Expectation-Maximization Bernoulli-Gaussian Approximate Message Passing

Jeremy Vila and Philip Schniter
Dept. of ECE, The Ohio State University, Columbus, OH 43210. (Email: vilaj@ece.osu.edu, schniter@ece.osu.edu)

***Abstract*—The approximate message passing (AMP) algorithm originally proposed by Donoho, Maleki, and Montanari yields a computationally attractive solution to the usual ℓ_1-regularized least-squares problem faced in compressed sensing, whose solution is known to be robust to the signal distribution. When the signal is drawn i.i.d from a marginal distribution that is not least-favorable, better performance can be attained using a Bayesian variation of AMP. The latter, however, assumes that the distribution is perfectly known. In this paper, we navigate the space between these two extremes by modeling the signal as i.i.d Bernoulli-Gaussian (BG) with unknown prior sparsity, mean, and variance, and the noise as zero-mean Gaussian with unknown variance, and we simultaneously reconstruct the signal while learning the prior signal and noise parameters. To accomplish this task, we embed the BG-AMP algorithm within an expectation-maximization (EM) framework. Numerical experiments confirm the excellent performance of our proposed EM-BG-AMP on a range of signal types.**[1][2]

I. Introduction

We consider the problem of recovering a signal $\boldsymbol{x} \in \mathbb{R}^N$ from noisy linear measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{R}^M$ in the "undersampled" regime where $M < N$. Roughly speaking, when $\boldsymbol{x}$ is sufficiently sparse (or compressible) and when the matrix $\boldsymbol{A}$ is sufficiently well-conditioned, accurate signal recovery is possible with polynomial-complexity algorithms.

One of the best-known approaches to this problem is known as "Lasso" [1], which minimizes the convex criterion

$$\hat{\boldsymbol{x}}_{\mathsf{lasso}} = \arg\min_{\hat{\boldsymbol{x}}} \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2^2 + \lambda_{\mathsf{lasso}}\|\hat{\boldsymbol{x}}\|_1, \qquad (1)$$

with λ_{lasso} a tuning parameter. When $\boldsymbol{A}$ is constructed from i.i.d Gaussian entries, the so-called phase transition curve (PTC) gives a sharp characterization of Lasso performance for K-sparse $\boldsymbol{x}$ in the large system limit, i.e., as $K, M, N \to \infty$ with fixed undersampling ratio M/N and sparsity ratio K/M [2]. (The Lasso PTC is illustrated in Figs. 1-3.) For noiseless observations, the PTC partitions the M/N-versus-K/M plane into two regions: one where Lasso reconstructs the signal perfectly (with high probability), and one where it does not. For noisy observations, the same curve indicates whether the noise sensitivity (i.e., the ratio of estimation-error power to measurement-noise power under the worst-case signal distribution) of Lasso remains bounded [3].

One remarkable feature of the noiseless Lasso PTC is that it is invariant to signal distribution. In other words, if we adopt a probabilistic viewpoint where the elements of $\boldsymbol{x}$ are drawn i.i.d from the marginal pdf $p_X(x) = \lambda f(x) + (1-\lambda)\delta(x)$, for Dirac delta $\delta(x)$, active-coefficient pdf $f(x)$, and $\lambda \triangleq K/N$, then the Lasso PTC is not affected by $f(\cdot)$. This PTC invariance implies that Lasso is robust, but that it cannot benefit from the restriction of $\boldsymbol{x}$ to an "easier" signal class. For example, if the coefficients in $\boldsymbol{x}$ are known to be non-negative, then there exists a polynomial-complexity algorithm whose PTC is better than that of Lasso [2].

Although, in some applications, robustness to worst-case signals may be the dominant concern, in many other applications the goal is to maximize average-case performance over an anticipated signal class. When the signal $\boldsymbol{x}$ is drawn i.i.d from an arbitrary *known* marginal distribution $p_X(\cdot)$ and the noise $\boldsymbol{w}$ is i.i.d Gaussian with *known* variance, there exist very-low-complexity iterative Bayesian algorithms to generate approximately maximum a posteriori (MAP) and minimum mean-squared error (MMSE) signal estimates, notably the Bayesian version of Donoho, Maleki, and Montanari's *approximate message passing* (AMP) algorithm [4]. AMP is formulated from a loopy-belief-propagation perspective, leveraging central-limit-theorem approximations that hold in the large system limit for suitably dense $\boldsymbol{A}$, and admits a rigorous analysis in the large-system limit [5]. Meanwhile, AMP's complexity is remarkably low, dominated by one application of $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathsf{T}}$ per iteration (which is especially cheap if $\boldsymbol{A}$ is an FFT or other fast operation), with typically < 50 iterations to convergence. More recently, a *generalized AMP* (GAMP) algorithm [6] was proposed that relaxes the requirements on the noise distribution and on the sensing matrix $\boldsymbol{A}$. (See Table I for a summary.)

Given that it is rare to know the signal and noise distributions perfectly, we take the approach of assuming signal and noise distributions that are known up to some statistical parameters, and then learning those unknown parameters while simultaneously recovering the signal. Examples of this "empirical Bayesian" approach include several algorithms based on Tipping's *relevance vector machine* [7]–[9]. Although the average-case performance of those algorithms is often quite good (depending on the signal class, of course), their complexities are generally much larger than that of (G)AMP.

In this paper, we propose a GAMP-based empirical-Bayesian algorithm. In particular, we treat the signal as Bernoulli-Gaussian (BG) signal with unknown sparsity, mean, and variance, and the noise as Gaussian with unknown variance, and then we then learn these statistical parameters using

[1]This work has been supported in part by NSF-I/UCRC grant IIP-0968910, by NSF grant CCF-1018368, and by DARPA/ONR grant N66001-10-1-4090.

[2]Portions of this work were presented in a poster at the Duke Workshop on Sensing and Analysis of High-Dimensional Data, July 2011.

an expectation-maximization (EM) approach [10] that calls BG-GAMP once per EM-iteration.

II. Bernoulli-Gaussian GAMP

A core component of our proposed method is the Bernoulli-Gaussian (BG) GAMP algorithm, which we now review. For BG-GAMP, the signal $\boldsymbol{x} = [x_1, \ldots, x_N]^\mathsf{T}$ is assumed to be i.i.d BG, i.e., to have marginal pdf

$$p_X(x; \lambda, \theta, \phi) = (1-\lambda)\delta(x) + \lambda \mathcal{N}(x; \theta, \phi), \tag{2}$$

where $\delta(\cdot)$ denotes the Dirac delta, λ the sparsity rate, θ the active-coefficient mean, and ϕ the active-coefficient variance. The noise $\boldsymbol{w}$ is assumed to be independent of $\boldsymbol{x}$ and i.i.d zero-mean Gaussian with variance ψ:

$$p_W(w; \psi) = \mathcal{N}(w; 0, \psi) \tag{3}$$

In our approach, the parameters $\boldsymbol{q} \triangleq [\lambda, \theta, \phi, \psi]$ that define these prior distributions are treated as deterministic unknowns, and learned through the EM algorithm, as detailed in Section III. Although above and in the sequel we assume real-valued Gaussians, all expressions can be converted to the circular-complex-Gaussian case by replacing all $\mathcal{N}$ with $\mathcal{CN}$ and removing all $\frac{1}{2}$'s.

GAMP can handle an arbitrary probabilistic relationship $p_{Y|Z}(y_m|z_m)$ between the observed output y_m and the noiseless output $z_m \triangleq \boldsymbol{a}_m^\mathsf{T}\boldsymbol{x}$, where $\boldsymbol{a}_m^\mathsf{T}$ is the m^{th} row of $\boldsymbol{A}$. Our additive Gaussian noise assumption implies $p_{Y|Z}(y|z) = \mathcal{N}(y; z, \psi)$. To complete our description, we need only to specify $g_{\mathsf{in}}(\cdot)$, $g'_{\mathsf{in}}(\cdot)$, $g_{\mathsf{out}}(\cdot)$, and $g'_{\mathsf{out}}(\cdot)$ in Table I. Using straightforward manipulations, our $p_{Y|Z}(\cdot|\cdot)$ yields [6]

$$g_{\mathsf{out}}(y, \hat{z}, \mu^z; \boldsymbol{q}) = \frac{y - \hat{z}}{\mu^z + \psi} \tag{4}$$

$$-g'_{\mathsf{out}}(y, \hat{z}, \mu^z; \boldsymbol{q}) = \frac{1}{\mu^z + \psi}, \tag{5}$$

and our BG signal prior (2) yields

$$g_{\mathsf{in}}(\hat{r}, \mu^r; \boldsymbol{q}) = \pi(\hat{r}, \mu^r; \boldsymbol{q})\, \gamma(\hat{r}, \mu^r; \boldsymbol{q}) \tag{6}$$

$$\begin{aligned}\mu^r g'_{\mathsf{in}}(\hat{r}, \mu^r; \boldsymbol{q}) &= \pi(\hat{r}, \mu^r; \boldsymbol{q}) \left(\nu(\hat{r}, \mu^r; \boldsymbol{q}) + |\gamma(\hat{r}, \mu^r; \boldsymbol{q})|^2\right) \\ &\quad - \left(\pi(\hat{r}, \mu^r; \boldsymbol{q})\right)^2 |\gamma(\hat{r}, \mu^r; \boldsymbol{q})|^2,\end{aligned} \tag{7}$$

where

$$\pi(\hat{r}, \mu^r; \boldsymbol{q}) \triangleq \frac{1}{1 + \left(\frac{\lambda}{1-\lambda}\frac{\mathcal{N}(\hat{r}; \theta, \phi+\mu^r)}{\mathcal{N}(\hat{r}; 0, \mu^r)}\right)^{-1}} \tag{8}$$

$$\gamma(\hat{r}, \mu^r; \boldsymbol{q}) \triangleq \frac{\hat{r}/\mu^r + \theta/\phi}{1/\mu^r + 1/\phi} \tag{9}$$

$$\nu(\hat{r}, \mu^r; \boldsymbol{q}) \triangleq \frac{1}{1/\mu^r + 1/\phi}. \tag{10}$$

Table I implies that BG-GAMP's marginal posteriors are

$$p(x_n|\boldsymbol{y}; \boldsymbol{q}) = \tfrac{1}{C_n}\, p_X(x_n; \boldsymbol{q})\, \mathcal{N}(x_n; \hat{r}_n, \mu^r_n) \tag{11}$$

$$\begin{aligned}&= \tfrac{1}{C_n}\big((1-\lambda)\delta(x_n) + \lambda\mathcal{N}(x_n; \theta, \phi)\big) \\ &\quad \times \mathcal{N}(x_n; \hat{r}_n, \mu^r_n)\end{aligned} \tag{12}$$

definitions:

$$p_{Z|Y}(z|y; \hat{z}, \mu^z) = \frac{p_{Y|Z}(y|z)\,\mathcal{N}(z; \hat{z}, \mu^z)}{\int_{z'} p_{Y|Z}(y|z')\,\mathcal{N}(z'; \hat{z}, \mu^z)} \quad \text{(D1)}$$

$$g_{\mathsf{out}}(y, \hat{z}, \mu^z) = \tfrac{1}{\mu^z}\left(\mathrm{E}_{Z|Y}\{z|y; \hat{z}, \mu^z\} - \hat{z}\right) \quad \text{(D2)}$$

$$g'_{\mathsf{out}}(y, \hat{z}, \mu^z) = \tfrac{1}{\mu^z}\left(\frac{\mathrm{var}_{Z|Y}\{z|y; \hat{z}, \mu^z\}}{\mu^z} - 1\right) \quad \text{(D3)}$$

$$p_{X|\boldsymbol{Y}}(x|\boldsymbol{y}; \hat{r}, \mu^r) = \frac{p_X(x)\,\mathcal{N}(x; \hat{r}, \mu^r)}{\int_{x'} p_X(x')\,\mathcal{N}(x'; \hat{r}, \mu^r)} \quad \text{(D4)}$$

$$g_{\mathsf{in}}(\hat{r}, \mu^r) = \textstyle\int_x x\, p_{X|\boldsymbol{Y}}(x|\boldsymbol{y}; \hat{r}, \mu^r) \quad \text{(D5)}$$

$$g'_{\mathsf{in}}(\hat{r}, \mu^r) = \tfrac{1}{\mu^r}\textstyle\int_x |x - g_{\mathsf{in}}(\hat{r}, \mu^r)|^2\, p_{X|\boldsymbol{Y}}(x|\boldsymbol{y}; \hat{r}, \mu^r) \quad \text{(D6)}$$

initialize:

$$\forall n: \hat{x}_n(1) = \textstyle\int_x x\, p_X(x) \quad \text{(I1)}$$

$$\forall n: \mu^x_n(1) = \textstyle\int_x |x - \hat{x}_n(1)|^2 p_X(x) \quad \text{(I2)}$$

$$\forall m: \hat{u}_m(0) = 0 \quad \text{(I3)}$$

for $t = 1, 2, 3, \ldots$

$$\forall m: \hat{z}_m(t) = \textstyle\sum_{n=1}^N A_{mn}\hat{x}_n(t) \quad \text{(R1)}$$

$$\forall m: \mu^z_m(t) = \textstyle\sum_{n=1}^N |A_{mn}|^2 \mu^x_n(t) \quad \text{(R2)}$$

$$\forall m: \hat{p}_m(t) = \hat{z}_m(t) - \mu^z_m(t)\,\hat{u}_m(t-1) \quad \text{(R3)}$$

$$\forall m: \hat{u}_m(t) = g_{\mathsf{out}}(y_m, \hat{p}_m(t), \mu^z_m(t)) \quad \text{(R4)}$$

$$\forall m: \mu^u_m(t) = -g'_{\mathsf{out}}(y_m, \hat{p}_m(t), \mu^z_m(t)) \quad \text{(R5)}$$

$$\forall n: \mu^r_n(t) = \left(\textstyle\sum_{n=1}^N |A_{mn}|^2 \mu^u_m(t)\right)^{-1} \quad \text{(R6)}$$

$$\forall n: \hat{r}_n(t) = \hat{x}_n(t) + \mu^r_n(t) \textstyle\sum_{m=1}^M A^*_{mn}\hat{u}_m(t) \quad \text{(R7)}$$

$$\forall n: \mu^x_n(t+1) = \mu^r_n(t) g'_{\mathsf{in}}(\hat{r}_n(t), \mu^r_n(t)) \quad \text{(R8)}$$

$$\forall n: \hat{x}_n(t+1) = g_{\mathsf{in}}(\hat{r}_n(t), \mu^r_n(t)) \quad \text{(R9)}$$

end

TABLE I
The GAMP Algorithm [6]

for scaling factor $C_n \triangleq \int p_X(x_n; \boldsymbol{q})\,\mathcal{N}(x_n; \hat{r}_n, \mu^r_n)$. From (12), it is straightforward to show that BG-GAMP yields the following posterior support probabilities:

$$\Pr\{x_n \neq 0 \,|\, \boldsymbol{y}; \boldsymbol{q}\} = \pi(\hat{r}_n, \mu^r_n; \boldsymbol{q}). \tag{13}$$

III. EM Learning of the Prior Parameters $\boldsymbol{q}$

We use the expectation-maximization (EM) algorithm [10] to learn the statistical parameters $\boldsymbol{q} \triangleq [\lambda, \theta, \phi, \psi]$. The EM algorithm is an iterative technique that increases the likelihood at each iteration, guaranteeing convergence to a local maximum of the likelihood $p(\boldsymbol{y}; \boldsymbol{q})$. In our case, we choose the "hidden data" to be $\{\boldsymbol{x}, \boldsymbol{w}\}$, which yields the EM update

$$\boldsymbol{q}^{i+1} = \arg\max_{\boldsymbol{q}} \mathrm{E}\left\{\ln p(\boldsymbol{x}, \boldsymbol{w}; \boldsymbol{q}) \,\middle|\, \boldsymbol{y}; \boldsymbol{q}^i\right\}, \tag{14}$$

where i denotes EM iteration and $\mathrm{E}\{\cdot|\boldsymbol{y}; \boldsymbol{q}^i\}$ denotes expectation conditioned on the observations $\boldsymbol{y}$ under the parameter hypothesis $\boldsymbol{q}^i$. Moreover, we use the well-established "incremental" updating schedule [11], where $\boldsymbol{q}$ is updated one element at a time while keeping the other elements fixed.

A. EM update for λ

We now derive the EM update for λ given previous parameters $\boldsymbol{q}^i = [\lambda^i, \theta^i, \phi^i, \psi^i]$. Because $\boldsymbol{x}$ is apriori independent of $\boldsymbol{w}$ and i.i.d, the joint pdf $p(\boldsymbol{x}, \boldsymbol{w}; \boldsymbol{q})$ decouples into $C\prod_{n=1}^N p_X(x_n; \lambda, \theta, \phi)$ for a λ-invariant constant C, and so

$$\lambda^{i+1} = \arg\max_{\lambda \in (0,1)} \sum_{n=1}^N \mathrm{E}\left\{\ln p_X(x_n; \lambda, \theta^i, \phi^i) \,\middle|\, \boldsymbol{y}; \boldsymbol{q}^i\right\}. \tag{15}$$

The maximizing value of λ in (15) is necessarily a value of λ that zeroes the derivative, i.e., that satisfies

$$\sum_{n=1}^N \int_{x_n} p(x_n|\boldsymbol{y}; \boldsymbol{q}^i) \frac{d}{d\lambda} \ln p_X(x_n; \lambda, \theta^i, \phi^i) = 0. \tag{16}$$

For the $p_X(x_n;\lambda,\theta,\phi)$ given in (2), it is readily seen that

$$\frac{d}{d\lambda}\ln p_X(x_n;\lambda,\theta^i,\phi^i) = \frac{\mathcal{N}(x_n;\theta^i,\phi^i)-\delta(x_n)}{p_X(x_n;\lambda,\theta^i,\phi^i)} = \begin{cases}\frac{1}{\lambda} & x_n\neq 0\\ \frac{-1}{1-\lambda} & x_n=0\end{cases}. \quad (17)$$

Plugging (17) and (12) into (16), it becomes evident that the neighborhood around the point $x_n=0$ should be treated differently than the remainder of $\mathbb{R}$. Thus, we define the closed ball $\mathcal{B}_\epsilon=[-\epsilon,\epsilon]$ and $\overline{\mathcal{B}_\epsilon}\triangleq\mathbb{R}\setminus\mathcal{B}_\epsilon$, and note that, in the limit $\epsilon\to 0$, the following is equivalent to (16):

$$\frac{1}{\lambda}\sum_{n=1}^N \underbrace{\int_{x_n\in\overline{\mathcal{B}_\epsilon}} p(x_n\,|\,\boldsymbol{y};\boldsymbol{q}^i)}_{\overset{\epsilon\to 0}{=}\pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)} = \frac{1}{1-\lambda}\sum_{n=1}^N \underbrace{\int_{x_n\in\mathcal{B}_\epsilon} p(x_n\,|\,\boldsymbol{y};\boldsymbol{q}^i)}_{\overset{\epsilon\to 0}{=}1-\pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)}. \quad (18)$$

To verify that the left integral converges to the $\pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)$ defined in (8), it suffices to plug (12) into (18) and apply the Gaussian-pdf multiplication rule;[3] meanwhile, for any ϵ, the right integral must equal one minus the left. Finally, the EM update for λ is the unique value satisfying (18) as $\epsilon\to 0$, i.e.,

$$\lambda^{i+1} = \frac{1}{N}\sum_{n=1}^N \pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i). \quad (19)$$

Conveniently, $\{\pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\}_{n=1}^N$ are GAMP outputs.

B. EM update for θ

Similar to (15), the EM update[4] for θ can be written as

$$\theta^{i+1} = \arg\max_{\theta\in\mathbb{R}} \sum_{n=1}^N \mathrm{E}\left\{\ln p_X(x_n;\lambda^i,\theta,\phi^i)\,\middle|\,\boldsymbol{y};\boldsymbol{q}^i\right\}, \quad (20)$$

The maximizing value of θ in (20) is necessarily a value of θ that zeroes the derivative, i.e., that satisfies

$$\sum_{n=1}^N \int_{x_n} p(x_n|\boldsymbol{y};\boldsymbol{q}^i)\frac{d}{d\theta}\ln p_X(x_n;\lambda^i,\theta,\phi^i) = 0. \quad (21)$$

For the $p_X(x_n;\lambda,\theta,\phi)$ given in (2), it is readily seen that

$$\frac{d}{d\theta}\ln p_X(x_n;\lambda^i,\theta,\phi^i) = \frac{(x_n-\theta)}{\phi^i}\frac{\lambda^i\mathcal{N}(x_n;\theta,\phi^i)}{p_X(x_n;\lambda^i,\theta,\phi^i)} = \begin{cases}\frac{x_n-\theta}{\phi^i} & x_n\neq 0\\ 0 & x_n=0\end{cases}. \quad (22)$$

Splitting the domain of integration in (21) into $\mathcal{B}_\epsilon$ and $\overline{\mathcal{B}_\epsilon}$, and then plugging in (22), we find that the following is equivalent to (21) in the limit of $\epsilon\to 0$:

$$\sum_{n=1}^N \int_{x_n\in\overline{\mathcal{B}_\epsilon}} (x_n-\theta)\,p(x_n|\boldsymbol{y};\boldsymbol{q}^i) = 0. \quad (23)$$

The unique value of θ satisfying (23) as $\epsilon\to 0$ is then

$$\theta^{i+1} = \frac{\sum_{n=1}^N \lim_{\epsilon\to 0}\int_{x_n\in\overline{\mathcal{B}_\epsilon}} x_n p(x_n|\boldsymbol{y};\boldsymbol{q}^i)}{\sum_{n=1}^N \lim_{\epsilon\to 0}\int_{x_n\in\overline{\mathcal{B}_\epsilon}} p(x_n|\boldsymbol{y};\boldsymbol{q}^i)} \quad (24)$$

$$= \frac{1}{\lambda^{i+1}N}\sum_{n=1}^N \pi(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\gamma(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i) \quad (25)$$

for the GAMP outputs $\{\gamma(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\}_{n=1}^N$ defined in (9). The equality in (25) can be verified by plugging the GAMP posterior expression from (12) into (24) and simplifying via the Gaussian-pdf multiplication rule.

C. EM update for ϕ

Similar to (15), the EM update for ϕ can be written as

$$\hat{\phi}^{i+1} = \arg\max_{\phi>0} \sum_{n=1}^N \mathrm{E}\left\{\ln p_X(x_n;\lambda^i,\theta^i,\phi)\,\middle|\,\boldsymbol{y};\boldsymbol{q}^i\right\}. \quad (26)$$

The maximizing value of ϕ in (26) is necessarily a value of ϕ that zeroes the derivative, i.e., that satisfies

$$\sum_{n=1}^N \int_{x_n} p(x_n|\boldsymbol{y};\boldsymbol{q}^i)\frac{d}{d\phi}\ln p_X(x_n;\lambda^i,\theta^i,\phi) = 0. \quad (27)$$

For the $p_X(x_n;\lambda,\theta,\phi)$ given in (2), it is readily seen that

$$\frac{d}{d\phi}\ln p_X(x_n;\lambda^i,\theta^i,\phi) = \frac{1}{2}\left(\frac{|x_n-\theta^i|^2}{(\phi)^2}-\frac{1}{\phi}\right)\frac{\lambda^i\mathcal{N}(x_n;\theta^i,\phi)}{p_X(x_n;\lambda^i,\theta^i,\phi)} \quad (28)$$

$$= \begin{cases}\frac{1}{2}\left(\frac{|x_n-\theta^i|^2}{(\phi)^2}-\frac{1}{\phi}\right) & x_n\neq 0\\ 0 & x_n=0\end{cases}. \quad (29)$$

Splitting the domain of integration in (27) into $\mathcal{B}_\epsilon$ and $\overline{\mathcal{B}_\epsilon}$, and then plugging in (29), we find that the following is equivalent to (27) in the limit of $\epsilon\to 0$:

$$\sum_{n=1}^N \int_{x_n\in\overline{\mathcal{B}_\epsilon}} \left(|x_n-\theta^i|^2-\phi\right)p(x_n|\boldsymbol{y};\boldsymbol{q}^i) = 0. \quad (30)$$

The unique value of ϕ satisfying (30) as $\epsilon\to 0$ is then

$$\phi^{i+1} = \frac{\sum_{n=1}^N \lim_{\epsilon\to 0}\int_{x_n\in\overline{\mathcal{B}_\epsilon}} |x_n-\theta^i|^2 p(x_n|\boldsymbol{y};\boldsymbol{q}^i)}{\sum_{n=1}^N \lim_{\epsilon\to 0}\int_{x_n\in\overline{\mathcal{B}_\epsilon}} p(x_n|\boldsymbol{y};\boldsymbol{q}^i)} \quad (31)$$

$$= \frac{1}{\lambda^{i+1}N}\sum_{n=1}^N \pi(\hat{r}_n,\mu_n^r,\boldsymbol{q}^i)\Big(\left|\theta^i-\gamma(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\right|^2 + \nu(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\Big) \quad (32)$$

for the GAMP outputs $\{\nu(\hat{r}_n,\mu_n^r;\boldsymbol{q}^i)\}_{n=1}^N$ defined in (10). The equality in (32) can be verified by plugging the GAMP posterior expression from (12) into (31) and simplifying using the Gaussian-pdf multiplication rule.

[3] $\mathcal{N}(x;a,A)\mathcal{N}(x;b,B)=\mathcal{N}(x;\frac{a/A+b/B}{1/A+1/B},\frac{1}{1/A+1/B})\mathcal{N}(0;a-b,A+B)$.

[4] If the user has good reason to believe that the true signal pdf is symmetric around zero, then they may consider fixing $\theta=0$ and avoiding this EM update.

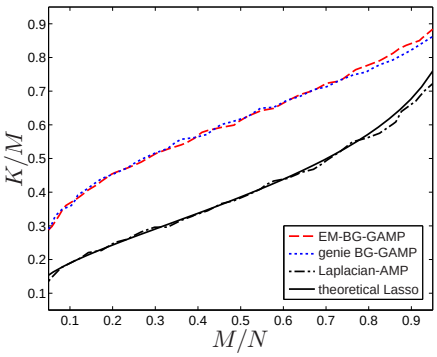

Fig. 1. Empirical noiseless PTCs for Bernoulli-Gaussian signals and theoretical PTC for Lasso.

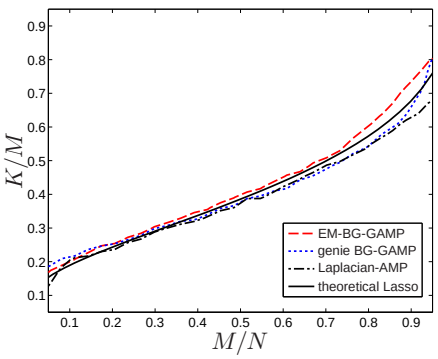

Fig. 2. Empirical noiseless PTCs for Bernoulli-Rademacher case and theoretical PTC for Lasso.

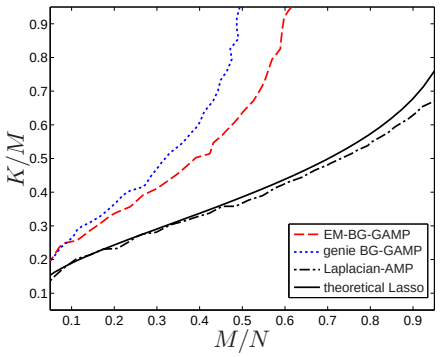

Fig. 3. Empirical noiseless PTCs for Bernoulli signals and theoretical PTC for Lasso.

D. EM update for ψ

Finally, we derive the EM update for ψ given previous parameters $\boldsymbol{q}^i$. Because $\boldsymbol{w}$ is apriori independent of $\boldsymbol{x}$ and i.i.d, the joint pdf $p(\boldsymbol{x},\boldsymbol{w};\boldsymbol{q})$ decouples into $C\prod_{m=1}^M p_W(w_m;\psi)$ for a ψ-invariant constant C, and so

$$\psi^{i+1} = \arg\max_{\psi>0} \sum_{m=1}^M \mathrm{E}\left\{ \ln p_W(w_m;\psi) \,\middle|\, \boldsymbol{y};\boldsymbol{q}^i \right\}. \tag{33}$$

The maximizing value of ψ in (33) is necessarily a value of ψ that zeroes the derivative, i.e., that satisfies

$$\sum_{m=1}^M \int_{w_m} p(w_m|\boldsymbol{y};\boldsymbol{q}^i) \frac{d}{d\psi} \ln p_W(w_m;\psi) = 0. \tag{34}$$

Because $p_W(w_m;\psi) = \mathcal{N}(w_m;0,\psi)$, it is readily seen that

$$\frac{d}{d\psi} \ln p_W(w_m;\psi) = \frac{1}{2}\left(\frac{|w_m|^2}{(\psi)^2} - \frac{1}{\psi} \right), \tag{35}$$

which, when plugged into (34), yields the unique solution

$$\psi^{i+1} = \frac{1}{M} \sum_{m=1}^M \int_{w_m} |w_m|^2 \, p(w_m|\boldsymbol{y};\boldsymbol{q}^i). \tag{36}$$

Since $w_m = y_m - z_m$ for $z_m \triangleq \boldsymbol{a}_m^{\mathsf{T}}\boldsymbol{x}$, we can also write[5]

$$\psi^{i+1} = \frac{1}{M} \sum_{m=1}^M \int_{z_m} |y_m - z_m|^2 \, p(z_m|\boldsymbol{y};\boldsymbol{q}^i) \tag{37}$$

$$= \frac{1}{M} \sum_{m=1}^M \left(|y_m - \hat{z}_m|^2 + \mu_m^z \right) \tag{38}$$

where $\hat{z}_m$ and μ_m^z, the posterior mean and variance of z_m, are available from GAMP (see steps (R1)-(R2) in Table I).

E. EM Initialization

Since the EM algorithm converges only to a local maximum of the likelihood function, proper initialization is essential. We initialize the sparsity as $\lambda^0 = \frac{M}{N}\rho_{\mathsf{SE}}(\frac{M}{N})$, where $\rho_{\mathsf{SE}}(\frac{M}{N})$ is the sparsity ratio $\frac{K}{M}$ achieved by the Lasso PTC [2]

$$\rho_{\mathsf{SE}}(\tfrac{M}{N}) = \max_{a\geq 0} \frac{1 - \frac{2N}{M}[(1+a^2)\Phi(a) - a\phi(a)]}{1 - a^2 - 2[(1+a^2)\Phi(a) - a\phi(a)]}. \tag{39}$$

[5]Empirically, we have observed that the EM update for ψ works better with the μ_m^z term in (38) weighted by $\frac{M}{N}$ and suppressed until later EM iterations. We conjecture that this is due to bias in the GAMP variance estimates μ_m^z.

We initialize the active mean as $\theta^0 = 0$, which effectively assumes that the active pdf $f(\cdot)$ is symmetric. Finally, noting that $\mathrm{E}\{\|\boldsymbol{y}\|_2^2\} = (\mathsf{SNR}+1)M\psi$ for $\mathsf{SNR} \triangleq \mathrm{tr}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})\lambda\phi/(M\psi)$, we see that the variances, ϕ and ψ, can be initialized based on $\|\boldsymbol{y}\|_2^2$ and a given hypothesis $\mathsf{SNR}^0 \geq 0$. In particular,

$$\psi^0 = \frac{\|\boldsymbol{y}\|_2^2}{(\mathsf{SNR}^0+1)M}, \quad \phi^0 = \frac{\|\boldsymbol{y}\|_2^2 - M\psi^0}{\mathrm{tr}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})\lambda^0}, \tag{40}$$

where, without other knowledge, we suggest $\mathsf{SNR}^0 = 100$.

IV. Numerical Results

A. Noiseless Phase Transitions

First, we describe the results of experiments that computed noiseless empirical phase transition curves (PTCs) under various sparse-signal distributions. To compute each empirical PTC, we constructed a 30×30 grid of oversampling ratio $\frac{M}{N} \in [0.05, 0.95]$ and sparsity ratio $\frac{K}{M} \in [0.05, 0.95]$ for fixed signal length $N = 1000$. At each grid point, we generated $R = 100$ independent realizations of K-sparse signal $\boldsymbol{x}$ and $M \times N$ i.i.d-Gaussian measurement matrix $\boldsymbol{A}$. From the measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, we attempted to reconstruct the signal $\boldsymbol{x}$ using various algorithms. A recovery $\hat{\boldsymbol{x}}$ from realization $r \in \{1,\ldots,R\}$ was considered a success (i.e., $S_r = 1$) if NMSE $\triangleq \|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2/\|\boldsymbol{x}\|_2^2 < 10^{-4}$, where the average success rate is defined as $\overline{S} \triangleq \frac{1}{R}\sum_{r=1}^R S_r$. The empirical PTC was then plotted, using Matlab's `contour` command, as the $\overline{S} = 0.5$ contour over the sparsity-undersampling grid.

Figures 1–3 show the empirical PTCs for three recovery algorithms: the proposed EM-BG-GAMP algorithm,[6] a "genie-aided" BG-GAMP that knew the true $[\lambda,\theta,\phi,\psi]$, and the Laplacian-AMP from [2]. For comparison, Figs. 1–3 also display the theoretical Lasso PTC (39). The signal was generated as BG with zero mean ($\theta = 0$) and unit variance ($\phi = 1$) in Fig. 1, as Bernoulli-Rademacher (BR) in Fig. 2 (i.e., non-zero coefficients chosen uniformly from $\{-1, 1\}$), and as Bernoulli in Fig. 3 (i.e., all non-zero coefficients set equal to 1 or, equivalently, BG with $\theta = 1$ and $\phi = 0$).

Figures 1–3 demonstrate that, for all three signal types, the empirical PTC of EM-BG-GAMP improves on that for Laplacian-AMP as well as the theoretical Lasso PTC. (The latter two are known to converge in the large system limit

[6]Matlab code available at http://www.ece.osu.edu/~schniter/EMturboGAMP

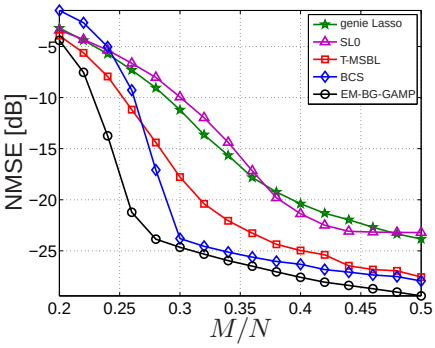

Fig. 4. NMSE for noisy recovery of a Bernoulli-Gaussian signal.

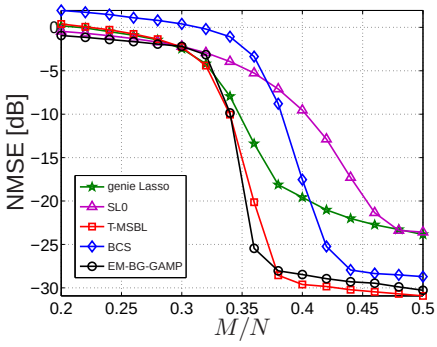

Fig. 5. NMSE for noisy recovery of a Bernoulli-Rademacher signal.

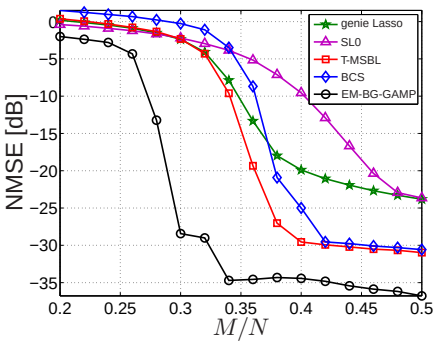

Fig. 6. NMSE for noisy recovery of a Bernoulli signal.

[2].) The smallest gains over Lasso appear when the signal is BR (i.e., the least-favorable distribution [3]), whereas the largest gains appear when the signal is Bernoulli. Amazingly, EM-BG-AMP perfectly recovered almost every Bernoulli realization when $\frac{M}{N} \geq 0.65$. The figures suggest that the EM algorithm does a decent job of learning the parameters λ, θ, ϕ. In fact, EM-BG-GAMP slightly outperforms genie-BG-GAMP in Figs. 1–2 due to realization-specific data fitting.

B. Noisy Signal Recovery

Figures 4–6 show NMSE for noisy recovery of the same three sparse signal types considered in Figs. 1–3. To construct these new plots, we fixed $N = 1000$, $K = 100$, SNR $= 25$dB, and varied M. Each data point represents NMSE averaged over $R = 500$ realizations. For comparison, we show the performance of the proposed EM-BG-GAMP, Bayesian Compressive Sensing (BCS) [9], Sparse Bayesian Learning [8] (via T-MSBL), debiased genie-aided[7] Lasso (via SPGL1 [12]), and Smoothed-ℓ_0 (SL0) [13]. All algorithms were run under the suggested defaults, with`'noise'='small'` in T-MSBL.

In Fig. 4 and Fig. 6, we see EM-BG-GAMP outperforming all other algorithms for all meaningful values of undersampling ratio $\frac{M}{N}$. In fact, for Bernoulli signals (Fig. 6), we see a significant improvement, especially when $\frac{M}{N} \in [0.3, 0.38]$. We have verified (in simulations not shown here) that similar behavior persists at lower SNRs. In Fig. 5, we see EM-BG-GAMP outperforming all algorithms except T-MSBL, which does $\approx$1 dB better for large values of $\frac{M}{N}$. Apparently, the prior assumed by T-MSBL is a better fit to BR than the BG prior.

Admittedly, the near-dominant EM-BG-GAMP performance observed for *perfectly* sparse signals in Figs. 1–6 does not hold for all signal classes. As an example, Fig. 7 shows noisy recovery NMSE for a Student's-t signal with pdf

$$p_X(x;q) \triangleq \frac{\Gamma((q+1)/2))}{\sqrt{2\pi}\Gamma(q/2)} \left(1+x^2\right)^{-(q+1)/2} \qquad (41)$$

under the *non-compressible* parameter choice $q = 1.67$ [14]. There, we see EM-BG-GAMP outperformed by SL0 and genie-aided Lasso, although not by T-MSBL and BCS. In fact, among the competing algorithms, those that performed best for exactly sparse signals seem to do worst for this non-compressible signal, and vice versa. We attribute these behaviors to a poor fit between the assumed and actual signal priors, motivating future work on an EM *Gaussian-Mixture* GAMP with automatic selection of the mixture order.

[7]We ran SPGL1 in 'BPDN' mode: $\min_{\hat{\boldsymbol{x}}} \|\boldsymbol{x}\|_1$ s.t. $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \sigma$, for tolerances $\sigma^2 \in \{0.1, 0.2, \ldots, 1.5\} \times M\psi$, and reported the lowest NMSE.

REFERENCES

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

[2] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. National Academy of Sciences*, vol. 106, pp. 18914–18919, Nov. 2009.

[3] D. L. Donoho, A. Maleki, and A. Montanari, "The noise-sensitivity phase transition in compressed sensing," *arXiv:1004.1218*, Apr. 2010.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. Inform. Theory Workshop*, (Cairo, Egypt), Jan. 2010.

[5] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, pp. 764–785, Feb. 2011.

[6] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *arXiv:1010.5141*, Oct. 2010.

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

[8] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, pp. 2153 – 2164, Aug. 2004.

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

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

[11] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models* (M. I. Jordan, ed.), pp. 355–368, MIT Press, 1999.

[12] E. van den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Scientific Comput.*, vol. 31, no. 2, pp. 890–912, 2008.

[13] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed norm," *IEEE Trans. Signal Process.*, vol. 57, pp. 289–301, Jan. 2009.

[14] V. Cevher, "Learning with compressible priors," in *Proc. Neural Inform. Process. Syst. Conf.*, (Vancouver, B.C.), Dec. 2009.

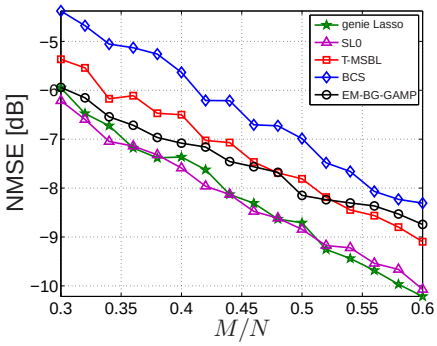

Fig. 7. NMSE for noisy recovery of a non-compressible Student's-t signal.