Approximate Message Passing with Consistent Parameter Estimation and Applications to Sparse Learning

Ulugbek S. Kamilov
EPFL
ulugbek.kamilov@epfl.ch

Sundeep Rangan
Polytechnic Institute of New York University
srangan@poly.edu

Alyson K. Fletcher
University of California, Santa Cruz
afletcher@soe.ucsc.edu

Michael Unser
EPFL
michael.unser@epfl.ch

Abstract

We consider the estimation of an i.i.d. vector $\mathbf{x} \in \mathbb{R}^n$ from measurements $\mathbf{y} \in \mathbb{R}^m$ obtained by a general cascade model consisting of a known linear transform followed by a probabilistic componentwise (possibly nonlinear) measurement channel. We present a method, called adaptive generalized approximate message passing (Adaptive GAMP), that enables joint learning of the statistics of the prior and measurement channel along with estimation of the unknown vector $\mathbf{x}$. Our method can be applied to a large class of learning problems including the learning of sparse priors in compressed sensing or identification of linear-nonlinear cascade models in dynamical systems and neural spiking processes. We prove that for large i.i.d. Gaussian transform matrices the asymptotic componentwise behavior of the adaptive GAMP algorithm is predicted by a simple set of scalar state evolution equations. This analysis shows that the adaptive GAMP method can yield asymptotically consistent parameter estimates, which implies that the algorithm achieves a reconstruction quality equivalent to the oracle algorithm that knows the correct parameter values. The adaptive GAMP methodology thus provides a systematic, general and computationally efficient method applicable to a large range of complex linear-nonlinear models with provable guarantees.

1 Introduction

Consider the estimation of a random vector $\mathbf{x} \in \mathbb{R}^n$ from a measurement vector $\mathbf{y} \in \mathbb{R}^m$. As illustrated in Figure 1, the vector $\mathbf{x}$, which is assumed to have i.i.d. components $x_j \sim P_X$, is passed through a known linear transform that outputs $\mathbf{z} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$. The components of $\mathbf{y} \in \mathbb{R}^m$ are generated by a componentwise transfer function $P_{Y|Z}$. This paper addresses the cases where the distributions P_X and $P_{Y|Z}$ have some parametric uncertainty that must be learned so as to properly estimate $\mathbf{x}$.

This joint estimation and learning problem with linear transforms and componentwise nonlinearities arises in a range of applications, including empirical Bayesian approaches to inverse problems in signal processing, linear regression and classification [1, 2], and, more recently, Bayesian compressed

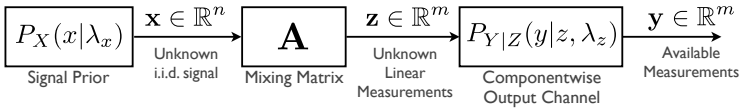

Figure 1: *Measurement model considered in this work. The vector $\mathbf{x} \in \mathbb{R}^n$ with an i.i.d. prior $P_X(x|\lambda_x)$ passes through the linear transform $\mathbf{A} \in \mathbb{R}^{m\times n}$ followed by a componentwise nonlinear channel $P_{Y|Z}(y|z,\lambda_z)$ to result in $\mathbf{y} \in \mathbb{R}^m$. The prior P_X and the nonlinear channel $P_{Y|Z}$ depend on the unknown parameters λ_x and λ_z, respectively. We propose adaptive GAMP to jointly estimate $\mathbf{x}$ and (λ_x, λ_z) given the measurements $\mathbf{y}$.*

sensing for estimation of sparse vectors $\mathbf{x}$ from underdetermined measurements [3–5]. Also, since the parameters in the output transfer function $P_{Y|Z}$ can model unknown nonlinearities, this problem formulation can be applied to the identification of linear-nonlinear cascade models of dynamical systems, in particular for neural spike responses [6–8].

In recent years, there has been considerable interest in so-called *approximate message passing* (AMP) methods for this estimation problem. The AMP techniques use Gaussian and quadratic approximations of loopy belief propagation (LBP) to provide estimation methods that are computationally efficient, general and analytically tractable. However, the AMP methods generally require that the distributions P_X and $P_{Y|Z}$ are known perfectly. When the parameters λ_x and λ_z are unknown, various extensions have been proposed including combining AMP methods with Expectation Maximization (EM) estimation [9–12] and hybrid graphical models approaches [13]. In this work, we present a novel method for joint parameter and vector estimation called *adaptive generalized AMP* (adaptive GAMP), that extends the GAMP method of [14]. We present two major theoretical results related to adaptive GAMP: We first show that, similar to the analysis of the standard GAMP algorithm, the componentwise asymptotic behavior of adaptive GAMP can be exactly described by a simple scalar *state evolution* (SE) equations [14–18]. An important consequence of this result is a theoretical justification to the EM-GAMP algorithm in [9–12] which is a special case of adaptive GAMP with a particular choice of adaptation functions. Our second result demonstrates the asymptotic consistency of adaptive GAMP when adaptation functions correspond to the *maximum-likelihood* (ML) parameter estimation. We show that when the ML estimation is computed exactly, the estimated parameters converge to the true values and the performance of adaptive GAMP asymptotically coincides with the performance of the oracle GAMP algorithms that knows correct parameter values. Adaptive GAMP thus provides a computationally-efficient method for solving a wide variety of joint estimation and learning problems with a simple, exact performance characterization and provable conditions for asymptotic consistency.

All proofs and some technical details that have been omitted for space appear in the full paper [19] that also provides more background and simulations.

2 Adaptive GAMP

Approximate message passing (AMP) refers to a class of algorithms based on Gaussian approximations of loopy belief propagation (LBP) for the estimation of the vectors $\mathbf{x}$ and $\mathbf{z}$ according to the model described in Section 1. These methods originated from CDMA multiuser detection problems in [15, 20, 21]; more recently, they have attracted considerable attention in compressed sensing [17, 18, 22]. The Gaussian approximations used in AMP are closely related to standard expectation propagation techniques [23, 24], but with additional simplifications that exploit the linear coupling between the variables $\mathbf{x}$ and $\mathbf{z}$. The key benefits of AMP methods are their computational performance, their large domain of application, and, for certain large random $\mathbf{A}$, their exact asymptotic performance characterizations with testable conditions for optimality [15–18]. This paper considers an adaptive version of the so-called generalized AMP (GAMP) method of [14] that extends the algorithm in [22] to arbitrary output distributions $P_{Y|Z}$.

The original GAMP algorithm of [14] requires that the distributions P_X and $P_{Y|Z}$ are known. We propose an *adaptive GAMP*, shown in Algorithm 1, to allow for simultaneous estimation of the distributions P_X and $P_{Y|Z}$ along with the estimation of $\mathbf{x}$ and $\mathbf{z}$. The algorithm assumes that distributions P_X and $P_{Y|Z}$ have some parametric forms

$$P_X(x|\lambda_x), \quad P_{Y|Z}(y|z,\lambda_z), \tag{1}$$

for parameters $\lambda_x \in \Lambda_x$ and $\lambda_z \in \Lambda_z$ and for parameter sets Λ_x and Λ_z. Algorithm 1 produces a sequence of estimates $\widehat{\mathbf{x}}^t$ and $\widehat{\mathbf{z}}^t$ for $\mathbf{x}$ and $\mathbf{z}$ along with parameter estimates $\widehat{\lambda}_x^t$ and $\widehat{\lambda}_z^t$, The precise value of these estimates depends on several factors in the algorithm including the termination criteria and the choice of what we will call *estimation functions* G_x^t, G_z^t and G_s^t, and *adaptation functions* H_x^t and H_z^t.

Algorithm 1 Adaptive GAMP

Require: Matrix $\mathbf{A}$, estimation functions G_x^t, G_s^t and G_z^t and adaptation functions H_x^t and H_z^t.

Initialize $t \leftarrow 0$, $\mathbf{s}^{-1} \leftarrow 0$ and some values for $\widehat{\mathbf{x}}^0$, τ_x^0,
repeat
  {Output node update}
  $\tau_p^t \leftarrow \|\mathbf{A}\|_F^2 \tau_x^t / m$
  $\mathbf{p}^t \leftarrow \mathbf{A}\widehat{\mathbf{x}}^t - \mathbf{s}^{t-1}\tau_p^t$
  $\widehat{\lambda}_z^t \leftarrow H_z^t(\mathbf{p}^t, \mathbf{y}, \tau_p^t)$
  $\widehat{z}_i^t \leftarrow G_z^t(p_i^t, y_i, \tau_p^t, \widehat{\lambda}_z^t)$ for all $i = 1, \ldots, m$
  $s_i^t \leftarrow G_s^t(p_i^t, y_i, \tau_p^t, \widehat{\lambda}_z^t)$ for all $i = 1, \ldots, m$
  $\tau_s^t \leftarrow -(1/m)\sum_i \partial G_s^t(p_i^t, y_i, \tau_p^t, \widehat{\lambda}_z^t)/\partial p_i^t$

  {Input node update}
  $1/\tau_r^t \leftarrow \|\mathbf{A}\|_F^2 \tau_s^t / n$
  $\mathbf{r}^t = \mathbf{x}^t + \tau_r^t \mathbf{A}^{\mathrm{T}} \mathbf{s}^t$
  $\widehat{\lambda}_x^t \leftarrow H_x^t(\mathbf{r}^t, \tau_r^t)$
  $\widehat{x}_j^{t+1} \leftarrow G_x^t(r_j^t, \tau_r^t, \widehat{\lambda}_x^t)$ for all $j = 1, \ldots, n$
  $\tau_x^{t+1} \leftarrow (\tau_r^t/n)\sum_j \partial G_x^t(r_j^t, \tau_r^t, \widehat{\lambda}_x^t)/\partial r_j$
until Terminated

The choice of the estimation and adaptation functions allows for considerable flexibility in the algorithm. For example, it is shown in [14] that G_x^t, G_z^t, and G_s^t can be selected such that the GAMP algorithm implements Gaussian approximations of either max-sum LBP or sum-product LBP that approximate the maximum-a-posteriori (MAP) or minimum-mean-squared-error (MMSE) estimates of $\mathbf{x}$ given $\mathbf{y}$, respectively. The adaptation functions can also be selected for a number of different parameter-estimation strategies. Because of space limitation, we present only the estimation functions for the sum-product GAMP algorithm from [14] along with an ML-type adaptation. Some of the analysis below, however, applies more generally.

As described in [14], the sum-product estimation can be implemented with the functions

$$G_x^t(r, \tau_r, \widehat{\lambda}_x) := E[X|R = r, \tau_r, \widehat{\lambda}_x], \tag{2a}$$

$$G_z^t(p, y, \tau_p, \widehat{\lambda}_z) := E[Z|P = p, Y = y, \tau_p, \widehat{\lambda}_z], \tag{2b}$$

$$G_s^t(p, y, \tau_p, \widehat{\lambda}_z) := \frac{1}{\tau_p}\left(G_z^t(p, y, \tau_p, \widehat{\lambda}_z) - p\right), \tag{2c}$$

where the expectations are with respect to the scalar random variables

$$R = X + V_x, \quad V_x \sim \mathcal{N}(0, \tau_r), \quad X \sim P_X(\cdot|\widehat{\lambda}_x), \tag{3a}$$

$$Z = P + V_z, \quad V_z \sim \mathcal{N}(0, \tau_p), \quad Y \sim P_{Y|Z}(\cdot|Z, \widehat{\lambda}_z). \tag{3b}$$

The estimation functions (2) correspond to scalar estimates of random variables in additive white Gaussian noise (AWGN). A key result of [14] is that, when the parameters are set to the true values (i.e. $(\widehat{\lambda}_x, \widehat{\lambda}_z) = (\lambda_x, \lambda_z)$), the outputs $\widehat{\mathbf{x}}^t$ and $\widehat{\mathbf{z}}^t$ can be interpreted as sum products estimates of the conditional expectations $E(\mathbf{x}|\mathbf{y})$ and $E(\mathbf{z}|\mathbf{y})$. The algorithm thus reduces the vector-valued estimation problem to a computationally simple sequence of scalar AWGN estimation problems along with linear transforms.

The estimation functions H_x^t and H_z^t in Algorithm 1 produce the estimates for the parameters λ_x and λ_z. In the special case when H_x^t and H_z^t produce fixed outputs

$$H_z^t(\mathbf{p}^t, \mathbf{y}^t, \tau_p^t) = \overline{\lambda}_z^t, \quad H_x^t(\mathbf{r}^t, \tau_r^t) = \overline{\lambda}_x^t,$$

for *pre-computed values* of $\overline{\lambda}_z^t$ and $\overline{\lambda}_x^t$, the adaptive GAMP algorithm reduces to the standard (non-adaptive) GAMP algorithm of [14]. The non-adaptive GAMP algorithm can be used when the parameters λ_x and λ_z are known.

When the parameters λ_x and λ_z are unknown, it has been proposed in [9–12] that they can be estimated via an EM method that exploits that fact that GAMP provides estimates of the posterior distributions of $\mathbf{x}$ and $\mathbf{z}$ given the current parameter estimates. As described in the full paper [19], this EM-GAMP method corresponds to a special case of the Adaptive GAMP method for a particular choice of the adaptation functions H_x^t and H_z^t.

However, in this work, we consider an alternate parameter estimation method based on ML adaptation. The ML adaptation uses the following fact that we will rigorously justify below: For certain large random $\mathbf{A}$, at any iteration t, the components of the vectors $\mathbf{r}^t$ and the joint vectors $(\mathbf{p}^t, \mathbf{y}^t)$ will be distributed as

$$R = \alpha_r X + V_x, \quad V_x \sim \mathcal{N}(0, \xi_r), \quad X \sim P_X(\cdot|\lambda_x^*), \tag{4a}$$

$$Z = P + V_z, \quad (Z, P) \sim \mathcal{N}(0, K_p), \quad Y \sim P_{Y|Z}(\cdot|Z, \lambda_z^*), \tag{4b}$$

where λ_x^* and λ_z^* are the "true" parameters and the scalars α_r and ξ_r and the covariance matrix $\mathbf{K}_p$ are some parameters that depend on the estimation and adaptation functions used in the previous iterations. Remarkably, the distributions of the components of $\mathbf{r}^t$ and $(\mathbf{p}^t, \mathbf{y}^t)$ will follow (4) even if the estimation functions in the iterations prior to t used the incorrect parameter values. The adaptive GAMP algorithm can thus attempt to estimate the parameters via a maximum likelihood (ML) estimation:

$$H_x^t(\mathbf{r}^t, \tau_r^t) := \arg\max_{\lambda_x \in \Lambda_x} \max_{(\alpha_r, \xi_r) \in S_x(\tau_r^t)} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \phi_x(r_j^t, \lambda_x, \alpha_r, \xi_r) \right\}, \tag{5a}$$

$$H_z^t(\mathbf{p}^t, \mathbf{y}, \tau_p^t) := \arg\max_{\lambda_z \in \Lambda_z} \max_{\mathbf{K}_p \in S_z(\tau_p^t)} \left\{ \frac{1}{m} \sum_{i=0}^{m-1} \phi_z(p_i^t, y_i, \mathbf{K}_p) \right\}, \tag{5b}$$

where S_x and S_z are sets of possible values for the parameters α_r, ξ_r and $\mathbf{K}_p$, ϕ_x and ϕ_z are the log-likelihoods

$$\phi_x(r, \lambda_x, \alpha_r, \xi_r) = \log p_R(r|\lambda_x, \alpha_r, \xi_r), \tag{6a}$$

$$\phi_z(p, y, \lambda_z, \mathbf{K}_p) = \log p_{P,Y}(p, y|\lambda_z, \mathbf{K}_p) \tag{6b}$$

and p_R and $p_{P,Y}$ are the probability density functions corresponding to the distributions in (4).

3 Convergence and Asymptotic Consistency with Gaussian Transforms

3.1 General State Evolution Analysis

Before proving the asymptotic consistency of the adaptive GAMP method with ML adaptation, we first prove a more general convergence result. Among other consequences, the result will justify the distribution model (4) assumed by the ML adaptation. Similar to the SE analyses in [14, 18] we consider the asymptotic behavior of the adaptive GAMP algorithm with large i.i.d. Gaussian matrices. The assumptions are summarized as follows. Details can be found in the full paper [19, Assumption 2].

Assumption 1 *Consider the adaptive GAMP algorithm running on a sequence of problems indexed by the dimension n, satisfying the following:*

(a) For each n, the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has i.i.d. components with $A_{ij} \sim \mathcal{N}(0, 1/m)$ and the dimension $m = m(n)$ is a deterministic function of n satisfying $n/m \to \beta$ for some $\beta > 0$ as $n \to \infty$.

(b) The input vectors $\mathbf{x}$ and initial condition $\widehat{\mathbf{x}}^0$ are deterministic sequences whose components converge empirically with bounded moments of order $s = 2k - 2$ as

$$\lim_{n \to \infty} (\mathbf{x}, \widehat{\mathbf{x}}^0) \overset{\text{PL}(s)}{=} (X, \widehat{X}^0), \tag{7}$$

to some random vector $(X, \widehat{X}^0)$ for $k = 2$. See [19] for a precise statement of this type of convergence.

(c) The output vectors $\mathbf{z}$ *and* $\mathbf{y} \in \mathbb{R}^m$ *are generated by*

$$\mathbf{z} = \mathbf{A}\mathbf{x}, \quad \mathbf{y} = h(\mathbf{z}, \mathbf{w}), \tag{8}$$

for some scalar function $h(z, w)$ where the disturbance vector $\mathbf{w}$ is deterministic, but empirically converges as

$$\lim_{n\to\infty} \mathbf{w} \overset{\text{PL}(s)}{=} W, \tag{9}$$

with $s = 2k-2$, $k = 2$ and W is some random variable. We let $P_{Y|Z}$ denote the conditional distribution of the random variable $Y = h(Z, W)$.

(d) Suitable continuity assumptions on the estimation functions G_x^t, G_z^t and G_s^t and adaptation functions H_x^t and H_z^t – see [19] for details.

Now define the sets of vectors

$$\theta_x^t := \{(x_j, r_j^t, \widehat{x}_j^{t+1}), j = 1, \ldots, n\}, \quad \theta_z^t := \{(z_i, \widehat{z}_i^t, y_i, p_i^t), i = 1, \ldots, m\}. \tag{10}$$

The first vector set, θ_x^t, represents the components of the the "true," but unknown, input vector $\mathbf{x}$, its adaptive GAMP estimate $\widehat{\mathbf{x}}^t$ as well as $\mathbf{r}^t$. The second vector, θ_z^t, contains the components of the "true," but unknown, output vector $\mathbf{z}$, its GAMP estimate $\widehat{\mathbf{z}}^t$, as well as $\mathbf{p}^t$ and the observed input $\mathbf{y}$.

The sets θ_x^t and θ_z^t are implicitly functions of the dimension n. Our main result, Theorem 1 below, characterizes the asymptotic joint distribution of the components of these two sets as $n \to \infty$. Specifically, we will show that the empirical distribution of the components of θ_x^t and θ_z^t converge to a random vectors of the form

$$\overline{\theta}_x^t := (X, R^t, \widehat{X}^{t+1}), \qquad \overline{\theta}_z^t := (Z, \widehat{Z}^t, Y, P^t), \tag{11}$$

where X is the random variable in the initial condition (7). R^t and $\widehat{X}^{t+1}$ are given by

$$R^t = \alpha_r^t X + V^t, \qquad V^t \sim \mathcal{N}(0, \xi_r^t), \qquad \widehat{X}^{t+1} = G_x^t(R^t, \overline{\tau}_r^t, \overline{\lambda}_x^t) \tag{12}$$

for some deterministic constants α_r^t, ξ_r^t, $\overline{\tau}_r^t$ and $\overline{\lambda}_x^t$ that will be defined momentarily. Similarly, $(Z, P^t) \sim \mathcal{N}(0, \mathbf{K}_p^t)$, and

$$Y \sim P_{Y|Z}(\cdot|Z), \qquad \widehat{Z}^t = G_z^t(P^t, Y, \overline{\tau}_p^t, \overline{\lambda}_z^t), \tag{13}$$

where W is the random variable in (9) and $\mathbf{K}_p^t$ and $\overline{\lambda}_z^t$ are also deterministic constants. The deterministic constants above can be computed iteratively with the following *state evolution* (SE) equations shown in Algorithm 2.

Theorem 1 *Consider the random vectors θ_x^t and θ_z^t generated by the outputs of GAMP under Assumption 1. Let $\overline{\theta}_x^t$ and $\overline{\theta}_z^t$ be the random vectors in* (11) *with the parameters determined by the SE equations in Algorithm 2. Then, for any fixed t, almost surely, the components of θ_x^t and θ_z^t converge empirically with bounded moments of order $k = 2$ as*

$$\lim_{n\to\infty} \theta_x^t \overset{\text{PL}(k)}{=} \overline{\theta}_x^t, \qquad \lim_{n\to\infty} \theta_z^t \overset{\text{PL}(k)}{=} \overline{\theta}_z^t. \tag{17}$$

where $\overline{\theta}_x^t$ and $\overline{\theta}_z^t$ are given in (11)*. In addition, for any t, the limits*

$$\lim_n \lambda_x^t = \overline{\lambda}_x^t, \qquad \lim_n \lambda_z^t = \overline{\lambda}_z^t, \qquad \lim_n \tau_r^t = \overline{\tau}_r^t, \qquad \lim_n \tau_p^t = \overline{\tau}_p^t, \tag{18}$$

also hold almost surely.

Similar to several other analyses of AMP algorithms such as [14–18], the theorem provides a *scalar equivalent model* for the componentwise behavior of the adaptive GAMP method. That is, asymptotically the components of the sets θ_x^t and θ_z^t in (10) are distributed identically to simple scalar random variables. The parameters in these random variables can be computed via the SE equations

Algorithm 2 Adaptive GAMP State Evolution

Given the distributions in Assumption 1, compute the sequence of parameters as follows:

- *Initialization*: Set $t = 0$ with

$$\mathbf{K}_x^0 = \text{cov}(X, \widehat{X}^0), \qquad \overline{\tau}_x^0 = \tau_x^0, \tag{14}$$

  where the expectation is over the random variables $(X, \widehat{X}^0)$ in Assumption 1(b) and τ_x^0 is the initial value in the GAMP algorithm.

- *Output node update:* Compute the variables associated with $\overline{\theta}_z^t$:

$$\overline{\tau}_p^t = \beta\overline{\tau}_x^t, \qquad \mathbf{K}_p^t = \beta\mathbf{K}_x^t, \quad \overline{\lambda}_z^t = H_z^t(P^t, \overline{\tau}_p^t), \tag{15a}$$

$$\overline{\tau}_r^t = -\mathbb{E}^{-1}\left[\frac{\partial}{\partial p}G_s^t(P^t, Y, \overline{\tau}_p^t, \overline{\lambda}_z^t)\right], \quad \xi_r^t = (\overline{\tau}_r^t)^2\mathbb{E}\left[G_s^t(P^t, Y, \overline{\tau}_p^t, \overline{\lambda}_z^t)\right], \tag{15b}$$

$$\alpha_r^t = \overline{\tau}_r^t\mathbb{E}\left[\frac{\partial}{\partial z}G_s^t(\widehat{P}, h(z, W), \overline{\tau}_p^t, \overline{\lambda}_z^t)\Big|_{z=Z}\right]. \tag{15c}$$

  where the expectations are over the random variables (P^t, Y, W).

- *Input node update:* Compute the variables associated with $\overline{\theta}_x^t$:

$$\overline{\lambda}_x^t = H_x^t(R^t, \overline{\tau}_r^t), \tag{16a}$$

$$\overline{\tau}_x^{t+1} = \overline{\tau}_r^t\mathbb{E}\left[\frac{\partial}{\partial r}G_x^t(R^t, \overline{\tau}_r^t, \overline{\lambda}_x^t)\right], \quad \mathbf{K}_x^{t+1} = \text{cov}(X, \widehat{X}^{t+1}), \tag{16b}$$

  where the expectation is over the random variable $(X, \widehat{X}^{t+1})$.

(14), (15) and (16), which can be evaluated with one or two-dimensional integrals. From this scalar equivalent model, one can compute a large class of componentwise performance metrics such as mean-squared error (MSE) or detection error rates. Thus, the SE analysis shows that for, essentially arbitrary estimation and adaptation functions, and distributions on the true input and disturbance, we can exactly evaluate the asymptotic behavior of the adaptive GAMP algorithm. In addition, when the parameter values $\overline{\lambda}_x$ and $\overline{\lambda}_z$ are fixed, the SE equations in Algorithm 2 reduce to SE equations for the standard (non-adaptive) GAMP algorithm described in [14].

3.2 Asymptotic Consistency with ML Adaptation

The general result, Theorem 1, can be applied to the adaptive GAMP algorithm with arbitrary estimation and adaptation function. In particular, the result can be used to rigorously justify the SE analysis of the EM-GAMP presented in [11, 12]. Here, we use the result to prove the asymptotic parameter consistency of Adaptive GAMP with ML adaptation. The key point is to realize that the distributions (12) and (13) exactly match the distributions (4) assumed by the ML adaptation functions (5). Thus, the ML adaptation should work provided that the maximizations in (5) yield the correct parameter estimates. This condition is essentially an *identifiability* requirement that we make precise with the following definitions.

Definition 1 *Consider a family of distributions, $\{P_X(x|\lambda_x), \lambda_x \in \Lambda_x\}$, a set S_x of parameters (α_r, ξ_r) of a Gaussian channel and function $\phi_x(r, \lambda_x, \alpha_r, \xi_r)$. We say that $P_X(x|\lambda_x)$ is* identifiable with Gaussian outputs *with parameter set S_x and function ϕ_x if:*

(a) The sets S_x and Λ_x are compact.

(b) For any "true" parameters $\lambda_x^* \in \Lambda_x$, and $(\alpha_r, \xi_r) \in S_x$, the maximization

$$\widehat{\lambda}_x = \arg\max_{\lambda_x \in \Lambda_x} \max_{(\alpha_r, \xi_r) \in S_x} \mathbb{E}\left[\phi_x(\alpha_r^* X + V, \lambda_x, \alpha_r, \xi_r)|\lambda_x^*, \alpha_r^*, \xi_r^*\right], \tag{19}$$

is well-defined, unique and returns the true value, $\widehat{\lambda}_x = \lambda_x^*$. The expectation in (19) *is with respect to $X \sim P_X(\cdot|\lambda_x^*)$ and $V \sim \mathcal{N}(0, \xi_r^*)$.*

(c) Suitable continuity assumptions – see [19] for details.

Definition 2 *Consider a family of conditional distributions, $\{P_{Y|Z}(y|z,\lambda_z), \lambda_z \in \Lambda_z\}$ generated by the mapping $Y = h(Z,W,\lambda_z)$ where $W \sim P_W$ is some random variable and $h(z,w,\lambda_z)$ is a scalar function. Let S_z be a set of covariance matrices $\mathbf{K}_p$ and let $\phi_z(y,p,\lambda_z,\mathbf{K}_p)$ be some function. We say that conditional distribution family $P_{Y|Z}(\cdot|\cdot,\lambda_z)$ is* identifiable with Gaussian inputs *with covariance set S_z and function ϕ_z if:*

(a) The parameter sets S_z and Λ_z are compact.

(b) For any "true" parameter $\lambda_z^* \in \Lambda_z$ and true covariance $\mathbf{K}_p^*$, the maximization

$$\widehat{\lambda}_z = \arg\max_{\lambda_z \in \Lambda_z} \max_{\mathbf{K}_p \in S_z} \mathbb{E}\left[\phi_z(Y,P,\lambda_z,\mathbf{K}_p)|\lambda_z^*,\mathbf{K}_p^*\right], \tag{20}$$

is well-defined, unique and returns the true value, $\widehat{\lambda}_z = \lambda_z^*$, The expectation in (20) *is with respect to $Y|Z \sim P_{Y|Z}(y|z,\lambda_z^*)$ and $(Z,P) \sim \mathcal{N}(0,\mathbf{K}_p^*)$.*

(c) Suitable continuity assumptions – see [19] for details.

Definitions 1 and 2 essentially require that the parameters λ_x and λ_z can be identified through a maximization. The functions ϕ_x and ϕ_z can be the log likelihood functions (6a) and (6b), although we permit other functions as well. See [19] for further discussion of the likelihood functions as well as the choice of the parameter sets S_x and S_z.

Theorem 2 *Let $P_X(\cdot|\lambda_x)$ and $P_{Y|Z}(\cdot|\cdot,\lambda_z)$ be families of input and output distributions that are identifiable in the sense of Definitions 1 and 2. Consider the outputs of the adaptive GAMP algorithm using the ML adaptation functions* (5) *using the functions ϕ_x and ϕ_z and parameter sets in Definitions 1 and 2. In addition, suppose Assumption 1(a) to (c) hold where the distribution of X is given by $P_X(\cdot|\lambda_x^*)$ for some "true" parameter $\lambda_x^* \in \Lambda_x$ and the conditional distribution of Y given Z is given by $P_{Y|Z}(y|z,\lambda_z^*)$ for some "true" parameter $\lambda_z^* \in \Lambda_z$. Then, under suitable continuity conditions (see [19] for details), for any fixed t,*

(a) The components of θ_x^t and θ_z^t in (10) *converge empirically with bounded moments of order $k = 2$ as in* (17) *and the limits* (18) *hold almost surely.*

(b) If $(\alpha_r^t, \xi_r^t) \in S_x(\tau_r^t)$ for some t, then $\lim_{n\to\infty} \widehat{\lambda}_x^t = \overline{\lambda}_x^t = \lambda_x^*$ almost surely.

(c) If $\mathbf{K}_p^t \in S_z(\tau_p^t)$ for some t, then $\lim_{n\to\infty} \widehat{\lambda}_z^t = \overline{\lambda}_z^t = \lambda_z^*$ almost surely.

The theorem shows, remarkably, that for a very large class of the parameterized distributions, the adaptive GAMP algorithm with ML adaptation is able to asymptotically estimate the correct parameters. Also, once the consistency limits in (b) and (c) hold, the SE equations in Algorithm 2 reduce to the SE equations for the non-adaptive GAMP method running with the true parameters. Thus, we conclude there is asymptotically no performance loss between the adaptive GAMP algorithm and a corresponding oracle GAMP algorithm that knows the correct parameters in the sense that the empirical distributions of the algorithm outputs are described by the same SE equations.

4 Numerical Example: Estimation of a Gauss-Bernoulli input

Recent results suggest that there is considerable value in learning of priors P_X in the context of compressed sensing [25], which considers the estimation of sparse vectors $\mathbf{x}$ from underdetermined measurements ($m < n$) . It is known that estimators such as LASSO offer certain optimal min-max performance over a large class of sparse distributions [26]. However, for many particular distributions, there is a potentially large performance gap between LASSO and MMSE estimator with the correct prior. This gap was the main motivation for [9, 10] which showed large gains of the EM-GAMP method due to its ability to learn the prior. Here, we present a simple simulation to illustrate the performance gain of adaptive GAMP and its asymptotic consistency. Specifically, Fig. 2 compares the performance of adaptive GAMP for estimation of a sparse Gauss-Bernoulli signal $\mathbf{x} \in \mathbb{R}^n$ from m noisy measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w},$$

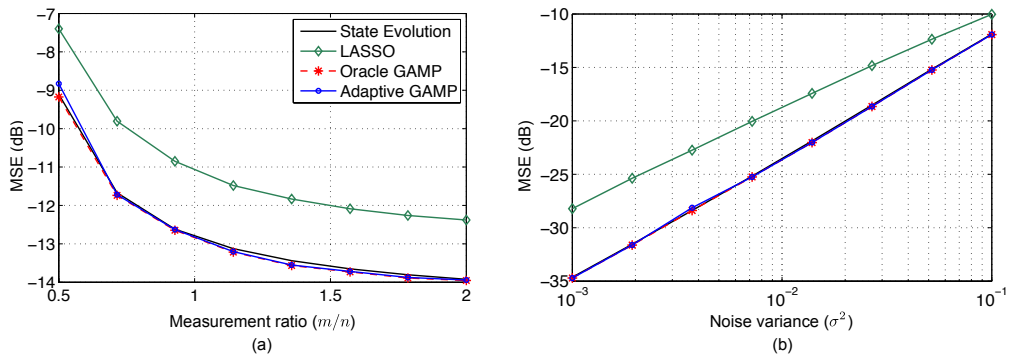

Figure 2: *Reconstruction of a Gauss-Bernoulli signal from noisy measurements. The average reconstruction MSE is plotted against (a) measurement ratio m/n and (b) AWGN variance σ^2. The plots illustrate that adaptive GAMP yields considerable improvement over ℓ_1-based LASSO estimator. Moreover, it exactly matches the performance of oracle GAMP that knows the prior parameters.*

where the additive noise $\mathbf{w}$ is random with i.i.d. entries $w_i \sim \mathcal{N}(0, \sigma^2)$. The signal of length $n = 400$ has 20% nonzero components drawn from the Gaussian distribution of variance 5. Adaptive GAMP uses EM iterations, which are used to approximate ML parameter estimation, to jointly recover the unknown signal $\mathbf{x}$ and the true parameters $\lambda_x = (\rho = 0.2, \sigma_x^2 = 5)$. The performance of adaptive GAMP is compared to that of LASSO with MSE optimal regularization parameter, and oracle GAMP that knows the parameters of the prior exactly. For generating the graphs, we performed 1000 random trials by forming the measurement matrix $\mathbf{A}$ from i.i.d. zero-mean Gaussian random variables of variance $1/m$. In Figure 2(a), we keep the variance of the noise fixed to $\sigma^2 = 0.1$ and plot the average MSE of the reconstruction against the measurement ratio m/n. In Figure 2(b), we keep the measurement ratio fixed to $m/n = 0.75$ and plot the average MSE of the reconstruction against the noise variance σ^2. For completeness, we also provide the asymptotic MSE values computed via SE recursion. The results illustrate that GAMP significantly outperforms LASSO over the whole range of m/n and σ^2. Moreover, the results corroborate the consistency of adaptive GAMP which achieves nearly identical quality of reconstruction with oracle GAMP. The performance results here and in [19] indicate that adaptive GAMP can be an effective method for estimation when the parameters of the problem are difficult to characterize and must be estimated from data.

5 Conclusions and Future Work

We have presented an adaptive GAMP method for the estimation of i.i.d. vectors $\mathbf{x}$ observed through a known linear transforms followed by an arbitrary, componentwise random transform. The procedure, which is a generalization of EM-GAMP methodology of [9, 10], estimates both the vector $\mathbf{x}$ as well as parameters in the source and componentwise output transform. In the case of large i.i.d. Gaussian transforms with ML parameter estimation, it is shown that the adaptive GAMP method is provably asymptotically consistent in that the parameter estimates converge to the true values. This convergence result holds over a large class of models with essentially arbitrarily complex parameterizations. Moreover, the algorithm is computationally efficient since it reduces the vector-valued estimation problem to a sequence of scalar estimation problems in Gaussian noise. We believe that this method is applicable to a large class of linear-nonlinear models with provable guarantees and that it can have applications in a wide range of problems. We have mentioned the use of the method for learning sparse priors in compressed sensing. Future work will include possible extensions to non-Gaussian matrices.

References

[1] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research*, vol. 1, pp. 211–244, Sep. 2001.

[2] M. West, "Bayesian factor regressionm models in the "large p, small n" paradigm," *Bayesian Statistics*, vol. 7, 2003.

[3] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[4] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, pp. 2346–2356, Jun. 2008.

[5] V. Cevher, "Learning with compressible priors," in *Proc. NIPS*, Vancouver, BC, Dec. 2009.

[6] S. Billings and S. Fakhouri, "Identification of systems containing linear dynamic and static nonlinear elements," *Automatica*, vol. 18, no. 1, pp. 15–26, 1982.

[7] I. W. Hunter and M. J. Korenberg, "The identification of nonlinear biological systems: Wiener and Hammerstein cascade models," *Biological Cybernetics*, vol. 55, no. 2–3, pp. 135–144, 1986.

[8] O. Schwartz, J. W. Pillow, N. C. Rust, and E. P. Simoncelli, "Spike-triggered neural characterization," *J. Vision*, vol. 6, no. 4, pp. 484–507, Jul. 2006.

[9] J. P. Vila and P. Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in *Conf. Rec. 45th Asilomar Conf. Signals, Syst. & Comput.*, Pacific Grove, CA, Nov. 2011, pp. 799–803.

[10] ——, "Expectation-maximization Gaussian-mixture approximate message passing," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2012.

[11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Statistical physics-based reconstruction in compressed sensing," arXiv:1109.4424, Sep. 2011.

[12] ——, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," arXiv:1206.3953, Jun. 2012.

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

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

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

[16] ——, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," in *Proc. IEEE Int. Symp. Inform. Theory*, Nice, France, Jun. 2007, pp. 946–950.

[17] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2010, pp. 1–6.

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

[19] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," arXiv:1207.3859 [cs.IT], Jul. 2012.

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

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

[22] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[23] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[24] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Machine Learning Research*, vol. 9, pp. 759–813, Sep. 2008.

[25] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[26] D. Donoho, I. Johnstone, A. Maleki, and A. Montanari, "Compressed sensing over ℓ^p-balls: Minimax mean square error," in *Proc. ISIT*, St. Petersburg, Russia, Jun. 2011.