boldsymbol{\xi}_z$ *and* $\boldsymbol{\xi}_x$ *to* (35) *will be called the* approximate diagonals *of* $\mathbf{Q}_z$ *and* $\mathbf{Q}_x$*, respectively.*

APPENDIX B
PROOF OF THEOREM 1

To prove (20b), observe that

$$\begin{aligned}\arg\min_{\mathbf{z}}&\left[L(\mathbf{x}^t, \mathbf{z}, \mathbf{s}^{t-1}) + \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\right]\\ &\overset{(a)}{=} \arg\min_{\mathbf{z}}\left[f_z(\mathbf{z}) + (\mathbf{s}^{t-1})^T\mathbf{z} + \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\right]\\ &\overset{(b)}{=} \arg\min_{\mathbf{z}}\left[f_z(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{p}^t\|^2_{\boldsymbol{\tau}_p^t}\right] \overset{(c)}{=} \mathbf{z}^t,\end{aligned}$$

where (a) follows from substituting (2) and (14) into (20b) and eliminating the terms that do not depend on $\mathbf{z}$; (b) follows from the definition of $\mathbf{p}^t$ in line 8; and (c) follows from the definition of $\mathbf{z}^t$ in line 10. This proves (20b). The update (20a) can be proven similarly. To prove (20c), observe that

$$\mathbf{s}^t \overset{(a)}{=} (\mathbf{z}^t - \mathbf{p}^t)./\boldsymbol{\tau}_p^t \quad \overset{(b)}{=} \mathbf{s}^{t-1} + (\mathbf{z}^t - \mathbf{A}\mathbf{x}^t)./\boldsymbol{\tau}_p^t,$$

where (a) follows from the update of $\mathbf{s}^t$ in line 16 in Algorithm 1 (recall that the division is componentwise); and (b) follows from the update for $\mathbf{p}^t$ in line 8. We have thus proven the equivalence of the max-sum GAMP algorithm with the Lagrangian updates (20).

Now consider any fixed point $(\widehat{\mathbf{z}}, \widehat{\mathbf{x}}, \mathbf{s})$ of the max-sum GAMP algorithm. A fixed point of (20c) requires that

$$\widehat{\mathbf{z}} = \mathbf{A}\widehat{\mathbf{x}}, \tag{36}$$

so the fixed point satisfies the constraint of the optimization (1). Now, using (36) and the fact that $\widehat{\mathbf{z}}$ is the minima of (20b), we have that

$$\nabla_{\mathbf{z}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}.$$

Similarly, since $\mathbf{x}$ is the minima of (20a), we have that

$$\nabla_{\mathbf{z}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}.$$

Thus, the fixed point $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s})$ is a critical point of the Lagrangian (14).

Finally, consider the quadratic terms $(\boldsymbol{\tau}_x, \boldsymbol{\tau}_r, \boldsymbol{\tau}_s)$ at the fixed point. From the updates of $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_r$ in Algorithm 1 [see also (7)] and the definition of $\mathbf{d}_x$ in (19), we obtain

$$\mathbf{1}./\boldsymbol{\tau}_x = \mathbf{d}_x + \mathbf{1}./\boldsymbol{\tau}_r = \mathbf{d}_x + \mathbf{S}^T\boldsymbol{\tau}_s. \tag{37}$$

Similarly, the updates of $\boldsymbol{\tau}_s$ and $\boldsymbol{\tau}_p$ show that

$$\mathbf{1}./\boldsymbol{\tau}_s = \mathbf{1}./\mathbf{d}_z + \boldsymbol{\tau}_p = \mathbf{1}./\mathbf{d}_z + \mathbf{S}\boldsymbol{\tau}_x. \tag{38}$$

Then, according to Definition 1, $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_s$ are the approximate diagonals of $\mathbf{Q}_x$ and $\mathbf{Q}_z$ in (18), respectively.

APPENDIX C
PROOF OF THEOREM 2

We prove this theorem in two parts. First we show that the sum-product GAMP updates are equivalent to (33) and (34). Then we show that any fixed points of these updates are critical points of the constrained optimization (29).

A. Equivalence of the Updates

We begin by proving (33). Define b_x^{t+1} as the solution to the minimization (33). So, we must show that this solution is given by the equation for $b_x^{t+1}(\mathbf{x})$ in (32). We use induction: Suppose that b_x^{t+1} in (32) is the solution to (33) for some t. We will then show that it is the solution for $t+1$.

First, combining the induction hypothesis that b_x^{t+1} is given in (32) with lines 26 and 27 of Algorithm 1, we have

$$\mathbf{x}^t = \mathbb{E}(\mathbf{x}|b_x^t), \quad \boldsymbol{\tau}_x^t = \mathrm{var}(\mathbf{x}|b_x^t). \tag{39}$$

That is, $\mathbf{x}^t$ and $\boldsymbol{\tau}_x^t$ are the mean and variance vectors of the density b_x^t. We next simplify the right hand side of (33) to remove terms that do not depend on b_x:

$$\begin{aligned}
&L_{\mathrm{SP}}(b_x, b_z^t, \boldsymbol{\tau}_p^t, \mathbf{s}^t) + \frac{1}{2}(\boldsymbol{\tau}_s^t)^T \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x) \\
&\quad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbb{E}(\mathbf{x}|b_x^t)\right\|_{\boldsymbol{\tau}_r^t}^2 \\
&\overset{(a)}{=} D(b_x\|e^{-f_x}) - (\mathbf{s}^t)^T\mathbf{A}\mathbb{E}(\mathbf{x}|b_x) + \frac{1}{2}(\boldsymbol{\tau}_s^t)^T\mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x) \\
&\quad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbb{E}(\mathbf{x}|b_x^t)\right\|_{\boldsymbol{\tau}_r^t}^2 + \mathrm{const} \\
&\overset{(b)}{=} D(b_x\|e^{-f_x}) - (\mathbf{s}^t)^T\mathbf{A}\mathbb{E}(\mathbf{x}|b_x) + \left(\frac{1}{2\boldsymbol{\tau}_r^t}\right)^T \mathrm{var}(\mathbf{x}|b_x) \\
&\quad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbf{x}^t\right\|_{\boldsymbol{\tau}_r^t}^2 + \mathrm{const} \qquad (40) \\
&\overset{(c)}{=} D(b_x\|e^{-f_x}) + \left(\frac{1}{2\boldsymbol{\tau}_r^t}\right)^T \mathrm{var}(\mathbf{x}|b_x) \\
&\quad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbf{r}^t\right\|_{\boldsymbol{\tau}_r^t}^2 + \mathrm{const} \\
&\overset{(d)}{=} D(b_x\|e^{-f_x}) + \frac{1}{2}\mathbb{E}\left(\|\mathbf{x}-\mathbf{r}^t\|_{\boldsymbol{\tau}_r^t}^2 \,\middle|\, b_x\right) + \mathrm{const}, \qquad (41)
\end{aligned}$$

where in all the steps "const" denotes any terms that do not depend on b_x, and (a) follows from the definition of the Lagrangian (31) and the objective function (30); (b) follows from (39) and the fact that $\boldsymbol{\tau}_r^t = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s^t)$ in line 20 of Algorithm 1; (c) follows from the definition of $\mathbf{r}^t$ in line 21; and finally (d) follows from the simplification:

$$\begin{aligned}
&\left(\mathbf{1}./\boldsymbol{\tau}_r^t\right)^T \mathrm{var}(\mathbf{x}|b_x) + \left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbf{r}^t\right\|_{\boldsymbol{\tau}_r^t}^2 \\
&\quad = \sum_{j=1}^n \left[\frac{1}{\tau_{r_j^t}}\Big(\mathrm{var}(x_j|b_{x_j}) + (\mathbb{E}(x_j|b_{x_j}) - r_j^t)^2\Big)\right] \\
&\quad = \sum_{j=1}^n \left[\frac{1}{\tau_{r_j}^t}\Big(\mathbb{E}(x_j^2|b_{x_j}) - 2r_j^t\mathbb{E}(x_j|b_{x_j})\Big)\right] + \mathrm{const} \\
&\quad = \mathbb{E}\left(\|\mathbf{x}-\mathbf{r}^t\|_{\boldsymbol{\tau}_r^t}^2 \,\middle|\, b_x\right) + \mathrm{const}.
\end{aligned}$$

Substituting (41) into (33), and using the definition of $p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r)$ in (13),

$$\begin{aligned}
b_x^{t+1} &= \arg\min_{b_x} D(b_x\|e^{-f_x}) + \frac{1}{2}\mathbb{E}\left(\|\mathbf{x}-\mathbf{r}^t\|_{\boldsymbol{\tau}_r^t}^2 \,\middle|\, b_x\right) \\
&= \arg\min_{b_x} -H(b_x) + \mathbb{E}\left(f_x(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{r}^t\|_{\boldsymbol{\tau}_r^t}^2 \,\middle|\, b_x\right) \\
&= \arg\min_{b_x} -H(b_x) - \mathbb{E}\left(\log p(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t) \mid b_x\right) \\
&= \arg\min_{b_x} D\left(b_x \parallel p(\cdot|\mathbf{r}^t, \boldsymbol{\tau}_r^t)\right), \qquad (42)
\end{aligned}$$

which proves that b_x^{t+1} satisfies (32).

Similarly, one can show that the solution b_z^t in (34b) is given by (32). In addition, $\mathbf{z}^t$ and $\boldsymbol{\tau}_z^t$ in lines 13 and 14 of Algorithm 1 are the mean and variances of the estimated densities,

$$\mathbf{z}^t = \mathbb{E}(\mathbf{z}|b_z^t), \quad \boldsymbol{\tau}_z^t = \mathrm{var}(\mathbf{z}|b_z^t). \tag{43}$$

Equation (34a) follows directly from line 7 and (39). Also, combining lines 8 and 16, we obtain (34c).

Finally, to prove (34d), we take the derivatives

$$\begin{aligned}
&\nabla_{\boldsymbol{\tau}_p} L_{\mathrm{SP}}(b_x^t, b_z^t, \boldsymbol{\tau}_p^t, \mathbf{s}^t) \\
&\quad \overset{(a)}{=} \nabla_{\boldsymbol{\tau}_p} H_{\mathrm{gauss}}(b_z^t, \boldsymbol{\tau}_p^t) \overset{(b)}{=} \frac{1}{2}\Big[\mathbf{1}./\boldsymbol{\tau}_p^t - \boldsymbol{\tau}_z^t./(\boldsymbol{\tau}_p^t.\boldsymbol{\tau}_p^t)\Big] \\
&\quad \overset{(c)}{=} \frac{1}{2}\boldsymbol{\tau}_s^t,
\end{aligned}$$

where (a) follows from removing the terms in (31) that do not depend on $\boldsymbol{\tau}_p$; (b) can be verified by simply taking the derivative of H_{gauss} in (28) with respect to each component τ_{p_i}; and (c) follows from the definition of $\boldsymbol{\tau}_s^t$ in line 17 of Algorithm 1. This proves (34d), and we have established that the sum-product GAMP updates are equivalent to (33) and (34).

B. Characterization of the Fixed Points

First, by substituting the constraint $\boldsymbol{\tau}_p = \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x)$, we can rewrite the optimization (29) as

$$\begin{aligned}
\min_{b_x, b_z} \quad & J_{\mathrm{SP}}(b_x, b_z, \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x)) \\
\text{s.t.} \quad & \mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x).
\end{aligned} \tag{44}$$

Corresponding to this optimization, define the Lagrangian

$$\begin{aligned}
\widetilde{L}_{\mathrm{SP}}(b_x, b_z, \mathbf{s}) = {} & J_{\mathrm{SP}}(b_x, b_z, \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x)) \\
& + \mathbf{s}^T\left(\mathbb{E}(\mathbf{z}|b_x) - \mathbf{A}\mathbb{E}(\mathbf{x}|b_x)\right),
\end{aligned} \tag{45}$$

where $\mathbf{s}$ are the dual parameters. Now, let $(\widehat{b}_x, \widehat{b}_z, \mathbf{s})$ be any fixed points of the updates (33) and (34). To show that $(\widehat{b}_x, \widehat{b}_z)$ are critical points of the optimization (44), we need to show that they satisfy the constraint $\mathbb{E}(\mathbf{z}|\widehat{b}_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|\widehat{b}_x)$ and that $(\widehat{b}_x, \widehat{b}_z)$ are stationary points of the Lagrangian $\widetilde{L}_{\mathrm{SP}}(b_x, b_z, \mathbf{s})$.

From (34c), we have that, at any fixed point $(\widehat{b}_x, \widehat{b}_z)$

$$\mathbb{E}(\mathbf{z}|\widehat{b}_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|\widehat{b}_x), \tag{46}$$

and so the linear constraint is satisfied.

To show that $(\widehat{b}_x, \widehat{b}_z)$ are stationary points of the Lagrangian, we introduce the following notation: suppose that

$V(b)$ is a scalar-valued or vector-valued functional of a density $b(\mathbf{u})$, and that $\Delta b(\mathbf{u})$ is a perturbation direction of that density. That is, $\Delta b(\mathbf{u})$ is in the tangent plane of the set of densities, so thatÂ $\int \Delta b(\mathbf{u})\,\mathrm{d}\mathbf{u} = 0$ and $\Delta b(\mathbf{u}) = 0$ when $b_0(\mathbf{u}) = 0$. We denote the differential of the functional $V(b)$ in the direction Δb at the point $b = b_0$ by

$$\left.\frac{\partial V(b)}{\partial b}\right|_{b=b_0} \bullet \Delta b = \lim_{\epsilon\to 0} \frac{1}{\epsilon}\left[V\left(b_0 + \epsilon\,\Delta b\right) - V(b_0)\right],$$

which is defined when the limit exists. See [69] for a complete treatment of differentials of functionals. Using this notation, we need to show that

$$\left.\frac{\partial}{\partial b_x}\widetilde{L}_{\mathrm{SP}}(b_x, \widehat{b}_z, \mathbf{s})\right|_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0, \tag{47a}$$

$$\left.\frac{\partial}{\partial b_z}\widetilde{L}_{\mathrm{SP}}(\widehat{b}_x, b_z, \mathbf{s})\right|_{b_z=\widehat{b}_z} \bullet \Delta b_z = 0, \tag{47b}$$

for all perturbation directions Δb_x and Δb_z.

To prove (47a), first note that, for any Δb_x, the partial derivative of the augmenting term in (33) is given by

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial b_x}&\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbb{E}(\mathbf{x}|\widehat{b}_x)\right\|^2_{\boldsymbol{\tau}_r}\Big|_{b_x=\widehat{b}_x} \bullet \Delta b_x \\ &= \left(\mathbb{E}(\mathbf{x}|\widehat{b}_x) - \mathbb{E}(\mathbf{x}|\widehat{b}_x)\right)^T \mathrm{Diag}(\boldsymbol{\tau}_r)^{-1} \\ &\quad\times \frac{\partial}{\partial b_x}\mathbb{E}(\mathbf{x}|\widehat{b}_x) \bullet \Delta b_x = 0.\end{aligned} \tag{48}$$

Also, since $\widehat{b}_x$ is a minima of (33), it is a stationary point of the function. Hence, for any perturbation direction Δb_x,

$$\begin{aligned}&\frac{\partial}{\partial b_x}\Big[L_{\mathrm{SP}}(b_x, \widehat{b}_z, \boldsymbol{\tau}_p, \mathbf{s}) + \frac{1}{2}(\boldsymbol{\tau}_s)^T \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x) \\ &\qquad + \frac{1}{2}\left\|\mathbb{E}(\mathbf{x}|b_x) - \mathbb{E}(\mathbf{x}|\widehat{b}_x)\right\|^2_{\boldsymbol{\tau}_r}\Big]_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(a)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}\Big[L_{\mathrm{SP}}(b_x, \widehat{b}_z, \boldsymbol{\tau}_p, \mathbf{s}) \\ &\qquad + \frac{1}{2}(\boldsymbol{\tau}_s)^T \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x)\Big]_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(b)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}\Big[L_{\mathrm{SP}}(b_x, \widehat{b}_z, \boldsymbol{\tau}_p, \mathbf{s}) \\ &\qquad + \frac{\partial}{\partial \boldsymbol{\tau}_p}L_{\mathrm{SP}}(\widehat{b}_x, \widehat{b}_z, \boldsymbol{\tau}_p, \widehat{\mathbf{s}})^T \boldsymbol{\tau}_p(b_x)\Big]_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(c)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}\Big[L_{\mathrm{SP}}(b_x, \widehat{b}_z, \boldsymbol{\tau}_p, \mathbf{s})\Big]_{b_x=\widehat{b}_x} \bullet \Delta b_x \\ &\qquad + \frac{\partial}{\partial \boldsymbol{\tau}_p}L_{\mathrm{SP}}(\widehat{b}_x, \widehat{b}_z, \boldsymbol{\tau}_p, \widehat{\mathbf{s}})^T \frac{\partial}{\partial b_x}\boldsymbol{\tau}_p(b_x)\Big|_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(d)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}L_{\mathrm{SP}}(b_x, \widehat{b}_z, \boldsymbol{\tau}_p(b_x), \mathbf{s})\Big|_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(e)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}L_{\mathrm{SP}}(b_x, \widehat{b}_z, \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x), \mathbf{s})\Big|_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0 \\ \overset{(f)}{\Longleftrightarrow}\ &\frac{\partial}{\partial b_x}\widetilde{L}_{\mathrm{SP}}(b_x, \widehat{b}_z, \mathbf{s})\Big|_{b_x=\widehat{b}_x} \bullet \Delta b_x = 0,\end{aligned} \tag{49}$$

where (a) follows from (48); (b) follows from the fixed points (34a) and (34d) and the clarifying notation $\boldsymbol{\tau}_p(b_x) = \mathbf{S}\,\mathrm{var}(\mathbf{x}|b_x)$; (c) follows from straightforward calculus; (d) follows from the multivariable chain rule; (e) follows from the definition of $\boldsymbol{\tau}_p(b_x)$; and (f) follows from the definition of the modified Lagrangian in (45). This proves (47a). The proof of (47b) is similar.

REFERENCES

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

[2] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *J. Roy. Stat. Soc. Ser. A*, vol. 135, no. 3, pp. 370–384, 1972.

[3] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London, U.K.: Chapman & Hall, 1989.

[4] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1902–1923, Mar. 2012.

[5] Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications*. New York, NY, USA: Cambridge Univ. Press, 2012.

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

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

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

[9] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[10] Y. Nesterov, "Gradient methods for minimizing composite objective function," Center Oper. Res. Econometrics (CORE), Catholic Univ. Louvain, Louvain-la-Neuve, Belgium, CORE Discussion Paper 2007/76, 2007.

[11] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[13] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 5, no. 1, pp. 293–318, Apr. 1992.

[14] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[15] X. Zhang, M. Burger, and S. Osher, "A unified primal-dual algorithm framework based on Bregman iteration," *SIAM J. Sci. Comput.*, vol. 46, no. 1, pp. 20–46, Jan. 2011.

[16] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[17] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control Optim.*, vol. 29, no. 1, pp. 119–138, Jan. 1991.

[18] M. Zhu and T. Chan, "An efficient primal-dual hybrid gradient algorithm for total variation image restoration," Dept. Comput. Appl. Math., Univ. California, Los Angeles, Los Angeles, CA, USA, Tech. Rep. 08-34, 2008.

[19] J. E. Esser, "Primal dual algorithms for convex models and applications to image restoration, registration and nonlocal inpainting," Ph.D. dissertation, Dept. Math., Univ. California, Los Angeles, Los Angeles, CA, USA, 2010.

[20] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[21] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective," *SIAM J. Imag. Sci.*, vol. 5, no. 1, pp. 119–149, 2012.

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

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

[24] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: II. Analysis and validation," in *Proc. IEEE Inf. Theory Workshop*, Jan. 2010, pp. 1–5.

[25] 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.

[26] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 2010, pp. 1–6.

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

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

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

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

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

[32] A. K. Fletcher, S. Rangan, L. R. Varshney, and A. Bhargava, "Neural reconstruction with approximate message passing (NeuRAMP)," in *Proc. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 2555–2563.

[33] S. Y. Chen, H. Tong, Z. Wang, S. Liu, M. Li, and B. Zhang, "Improved generalized belief propagation for vision processing," *Math. Problems Eng.*, vol. 2011, Dec. 2011, Art. no. 416963, doi: 10.1155/2011/416963.

[34] U. S. Kamilov, V. K. Goyal, and S. Rangan, "Message-passing dequantization with applications to compressed sensing," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6270–6281, Dec. 2012.

[35] J. Vila, P. Schniter, and J. Meola, "Hyperspectral imaging via turbo bilinear approximate message passing," *IEEE Trans. Comput. Imag.*, vol. 1, no. 3, pp. 143–158, Sep. 2015.

[36] A. K. Fletcher and S. Rangan, "Scalable inference for neuronal connectivity from calcium imaging," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2843–2851.

[37] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *Proc. IEEE ICASSP*, Apr. 2015, pp. 2021–2025.

[38] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Sparse estimation with the swept approximated message-passing algorithm," in *Proc. ICML*, 2015, pp. 1123–1132.

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

[40] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Inf. Inference*, 2013, vol. 2, no. 2, pp. 115–144, doi: 10.1093/imaiai/iat004.

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

[42] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová. (2014). "Phase transitions and sample complexity in Bayes-optimal matrix factorization." [Online]. Available: https://arxiv.org/abs/1402.1298

[43] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, Jan. 2008.

[44] 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.

[45] T. Tanaka and M. Okada, "Approximate belief propagation, density evolution, and statistical neurodynamics for CDMA multiuser detection," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 700–706, Feb. 2005.

[46] D. Guo and C.-C. Wang, "Asymptotic mean-square optimality of belief propagation for sparse linear systems," in *Proc. IEEE Inf. Theory Workshop*, Chengdu, China, Oct. 2006, pp. 194–198.

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

[48] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Dept. of Comp. Sci. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

[49] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, Jun. 2008.

[50] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[51] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.

[52] D. L. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3396–3433, Jun. 2013.

[53] Y. Weiss and W. T. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 736–744, Feb. 2001.

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

[55] T. Heskes, "Stable fixed points of loopy belief propagation are local minima of the Bethe free energy," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2003, pp. 343–350.

[56] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical-physics-based reconstruction in compressed sensing," *Phys. Rev. X*, vol. 2, no. 2, p. 021005, 2012.

[57] 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.

[58] J. P. Vila and P. Schniter, "An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals," *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4689–4703, Sep. 2014.

[59] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2969–2985, May 2014.

[60] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a Markov-tree prior," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3439–3448, Jul. 2012.

[61] S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid generalized approximate message passing with applications to structured sparsity," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 1241–1245.

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

[63] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Inf. Theory*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

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

[65] F. Krzakala, M. Mézard, and L. Zdeborová, "Phase diagram and approximate message passing for blind calibration and dictionary learning," in *Proc. IEEE ISIT*, Jul. 2013, pp. 659–663.

[66] T. Lesieur, F. Krzakala, and L. Zdeborová, "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel," in *Proc. Allerton Conf. Commun. Control Comput.*, Sep./Oct. 2015, pp. 680–687.

[67] J. T. Parker and P. Schniter, "Parametric bilinear generalized approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 795–808, Jun. 2016.

[68] W. Hachem, P. Loubaton, and J. Najim, "Deterministic equivalents for certain functionals of large random matrices," *Ann. Appl. Probab.*, vol. 17, no. 3, pp. 875–930, Jun. 2007.

[69] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. North Chelmsford, MA, USA: Courier Corporation, 2000.

Sundeep Rangan (M'02–SM'14–F'16) received the B.A.Sc. degree from the University ofWaterloo, Waterloo, ON, Canada, and the M.Sc. and Ph.D. degrees from the University of California, Berkeley, Berkeley, CA, USA, all in electricalengineering. He has held postdoctoral appointments with the University of Michigan, Ann Arbor, MI, USA, and Bell Labs. In 2000, he cofounded (with four others) Flarion Technologies, a spin-off of Bell Labs, that developed Flash OFDM, the first cellular OFDM data system and precursor to 4G systems including LTE and WiMAX. In 2006, Flarion was acquired by Qualcomm Technologies. He was the Director of Engineering at Qualcomm involved in OFDM infrastructure products. He joined the Department of ECE, NYU, in 2010, where he is currently an Associate Professor and the Director of NYU WIRELESS. His research interests include wireless communications, signal processing, information theory, and control theory.

Philip Schniter (S'92–M'93–SM'05–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 Sup´elec, Gif-sur- Yvette, France. In 2016-2017 he was a Visiting Professor at Duke University, Durham, NC. His areas of interest currently include signal processing, wireless communications, and machine learning.

Erwin Riegler received the Dipl-Ing. degree in Technical Physics (with distinction) in 2001 and the Dr. techn. degree in Technical Physics (with distinction) in 2004 from Vienna University of Technology. From 2005 to 2006 he was a post-doc at the Institute for Analysis and Scientific Computing at Vienna University of Technology. From 2007 to 2010 he was a senior researcher at the Telecommunications Research Center Vienna (FTW). From 2010 to 2014 he was a post-doc at the Institute of Telecommunications at Vienna University of Technology. Since 2014 he has been a senior researcher at the Swiss Federal Institute of Technology in Zurich (ETHZ). He was a visiting researcher at ETHZ, Chalmers University of Technology, The Ohio State University, Aalborg University, and the Max Planck Institute for Mathematics in the Sciences. His research interests include information theory, noncoherent communications, statistical physics, and transceiver design. He is the co-author of a paper that won a student paper award at the International Symposium on Information Theory, 2012.

Alyson K. Fletcher (S'03–M'04) received the B.S. degree in mathematics from the University of Iowa. From the University of California, Berkeley, she received the M.S. degree in electrical engineering in 2002, and the M.A. degree in mathematics and Ph.D. degree in electrical engineering, both in 2006.

Dr. Fletcher is a member of SWE, SIAM, and Sigma Xi. In 2005, she received the University of California Eugene L. Lawler Award, the Henry Luce Foundations Clare Boothe Luce Fellowship, the Soroptimist Dissertation Fellowship, and University of California Presidents Postdoctoral Fellowship. Her research interests include signal processing, information theory, machine learning, and neuroscience.

Volkan Cevher (SM'10) received the B.Sc. (valedictorian) in electrical engineering from Bilkent University in Ankara, Turkey, in 1999 and the Ph.D. in electrical and computer engineering from the Georgia Institute of Technology in Atlanta, GA in 2005. He was a Research Scientist with the University of Maryland, College Park from 2006-2007 and also with Rice University in Houston, TX, from 2008-2009. Currently, he is an Associate Professor at the Swiss Federal Institute of Technology Lausanne and a Faculty Fellow in the Electrical and Computer Engineering Department at Rice University. His research interests include signal processing theory, machine learning, convex optimization, and information theory. Dr. Cevher was the recipient of a Best Paper Award at SPARS in 2009, a Best Paper Award at CAMSAP in 2015, and an ERC StG in 2011.