Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines

Eric Tramel, Andre Manoel, Francesco Caltagirone, Marylou Gabrié, Florent Krzakala

Inferring Sparsity: Compressed Sensing using Generalized Restricted Boltzmann Machines

Eric W. Tramel[†], Andre Manoel[†], Francesco Caltagirone[‡], Marylou Gabrié[†] and Florent Krzakala[†§]
[†]Laboratoire de Physique Statistique (CNRS UMR-8550),
École Normale Supérieure, PSL Research University, 24 rue Lhomond, 75005 Paris, France
[§]Université Pierre et Marie Curie, Sorbonne Universités, 75005 Paris, France
[‡]INRIA Paris, 2 rue Simone Iff, 75012 Paris, France

***Abstract*—In this work, we consider compressed sensing reconstruction from M measurements of K-sparse structured signals which do not possess a writable correlation model. Assuming that a generative statistical model, such as a Boltzmann machine, can be trained in an unsupervised manner on example signals, we demonstrate how this signal model can be used within a Bayesian framework of signal reconstruction. By deriving a message-passing inference for general distribution restricted Boltzmann machines, we are able to integrate these inferred signal models into approximate message passing for compressed sensing reconstruction. Finally, we show for the MNIST dataset that this approach can be very effective, even for $M < K$.**

I. Introduction

Over the past decade, the study of compressed sensing (CS) [1–3] has lead to many significant developments in the field of signal processing including novel sub-Nyquist sampling strategies [4, 5] and a veritable explosion of work in sparse approximation and representation [6]. The core problem in CS is the reconstruction of a sparse signal of dimensionality N from a set of M noisy observations for $M \ll N$. Here, a *sparse* signal is defined as one which possesses many fewer non-zero coefficients, K, than its ambient dimensionality, $K \ll N$. The theoretical foundations of CS recovery conditions are built upon the concept of *support identification*, finding the locations of these non-zero coefficients. Tackling the reconstruction directly is combinatorially hard, requiring a search over all $\binom{N}{K}$ possible support patterns for the one which best matches the given observations. In the noiseless setting, if such a search were possible, we require at least $M = K$ for the on-support values to be perfectly estimated. As we do not assume any particular distribution on the K non-zero values, this requirement follows directly from linear algebra. It was shown in [2] that a convex relaxation from a strict requirement on K-sparsity to an ℓ_1 regularization allows for efficient reconstruction, but requires $M \gtrsim K \log N$ for exact reconstruction. Greedy approaches [7, 8] retain K-sparsity, solving the reconstruction problem by iterating between support identification and on-support signal estimation. However, these robust and conceptually simple techniques come at the cost of increased M for exact reconstruction.

Since the identification of sharp thresholds between successful and unsuccessful signal recovery as a function of $\alpha = M/N$ and $\rho = K/N$ [9], these *phase transitions* in the space of possible CS reconstruction problems have been used as a tool for comparing different CS recovery strategies. In [10, 11], it was shown that using sum-product belief propagation (BP) in conjunction with a two-mode Gauss-Bernoulli sparse signal prior provided a much more favorable phase transition, as compared to ℓ_1 minimization, for K-sparse signal reconstruction. In [12], approximate message passing (AMP) was proposed as an efficient alternative to BP and was further refined in a series of papers [13–15].

While these Bayesian techniques have had a significant impact in improving the lower bound on M for exact reconstruction, the $M = K$ lower bound on general recoverability can only be approached in the case that directly designing the sampling strategy [16] is possible. As such designs are often not practically achievable, decreasing the requirement on M necessitates the use of more informative signal priors, e.g. a prior which leverages *a priori* known correlations for the signal class of interest. The works [17, 18] sought to model support correlation directly by leveraging the abstraction power of latent variable models via Boltzmann machines trained on examples of signal support. While these techniques demonstrated significant improvements in recovery performance for sparse signals, they are still fundamentally bound by the $M = K$ transition. The unification of machine learning approaches with CS was also addressed in the recent works [19–21], in which feed-forward deep neural network structures are used to aid CS signal reconstruction.

In this work, we investigate the possibility of modeling both signal and support, as in [22, 23], using however a trained latent variable model as prior for the AMP reconstruction. For this, we turn to real-valued restricted Boltzmann machines (RBMs). In order to utilize real-valued RBMs within the AMP framework, we propose an extended mean-field approximation similar in nature to [18, 24]. However, we extend this approximation to the case of *general distributions* on both hidden and visible units of the RBM, allowing us to model sparse signals directly. Given this trained RBM, we propose a CS reconstruction algorithm which amounts to two nested inference problems, one on the CS observation-matching problem, and the other on the RBM model. In our results, we show that this technique can provide good reconstructions even for $M < K$, as the RBM signal prior not only models the support correlation structure, but the joint distribution of the on-support values, as well.

II. Background

In the CS problem, we wish to recover some unknown K-sparse signal $\mathbf{x} \in \mathbb{R}^N$ given a set of observations $\mathbf{y} \in \mathbb{R}^M$, $M \ll N$, generated by $\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{w}$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Delta\mathbf{I})$ and the matrix $\mathbf{F} \in \mathbb{R}^{M\times N}$ is a random projection operator. While a number of different output channels of the form $\mathbf{y} = g(\mathbf{F}\mathbf{x})$ could be conceived [14], for clarity we focus on the case of an additive white Gaussian noise (AWGN) channel.

Following the Bayesian approach to signal reconstruction, we will focus on estimation techniques involving the posterior distribution

$$P(\mathbf{x}|\mathbf{F},\mathbf{y}) = \frac{e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2}\, P_0(\mathbf{x})}{\int \mathrm{d}\mathbf{x}\;\; e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{F}\mathbf{x}\|_2^2}\, P_0(\mathbf{x})}. \tag{1}$$

Even if computing the moments of (1) is intractable for some $P_0(\mathbf{x})$, [12, 15] show that the minimum mean-square-error (MMSE) estimator, $\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{F},\mathbf{y}) = \int \mathrm{d}\mathbf{x}\;\; \mathbf{x}\, P(\mathbf{x}|\mathbf{F},\mathbf{y})$, can be computed extremely efficiently using loopy BP or AMP whenever $P_0(\mathbf{x})$ is fully factorized.

The AMP algorithm [12, 14, 15], under stricter requirements of i.i.d. random $\mathbf{F}$, provides at each step of its iteration, an approximation to the posterior of the form

$$Q(\mathbf{x}|\boldsymbol{A},\boldsymbol{B}) = \frac{1}{\mathcal{Z}(\boldsymbol{A},\boldsymbol{B})}\, P_0(\mathbf{x})\, e^{-\frac{1}{2}\sum_i A_i x_i^2 + \sum_i B_i x_i}, \tag{2}$$

where $\mathcal{Z}(\boldsymbol{A},\boldsymbol{B})$ is a normalization, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are quantities obtained by iterating the following AMP equations,

$$V_m^{(\mathrm{t}+1)} = \sum_i F_{mi}^2 c_i^{(\mathrm{t})}, \tag{3}$$

$$\omega_m^{(\mathrm{t}+1)} = \sum_i F_{mi} a_i^{(\mathrm{t})} - V_m^{(\mathrm{t}+1)} \frac{y_m - \omega_m^{(\mathrm{t})}}{\Delta + V_m^{(\mathrm{t})}}, \tag{4}$$

$$A_i^{(\mathrm{t}+1)} = \sum_m \frac{F_{mi}^2}{\Delta + V_m^{(\mathrm{t}+1)}}, \tag{5}$$

$$B_i^{(\mathrm{t}+1)} = A_i^{(\mathrm{t}+1)} a_i^{(\mathrm{t})} + \sum_m F_{mi}\, \frac{y_m - \omega_m^{(\mathrm{t}+1)}}{\Delta + V_m^{(\mathrm{t}+1)}}, \tag{6}$$

where $\mathbf{a}^{(\mathrm{t})}$ and $\mathbf{c}^{(\mathrm{t})}$ are the mean and variance of $Q(\mathbf{x}|\boldsymbol{A}^{(\mathrm{t})},\boldsymbol{B}^{(\mathrm{t})})$, which after the convergence of the algorithm provide an approximation of the mean and variance of (1). These moments, given a computable $\mathcal{Z}(\boldsymbol{A},\boldsymbol{B}) = \int \mathrm{d}\mathbf{x}\, P_0(\mathbf{x})\, e^{-\frac{1}{2}\sum_i A_i x_i^2 + \sum_i B_i x_i}$, are easily obtainable from

$$a_i \triangleq \frac{\partial \ln \mathcal{Z}(\boldsymbol{A},\boldsymbol{B})}{\partial B_i}, \quad c_i \triangleq \frac{\partial^2 \ln \mathcal{Z}(\boldsymbol{A},\boldsymbol{B})}{\partial B_i^2}. \tag{7}$$

In particular, whenever the prior distribution is fully factorized, $P_0(\mathbf{x}) = \prod_i P_0(x_i)$, evaluating $\mathcal{Z}(\boldsymbol{A},\boldsymbol{B})$ amounts to solving N independent one-dimensional integrals. In contrast, for general $P_0(\mathbf{x})$, the normalization is intractable. In this case one must resort to further approximations.

In what follows, we use an RBM [25] to model the signal's prior distribution jointly with a set of latent, or *hidden*, variables $\mathbf{h}$,

$$P_0(\mathbf{x},\mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\; e^{\mathbf{x}^T\mathbf{W}\mathbf{h}} \prod_i P_0(x_i|\boldsymbol{\theta}_\mathbf{x}) \prod_\mu P_0(h_\mu|\boldsymbol{\theta}_\mathbf{h}), \tag{8}$$

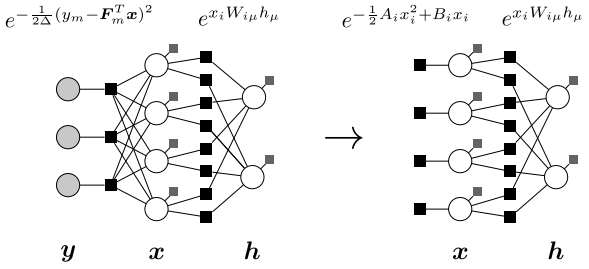

Fig. 1. *Left:* Factor graph representation of the posterior (1) given the RBM signal prior (8). Circles and squares represent variables and factors, respectively. Light gray circles are the observed signal measurements. Black factors are the factors induced by linear operators $\mathbf{F}$ and $\mathbf{W}$. Light gray factors represent prior distributions influencing their adjoining variables. *Right:* Factor graph for the approximating posterior (2). Factors on the left represent local effective potentials provided by AMP at each of its iterations.

and the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\theta}_\mathbf{x}, \boldsymbol{\theta}_\mathbf{h}\}$ can be obtained by training the RBM over a set of examples [24, 26]. This construction defines a *generalized* RBM (GRBM), in the sense that the visible and hidden variables are not strictly binary and may possess any distribution. Using a GRBM as a signal prior, the normalization of (2),

$$\mathcal{Z}(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{Z(\boldsymbol{\theta})} \int \left[\prod_i \mathrm{d}x_i\, P_0(x_i)\, e^{-\frac{A_i}{2}x_i^2 + B_i x_i}\right] \times \int \left[\prod_\mu \mathrm{d}h_\mu\, P_0(h_\mu)\right] e^{\mathbf{x}^T\mathbf{W}\mathbf{h}}, \tag{9}$$

is no longer factorized or tractable, thus requiring some approximation to calculate the necessary moments of Q. In the next section we introduce a message-passing algorithm to evaluate (9) and estimate these moments.

III. TAP Approximation of a General-case RBM

In order to use the RBM model (8) as a prior within AMP, we must perform inference over the RBM graphical model given on the right of Fig. 1. Specifically, we construct a message-passing scheme between the factors and the hidden and visible variables of the RBM. Loopy BP has been considered in the context of inference on RBMs in [27], where the authors assume a Bernoulli distribution on the hidden variables and rewrite them as factors. In contrast, we investigate a more general setting with arbitrary distributions on both the hidden and visible variables. Since all factors have degree 2, using BP we can write the messages from variable to variable,

$$\psi_{i\to\mu}(x_i) = \frac{1}{\mathcal{Z}_{i\to\mu}} P_0(x_i) e^{-\frac{1}{2}A_i^{\text{AMP}}x_i^2 + B_i^{\text{AMP}}x_i} \times \prod_{\nu\in\partial i/\mu} \int \mathrm{d}h_\nu \;\; \psi_{\nu\to i}(h_\nu) e^{x_i W_{i\nu} h_\nu}, \tag{10}$$

$$\psi_{\mu\to i}(h_\mu) = \frac{1}{\mathcal{Z}_{\mu\to i}} P_0(h_\mu) \times \prod_{j\in\partial\mu/i} \int \mathrm{d}x_j \;\; \psi_{j\to\mu}(x_j) e^{x_j W_{j\mu} h_\mu}, \tag{11}$$

where we index hidden units by μ, ν and visible units by i, j, and the notation $\partial i/\mu$ represents the set of all edge-connected neighbors of variable i *except* μ. If we assume that the magnitudes of the values of $\mathbf{W}$ are small, then we can perform an expansion on the messages [15]. This is in essence the relaxed BP of [11], transitioning from messages of continuous distributions to messages parameterized by their first two central moments, denoted by the letters a and c, respectively. Specifically,

$$\int \mathrm{d}h_\nu \;\; \psi_{\nu\to i}(h_\nu) e^{x_i W_{i\nu} h_\nu} = \exp\left\{ x_i W_{i\nu} a^{\mathrm{h}}_{\nu\to i} + \frac{1}{2} x_i^2 W_{i\nu}^2 c^{\mathrm{h}}_{\nu\to i} \right\} + O(\mathbf{W}^3), \quad (12)$$

$$\int \mathrm{d}x_j \;\; \psi_{j\to\mu}(x_j) e^{x_j W_{j\mu} h_\mu} = \exp\left\{ h_\mu W_{j\mu} a^{\mathrm{v}}_{j\to\mu} + \frac{1}{2} h_\mu^2 W_{j\mu}^2 c^{\mathrm{v}}_{j\to\mu} \right\} + O(\mathbf{W}^3), \quad (13)$$

where $O(\mathbf{W}^3)$ represents a vanishing third-order correction to the two-moment approximation of the message marginalizations which is dependent on the RBM parameters $\mathbf{W}$. As we are now able to move the product of incoming messages as a sum into the exponent, we define the following intermediate sum variables as in AMP,

$$A^{\mathrm{v}}_{i\to\mu} \triangleq -\sum_{\nu\in H/\mu} W_{i\nu}^2 c^{\mathrm{h}}_{\nu\to i}, \;\; B^{\mathrm{v}}_{i\to\mu} \triangleq \sum_{\nu\in H/\mu} W_{i\nu} a^{\mathrm{h}}_{\nu\to i}, \quad (14)$$

$$A^{\mathrm{h}}_{\mu\to i} \triangleq -\sum_{j\in V/i} W_{j\mu}^2 c^{\mathrm{v}}_{j\to\mu}, \;\; B^{\mathrm{h}}_{\mu\to i} \triangleq \sum_{j\in V/i} W_{j\mu} a^{\mathrm{v}}_{j\to\mu}, \quad (15)$$

where the sets H and V are the set of all hidden and visible variables, respectively. From these definitions, we can define the message moments explicitly and close the message passing equations on the edges of the RBM factor graph,

$$a^{\mathrm{h}}_{\mu\to i} = f^{\mathrm{h}}_{\mathrm{a}}(A^{\mathrm{h}}_{\mu\to i}, B^{\mathrm{h}}_{\mu\to i}), \;\; c^{\mathrm{h}}_{\mu\to i} = f^{\mathrm{h}}_{\mathrm{c}}(A^{\mathrm{h}}_{\mu\to i}, B^{\mathrm{h}}_{\mu\to i}), \quad (16)$$

$$a^{\mathrm{v}}_{i\to\mu} = f^{\mathrm{v}}_{\mathrm{a}}(A^{\mathrm{AMP}}_i + A^{\mathrm{v}}_{i\to\mu}, B^{\mathrm{AMP}}_i + B^{\mathrm{v}}_{i\to\mu}), \quad (17)$$

$$c^{\mathrm{v}}_{i\to\mu} = f^{\mathrm{v}}_{\mathrm{c}}(A^{\mathrm{AMP}}_i + A^{\mathrm{v}}_{i\to\mu}, B^{\mathrm{AMP}}_i + B^{\mathrm{v}}_{i\to\mu}), \quad (18)$$

where the prior-dependent functions for the visible and hidden variables, $(f^{\mathrm{v}}_{\mathrm{a}}, f^{\mathrm{v}}_{\mathrm{c}})$ and $(f^{\mathrm{h}}_{\mathrm{a}}, f^{\mathrm{h}}_{\mathrm{c}})$ respectively, are defined in a fashion similar to (7), i.e. as the moments of an approximating distribution similar to (2), but using the desired hidden and visible distributions in place of P_0. The marginal beliefs at each hidden and visible variable can be defined by summing over *all* incoming messages. One could run this message passing, as stated, until convergence on the beliefs in order to infer marginal distributions on both the hidden and visible variables of the RBM. The moments of the marginal distributions on the visible units, a^{v}_i and c^{v}_i, would give us exactly the moments required by the AMP iteration.

However, such a message passing on the edges of the factor graph can be quite memory and computationally intensive, especially if this inference occurs nested as an inner loop of AMP. If we assume that the entries of $\mathbf{W}$ are widely distributed, without any particular strong correlations in its structure, we can construct an algorithm which operates entirely on the beliefs, the nodes of the factor graph, rather than the messages, the edges. Such an algorithm is similar in spirit to AMP and also to the Thouless-Anderson-Palmer (TAP) equations from statistical physics. We now write these TAP self-consistency equations closed on the parameters of the marginal beliefs alone,

$$A^{\mathrm{h}}_\mu = -\sum_{i\in V} W_{i\mu}^2 c^{\mathrm{v}}_i, \quad B^{\mathrm{h}}_\mu = a^{\mathrm{h}}_\mu A^{\mathrm{h}}_\mu + \sum_{i\in V} W_{i\mu} a^{\mathrm{v}}_i, \quad (19)$$

$$a^{\mathrm{h}}_\mu = f^{\mathrm{h}}_{\mathrm{a}}(A^{\mathrm{h}}_\mu, B^{\mathrm{h}}_\mu), \quad c^{\mathrm{h}}_\mu = f^{\mathrm{h}}_{\mathrm{c}}(A^{\mathrm{h}}_\mu, B^{\mathrm{h}}_\mu), \quad (20)$$

$$A^{\mathrm{v}}_i = -\sum_{\mu\in H} W_{i\mu}^2 c^{\mathrm{h}}_\mu, \quad B^{\mathrm{v}}_i = a^{\mathrm{v}}_i A^{\mathrm{v}}_i + \sum_{\mu\in H} W_{i\mu} a^{\mathrm{h}}_\mu, \quad (21)$$

$$a_i = f^{\mathrm{v}}_{\mathrm{a}}(A^{\mathrm{AMP}}_i + A^{\mathrm{v}}_i, B^{\mathrm{AMP}}_i + B^{\mathrm{v}}_i), \quad (22)$$

$$c_i = f^{\mathrm{v}}_{\mathrm{c}}(A^{\mathrm{AMP}}_i + A^{\mathrm{v}}_i, B^{\mathrm{AMP}}_i + B^{\mathrm{v}}_i). \quad (23)$$

Algorithm 1 AMP with GRBM Signal Prior

Input: $\mathbf{F}$, $\mathbf{y}$, $\mathbf{W}$, θ^{v}, θ^{h}
Initialize: $\mathbf{a}$,$\mathbf{c}$, $\mathrm{t} = 1$
Outer AMP Inference Loop:
repeat
  AMP Update on $\{V_m, \omega_m\}$ as in (3), (4)
  AMP Update on $\{A^{\mathrm{AMP}}_i, B^{\mathrm{AMP}}_i\}$ as in (5), (6)
  (Re)Initialize: $a_i = f^{\mathrm{v}}_{\mathrm{a}}(A^{\mathrm{AMP}}_i, B^{\mathrm{AMP}}_i)\ \forall i$, $a^{\mathrm{h}}_\mu = c^{\mathrm{h}}_\mu = 0\ \forall \mu$
  Inner RBM Inference Loop:
  repeat
    Update $\{A^{\mathrm{v}}_i, B^{\mathrm{v}}_i\}$ as in (21)
    Update $\{a_i, c_i\}$ as in (22), (23)
    Update $\{A^{\mathrm{h}}_\mu, B^{\mathrm{h}}_\mu\}$ as in (19)
    Update $\{a^{\mathrm{h}}_\mu, c^{\mathrm{h}}_\mu\}$ as in (20)
  until Convergence
  $\mathbf{a}^{(\mathrm{t})} = \gamma \cdot \mathbf{a}^{(\mathrm{t}-1)} + (1-\gamma)\cdot \mathbf{a}$
  $\mathbf{c}^{(\mathrm{t})} = \gamma \cdot \mathbf{c}^{(\mathrm{t}-1)} + (1-\gamma)\cdot \mathbf{c}$
  $\mathrm{t} \leftarrow \mathrm{t}+1$
until Convergence on $\mathbf{a}$

IV. Implementation

Using the equations detailed in (19)–(23), we can construct a fixed-point iteration (FPI) which, given some arbitrary starting condition, can be run until convergence in order to obtain the GRBM-inferred distribution on the signal variables defined by $\mathbf{a}$, $\mathbf{c}$. These distributions are then passed back to the CS observational factors to complete the AMP iteration for CS reconstruction. We detail this procedure in Alg. 1. One important addition is the use of a damping step [28] on the RBM-inferred values of $\mathbf{a}$ and $\mathbf{c}$. We find that a fixed value of $\gamma = 0.5$, equally combining the previous and presently inferred moments, stabilizes the interaction between the GRBM and the outer AMP inference. This becomes especially important for α small, where oscillations between the two inference loops degrades reconstruction performance.

While the Hamiltonian for the RBM model used here is in agreement with the literature on real-valued RBMs, the TAP FPI on the RBM points out one flaw in the construction of the real-valued RBM model. Specifically, the unbounded nature of the energy for variables in $\mathbb{R}$ manifests in this context by allowing *negative* variance-like terms A^{v}_i and A^{h}_μ which

carry through into the prior-dependent functions. We propose to handle this dilemma of negative variance by forcing the truncation of $P_0(h_\mu)$ and $P_0(x_i)$. Truncation can gracefully handle negative variances by transforming these distributions to uniformity over the truncation bounds. In practice, truncation is an easy assumption to make on both the training data and the signal to be reconstructed. For example, image data naturally lies within a specific range of values as defined by the images' bit-depth. For sparse signals, such as those commonly studied in the context of CS, we propose the use of a *truncated* Gauss-Bernoulli distribution on the visible units, which only has non-zero probability density within a fixed range.

V. Experiments

We now present the results of our numerical studies of GRBM-AMP performance for the AWGN CS reconstruction task. For all reconstruction tasks, an AWGN noise variance $\Delta = 10^{-8}$ was used. Additionally, all elements of the sampling matrices $\mathbf{F}$ were drawn from a zero-mean Gaussian distribution of variance $1/\sqrt{N}$.

The results we present are based on the MNIST handwritten digit dataset [29] which consists of 28×28 gray-scale digit images split between 60,000 training samples and 10,000 test samples. While we train over the entire MNIST training partition, we conduct our reconstruction experiments for the first 1,000 digit images drawn from the test partition. We test three different approaches for this dataset. The first, termed non-i.i.d. AMP, consists of empirically estimating the per-coefficient prior hyper-parameters from the training data. This approach assumes a fully factorized model of the data, neglecting any covariance structure between the coefficients. The second approach is that of [18], here termed binary-support RBM (BRBM-AMP), which uses a binary RBM to model the correlation structure of the support, alone. Finally, we test the proposed GRBM-AMP, using a general RBM trained with binary hidden units and Gauss-Bernoulli visible units, which models the data in its ambient domain.

To train the GRBM parameters for MNIST, we use a GRBM with 784 binary hidden units. The RBM can be trained using either contrastive divergence, sampling the visible units from a truncated GB prior, or using the GRBM TAP iteration shown here in conjunction with the EMF strategy of [24], a strategy which we detail in a forthcoming work. For the specific GRBM model we use for these CS experiments, we train a GRBM using the EMF approach for 150 epochs using a learning rate of 0.01 with an ℓ_2 weight decay penalty of 0.001. Learning momentum of 0.5 was used. Finally, the truncation bounds were set to the range $[0, 1]$. In the case of the BRBM, where all variables are binary, EMF training [24] was used with a learning rate of 0.005, while other parameters have been set the same as for the GRBM.

We present the results of the three approaches in Fig. 2, where we evaluate reconstruction performance over the test set in terms of both MSE, measured in decibels, and correlation between the reconstructed and original images, where correlation is measured as $(\mathbf{x} - \bar{x})^T(\mathbf{a} - \bar{a})/\sigma_{\mathbf{x}}\sigma_{\mathbf{a}}$. In both of these comparisons, we show performance over the phase diagram, where α refers to the number of CS measurements observed and ρ refers to the overall sparsity for each digit image, K/N, where K is the number of non-zero pixels in the digit image. As many of the tested images posses few non-zeros, the test dataset is skewed towards small ρ, hence the increased variability at $\rho > 0.3$ for Fig. 2. From these results, we can see a clear progression of reconstruction performance as we move from an empirical factorized model, to a model of the support alone, to a model of the signal itself. For non-i.i.d. AMP, we see that the localized information from the training set allows for a transition curve which is parallel to $M = K$, which is in contrast to the well-known transition for AMP with an i.i.d. GB signal prior. For BRBM-AMP we observe that the reconstruction transition lies very close to the $M = K$ optimal line, as also observed in [18], showing the advantage of leveraging the strong support correlations which exist in the dataset. In the case of the GRBM-AMP, we observe that the maximal performance is no longer bounded by the $M = K$ line, with low MSE achievable even for $M < K$. This is an intuitive result, as the GRBM model provides information not only about the support, but also about the values of the signal on that support. This effect is most drastically observed in terms of the average reconstruction correlation, where we observe almost perfect correlation for the entire test set for $\alpha > 0.10$, independent of the signal sparsity.

Finally, we also see a visual comparison for one example of a reconstructed digit in Fig. 2 in the small α setting. In this extreme case, we can see that both the BRBM- and GRBM-AMP produce reconstructions whose support closely match the original '6' digit image. We note that the GRBM is able to capture the smoothness of the digit image where the BRBM cannot.

VI. Conclusion

In this work, we derived an AMP-based algorithm for CS signal reconstruction using an RBM to model the signal class in its ambient domain. To accomplish this modeling, we developed a model for a class of general RBMs, allowing for arbitrary distributions on the hidden and visible units. To allow the use of such a model within AMP, we proposed a TAP-based approximation of the RBM which we derived from belief propagation. By performing inference on the RBM under the influence of the outer AMP inference, we have developed a novel algorithm for CS reconstruction of sparse structured data. The proposed approach can be of great use in signal reconstruction contexts where there exists an abundance of data which lack developed correlation models. Additionally, the inference we propose for general RBMs could be used to develop novel generative models by varying its architecture and the distribution of the hidden variables.

Acknowledgment

This research was funded by European Research Council under the European Unions 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCS).

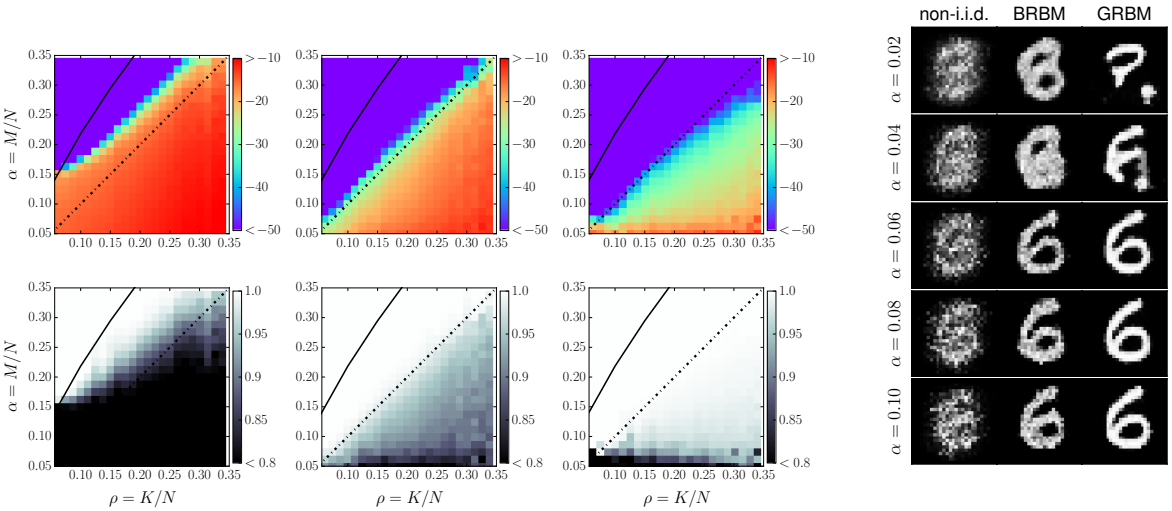

Fig. 2. **(Left)** CS reconstruction performance over first 1,000 digit images from the MNIST test partition. Results for non-i.i.d. AMP, support-based BRBM-AMP, and GRBM-AMP are on the left, center, and right, respectively. The $M = K$ oracle support transition is indicated by the black dotted line, and the spinodal transition [15] by the solid one. *Top:* Average reconstruction accuracy in MSE measured in dB. *Bottom:* Average reconstruction correlation with original digit image. **(Right)** Visual comparison of reconstructions for a single digit image ($\rho = 0.25$) for small values of α.

REFERENCES

[1] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[2] E. Candès and J. Romberg, "Signal recovery from random projections," in *Computational Imaging III*. Proc. SPIE 5674, Mar. 2005, pp. 76–86.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] D. Takhar, J. N. Laska, M. B. Wakin, M. F. Duarte, D. Baron, S. Sarvotham, K. F. Kelly, and R. G. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," in *Computational Imaging IV*. Proc. SPIE 6065, 2006, p. 606509.

[5] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices and Systems*, vol. 5, no. 1, pp. 8–20, January 2011.

[6] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 290–530, 2015.

[7] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," Stanford University, Tech. Rep., 2006.

[8] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, Oct. 2008, pp. 581–587.

[9] D. L. Donoho and J. Tanner, "Thresholds for the recovery of sparse solutions via l1 minimization," in *Information Sciences and Systems, Proc. Annual Conference on*, 2006, pp. 202–206.

[10] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, pp. 269–280, 2009.

[11] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Annual Conf. on Information Sciences and Systems*, 2010, pp. 1–6.

[12] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Academy of Sciences of the U.S.A.*, vol. 106, no. 45, p. 18914, 2009.

[13] ——, "Message passing algorithms for compressed sensing: II. analysis and validation," in *Proc. IEEE Info. Theory Workshop*, Cairo, Egypt, 2010, pp. 1–5.

[14] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Intl. Symp. on Info. Theory*, 2011, p. 2168.

[15] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 8, p. P08009, 2012.

[16] ——, "Statistical physics-based reconstruction in compressed sensing," *Phys. Rev. X*, vol. 2, p. 021005, 2012.

[17] A. Drémeau, C. Herzet, and L. Daudet, "Boltzmann machine and mean-field approximation for structured sparse decompositions," *IEEE Trans. on Signal Processing*, vol. 60, no. 7, pp. 3425–3438, 2012.

[18] E. W. Tramel, A. Drémeau, and F. Krzakala, "Approximate message passing with restricted Boltzmann machine priors," *Journal of Statistical Mechanics: Theory and Experiment*, 2016, to appear.

[19] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "ReconNet: Non-iterative reconstruction of images from compressively sensed random measurements," 2016, arXiv:1601.06892.

[20] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A deep learning approach to structured signal recovery," 2015, arXiv:1508.04065.

[21] U. S. Kamilov and H. Mansour, "Learning optimal nonlinearities for iterative thresholding algorithms," 2015, arXiv:1512.04754.

[22] P. Schniter, "Turbo reconstruction of structured sparse signals," in *Proc. Conf. on Info. Sciences and Systems*, 2010, pp. 1–6.

[23] S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid approximate message passing with applications to structured sparsity," 2011, arXiv:1111.2581.

[24] M. Gabrié, E. W. Tramel, and F. Krzakala, "Training restricted Boltzmann machines via the Thouless-Andreson-Palmer free energy," in *Advances in Neural Information Processing System*, vol. 28, Montreal, Canada, June 2015, pp. 640–648.

[25] P. Smolensky, *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, ser. Parallel Distributed Porcessing: Explorations in the Microsctructure of Cognition. MIT Press, 1986, ch. 6, pp. 194–281.

[26] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[27] H. Huang and T. Toyoizumi, "Advanced mean-field theory of restricted Boltzmann machine," *Physical Review E*, vol. 91, no. 5, p. 050101, 2015.

[28] T. Heskes, "Stable fixed points of loopy belief propagation are minima of the Bethe free energy," in *Advances in Neural Information Processing Systems*, vol. 15, 2002, pp. 359–366.

[29] Y. LeCun, L. Bottu, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.