Exploiting Structured Sparsity in Bayesian Experimental Design

Philip Schniter
Dept. ECE, The Ohio State University, Columbus OH 43210, schniter@ece.osu.edu

Abstract—**In this paper, we merge Bayesian experimental design with turbo approximate message passing (AMP) algorithms for the purpose of recovering structured-sparse signals using a multi-step adaptive compressive-measurement procedure. First, we show that, when the signal posterior is Gaussian, a waterfilling approach can be used to adapt the measurement matrix in a way that expected information gain is maximized. Next, we propose four methods of approximating AMP's non-Gaussian marginal posteriors by a Gaussian joint posterior. One of these methods requires only point estimates of the signal, and leads to a novel kernel adaptation scheme that works even with non-Bayesian signal recovery algorithms like LASSO. Finally, we demonstrate (empirically) that our adaptive turbo AMP yields estimation performance very close to the support-oracle bound.**[1]

I. Introduction

Many signals in nature are known to yield a sparse representation in an appropriate basis. Here we mean that, for all signals $\boldsymbol{u} \in \mathbb{R}^N$ in a given class $\mathcal{U}$, there exists a unitary $\boldsymbol{\Psi} \in \mathbb{R}^{N \times N}$ that gives $\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^N$ containing only a few (say K, where $K \ll N$) entries of significant magnitude. In such cases, it is known that $M = \mathcal{O}(K \log(N/K))$ random projections $\boldsymbol{y}$, collected using $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ via

$$\boldsymbol{y} \;=\; \boldsymbol{\Phi}\boldsymbol{u} + \boldsymbol{w} \;=\; \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

suffice (with high probability) for accurate recovery of $\boldsymbol{x}$ (and thus $\boldsymbol{u}$), even in the presence of noise $\boldsymbol{w} \in \mathbb{R}^M$. For example, given such $\boldsymbol{y}$, computationally efficient algorithms like LASSO [1] are capable of producing estimates $\hat{\boldsymbol{x}}$ with error $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq C\|\boldsymbol{w}\|_2$, where C is a constant [2].

In many applications, the sparse representation $\boldsymbol{x}$ has structure beyond simple sparsity. For example, if $\boldsymbol{u}$ is a natural scene and $\boldsymbol{\Psi}$ implements a 2D discrete wavelet transform, then the coefficients $\boldsymbol{x}$ are not only sparse but also exhibit persistence across scales [3], which manifests as correlation within the sparsity pattern. Similarly, if $\boldsymbol{u}$ is an impulse response from a wideband wireless communications channel and $\boldsymbol{\Psi} = \boldsymbol{I}$, then $\boldsymbol{x}$ is sparse with clustered dominant entries [4]. The advantage of structured sparsity is that accurate signal recovery can be accomplished using fewer random linear measurements, such as $M = \mathcal{O}(K)$ [2].

When the measurement kernel $\boldsymbol{\Phi}$ must be chosen prior to sparse signal recovery, there are good reasons to choose $\boldsymbol{\Phi}$ with i.i.d random entries (see, e.g., [5]). We, however, are interested in *adapting* the measurement kernel $\boldsymbol{\Phi}$ during the signal recovery process [6]–[10] based on the current knowledge of $\boldsymbol{u}$. In particular, say that $\underline{\boldsymbol{y}}_{t-1} = \underline{\boldsymbol{\Phi}}_{t-1}\boldsymbol{u} + \underline{\boldsymbol{w}}_{t-1}$ are the *cumulative* measurements available before the t^{th} measurement step, and that $\boldsymbol{y}_t = \boldsymbol{\Phi}_t\boldsymbol{u} + \boldsymbol{w}_t \in \mathbb{R}^{M_t}$ are the M_t *new* measurements taken during the t^{th} measurement step, so that

$$\underbrace{\begin{bmatrix} \underline{\boldsymbol{y}}_{t-1} \\ \boldsymbol{y}_t \end{bmatrix}}_{\underline{\boldsymbol{y}}_t} = \underbrace{\begin{bmatrix} \underline{\boldsymbol{\Phi}}_{t-1} \\ \boldsymbol{\Phi}_t \end{bmatrix}}_{\underline{\boldsymbol{\Phi}}_t} \boldsymbol{u} + \underbrace{\begin{bmatrix} \underline{\boldsymbol{w}}_{t-1} \\ \boldsymbol{w}_t \end{bmatrix}}_{\underline{\boldsymbol{w}}_t}. \tag{2}$$

The knowledge of $\boldsymbol{u}$ gained from $\underline{\boldsymbol{y}}_{t-1}$ can be used to design $\boldsymbol{\Phi}_t$ in an effort to make $\boldsymbol{y}_t$ most informative about $\boldsymbol{u}$.

In this paper, we first briefly review the *Bayesian experimental design* approach to measurement kernel adaptation that is well known for $M_t = 1$ new measurements per step, and show how it can be generalized to $M_t \geq 1$ using a waterfilling approach. We then propose a method to exploit structured sparsity in measurement kernel adaptation that leverages our prior work on "turbo" approximate message passing (AMP) algorithms. A key step in our approach involves joint Gaussian posterior approximation, for which we propose several options. One of these options requires only point estimates of the signal, and leads to a novel kernel adaptation scheme that works with generic non-Bayesian signal recovery algorithms like LASSO. Finally, we demonstrate the efficacy of our approaches using experiments with clustered-sparse signals.

II. Bayesian Experimental Design

We now review core principles from Bayesian experimental design [6]–[9]. In the sequel, we use $t = 0, 1, 2, \ldots, T-1$ to index the measurement step, and we assume the following procedure. At $t = 0$, M_0 initial (non-adaptive) measurements $\boldsymbol{y}_0 = \boldsymbol{\Phi}_0\boldsymbol{u} + \boldsymbol{w}_0$ are collected using the kernel $\boldsymbol{\Phi}_0 \in \mathbb{R}^{M_0 \times N}$. At each subsequent step $t > 0$, M_t new measurements $\boldsymbol{y}_t = \boldsymbol{\Phi}_t\boldsymbol{u} + \boldsymbol{w}_t$ are collected using a kernel $\boldsymbol{\Phi}_t \in \mathbb{R}^{M_t \times N}$ that is designed with the goal of maximizing the *expected information gain*, as defined in the sequel. Just after step t, a total of $\underline{M}_t \triangleq \sum_{\tau=0}^{t} M_\tau$ measurements have been taken. Upon termination, the total number of measurements will be $M \triangleq \underline{M}_{T-1}$.

A. Expected Information Gain

Suppose that, at the start of the t^{th} measurement step, the signal posterior is given by $p(\boldsymbol{u} \,|\, \underline{\boldsymbol{y}}_{t-1})$. After $\boldsymbol{\Phi}_t$ is chosen and the corresponding new measurements $\boldsymbol{y}_t$ have been taken,

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

the posterior will be $p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}, \boldsymbol{y}_t)$. For brevity, we write

$$q(\boldsymbol{u}) \triangleq p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) \tag{3}$$

$$q(\boldsymbol{u}\,|\,\boldsymbol{y}_t) \triangleq p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}, \boldsymbol{y}_t) \tag{4}$$

noting that $\underline{\boldsymbol{y}}_{t-1}$ is fixed and known here. The *information gain* provided by the measurements $\boldsymbol{y}_t$ is defined as the Kullback-Leibler (KL) divergence between $q(\boldsymbol{u})$ and $q(\boldsymbol{u}\,|\,\boldsymbol{y}_t)$, i.e.,

$$D(\boldsymbol{y}_t) \triangleq \int_{\boldsymbol{u}} q(\boldsymbol{u}\,|\,\boldsymbol{y}_t) \log \frac{q(\boldsymbol{u}\,|\,\boldsymbol{y}_t)}{q(\boldsymbol{u})}. \tag{5}$$

Since $\boldsymbol{y}_t$ is not yet known when designing $\boldsymbol{\Phi}_t$, one typically designs $\boldsymbol{\Phi}_t$ to maximize the *expected* information gain (EIG)

$$\mathrm{E}\{D(\boldsymbol{y}_t)\,|\,\underline{\boldsymbol{y}}_{t-1}\} = \int_{\boldsymbol{y}_t} \underbrace{p(\boldsymbol{y}_t\,|\,\underline{\boldsymbol{y}}_{t-1})}_{\triangleq\, q(\boldsymbol{y}_t)} \int_{\boldsymbol{u}} q(\boldsymbol{u}\,|\,\boldsymbol{y}_t) \log \frac{q(\boldsymbol{u}\,|\,\boldsymbol{y}_t)}{q(\boldsymbol{u})}. \tag{6}$$

From (6), it is easy to see that the expected information gain equals the mutual information $I(\boldsymbol{U};\boldsymbol{Y}_t)$ between random vectors $\boldsymbol{U} \sim q(\boldsymbol{u})$ and $\boldsymbol{Y}_t \sim q(\boldsymbol{y}_t)$, i.e.,

$$I(\boldsymbol{U};\boldsymbol{Y}_t) = \int_{\boldsymbol{y}_t} \int_{\boldsymbol{u}} q(\boldsymbol{u}, \boldsymbol{y}_t) \log \frac{q(\boldsymbol{u}, \boldsymbol{y}_t)}{q(\boldsymbol{u}) q(\boldsymbol{y}_t)}. \tag{7}$$

Thus, when designing $\boldsymbol{\Phi}_t$, maximizing EIG is equivalent to maximizing the ($\underline{\boldsymbol{y}}_{t-1}$-conditioned) mutual information between the signal $\boldsymbol{u}$ and the new measurements $\boldsymbol{y}_t$.

Although evaluation of the EIG expression (6) is generally difficult, it is simple when the noise prior $p(\boldsymbol{w})$ and the signal posterior $q(\boldsymbol{u})$ are both Gaussian. To see this, we first write

$$\mathrm{E}\{D(\boldsymbol{y}_t)\,|\,\underline{\boldsymbol{y}}_{t-1}\} = H(\boldsymbol{Y}_t) - \int_{\boldsymbol{u}} q(\boldsymbol{u})\, H(\boldsymbol{Y}_t\,|\,\boldsymbol{U} = \boldsymbol{u}), \tag{8}$$

for differential entropies

$$H(\boldsymbol{Y}_t) \triangleq -\int_{\boldsymbol{y}_t} q(\boldsymbol{y}_t) \log q(\boldsymbol{y}_t) \tag{9}$$

$$H(\boldsymbol{Y}_t\,|\,\boldsymbol{U}{=}\boldsymbol{u}) \triangleq -\int_{\boldsymbol{y}_t} q(\boldsymbol{y}_t\,|\,\boldsymbol{u}) \log q(\boldsymbol{y}_t\,|\,\boldsymbol{u}). \tag{10}$$

Differential entropies are easy to evaluate for Gaussian distributions: $H\big(\mathcal{N}(\boldsymbol{a};\hat{\boldsymbol{a}},\boldsymbol{C})\big) = \frac{1}{2}\log|2\pi e \boldsymbol{C}|$. So, if we assume

$$p(\underline{\boldsymbol{w}}_{t-1}) = \mathcal{N}(\underline{\boldsymbol{w}}_{t-1}; \boldsymbol{0}, v_w \boldsymbol{I}) \tag{11}$$

$$q(\boldsymbol{u}) = p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), \tag{12}$$

then we have

$$q(\boldsymbol{y}_t\,|\,\boldsymbol{u}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\Phi}_t \boldsymbol{u}, v_w \boldsymbol{I}) \tag{13}$$

$$q(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\Phi}_t \boldsymbol{\mu}_u, \boldsymbol{\Phi}_t \boldsymbol{\Sigma}_u \boldsymbol{\Phi}_t^{\mathsf{T}} + v_w \boldsymbol{I}), \tag{14}$$

and (8) can be easily evaluated. In particular,

$$\begin{aligned} &\mathrm{E}\{D(\boldsymbol{y}_t)\,|\,\underline{\boldsymbol{y}}_{t-1}\} \\ &= \tfrac{1}{2}\log\big|2\pi e(\boldsymbol{\Phi}_t \boldsymbol{\Sigma}_u \boldsymbol{\Phi}_t^{\mathsf{T}} + v_w \boldsymbol{I})\big| - \int_{\boldsymbol{u}} q(\boldsymbol{u})\, \tfrac{1}{2}\log|2\pi e v_w \boldsymbol{I}| \\ &= \tfrac{1}{2}\log\big|\tfrac{1}{v_w}\boldsymbol{\Phi}_t \boldsymbol{\Sigma}_u \boldsymbol{\Phi}_t^{\mathsf{T}} + \boldsymbol{I}\big|. \end{aligned} \tag{15}$$

B. Maximizing the Expected Information Gain

We now seek the measurement kernel $\boldsymbol{\Phi}_t$ that maximizes the EIG subject to a sensing energy constraint $\|\boldsymbol{\Phi}_t\|_F^2 \le \mathcal{E}$.

In the often discussed case that $M_t = 1$ (e.g., [7], [8], [10]), $\boldsymbol{\Phi}_t$ is a row vector and so the EIG maximizing choice is simply $\sqrt{\mathcal{E}}$ times the dominant eigenvector of $\boldsymbol{\Sigma}_u$.

To our knowledge, the general case $M_t \ge 1$ has not been discussed in the literature. To tackle this case, we first write

$$\mathrm{E}\{D(\boldsymbol{y}_t)\,|\,\underline{\boldsymbol{y}}_{t-1}\} = \tfrac{1}{2}\log\big|\boldsymbol{\Phi}_t^{\mathsf{T}}\boldsymbol{\Phi}_t + v_w \boldsymbol{\Sigma}_u^{-1}\big| + \tfrac{1}{2}\log\big|\tfrac{1}{v_w}\boldsymbol{\Sigma}_u\big|, \tag{16}$$

and seek the positive semidefinite (p.s.d.) $\boldsymbol{\Phi}_t^{\mathsf{T}}\boldsymbol{\Phi}_t \in \mathbb{R}^{N\times N}$ with rank $\le M_t$ that maximizes the first term in (16). Using the eigenvalue decomposition $v_w \boldsymbol{\Sigma}_u^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathsf{T}}$ and the definition $\boldsymbol{B} \triangleq \boldsymbol{V}^{\mathsf{T}}\boldsymbol{\Phi}_t^{\mathsf{T}}\boldsymbol{\Phi}_t\boldsymbol{V}$, we can write $|\boldsymbol{\Phi}_t^{\mathsf{T}}\boldsymbol{\Phi}_t + v_w\boldsymbol{\Sigma}_u^{-1}| = |\boldsymbol{B}+\boldsymbol{\Lambda}|$ and thus translate the design problem to

$$\max_{\text{symmetric p.s.d. } \boldsymbol{B}} |\boldsymbol{B}+\boldsymbol{\Lambda}| \text{ s.t. } \mathrm{tr}(\boldsymbol{B}) \le \mathcal{E} \text{ and } \mathrm{rank}(\boldsymbol{B}) \le M_t,$$

which has [11, p. 255] a “waterfilling” solution, i.e., $\boldsymbol{B}$ is the diagonal matrix whose nonzero elements B_{nn} satisfy

$$B_{nn} = \begin{cases} \max\{L - \lambda_n, 0\} & n = 1,\ldots,M_t \\ 0 & n = M_t+1,\ldots,N \end{cases} \tag{17}$$

for $L \in \mathbb{R}^+$ selected so that $\sum_{n=1}^{M_t} B_{nn} = \mathcal{E}$. In writing (17), we have assumed that the eigenvalues are ordered such that $\lambda_n \le \lambda_{n+1}\ \forall n$. Also, we note that the solution (17) is not unique under repeated eigenvalues. Translating back to the measurement-kernel domain, we conclude that the n^{th} row of the EIG-maximizing $\boldsymbol{\Phi}_t$ equals $\sqrt{B_{nn}}$ times the n^{th} column of the eigenvector matrix $\boldsymbol{V}$.

C. Gaussian Posterior Approximation

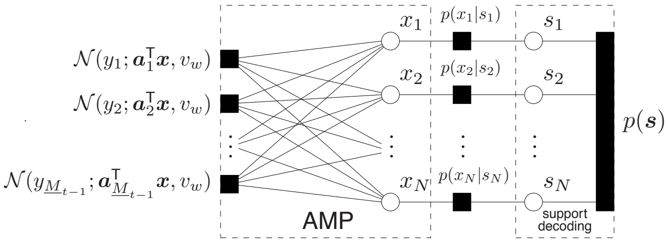
In Section II-A, it was observed that evaluation of the EIG is straightforward when the signal posterior $p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1})$ is Gaussian. However, for the sparsely represented signals $\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{x}$ that we target, the prior $p(\boldsymbol{u})$ is decidedly non-Gaussian, implying the same for the posterior $p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) \propto \mathcal{N}(\underline{\boldsymbol{y}}_{t-1}; \underline{\boldsymbol{\Phi}}_{t-1}\boldsymbol{u}, v_w\boldsymbol{I})p(\boldsymbol{u})$. Thus, it is common practice to *approximate* $p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1})$ as Gaussian.

Various examples of Gaussian posterior approximations can be found in the literature. For example, in the “Bayesian compressive sensing” (BCS) approach [8], the sparse representation $\boldsymbol{x}$ is assumed to have the *conditionally* Gaussian prior $p(\boldsymbol{x}\,|\,\boldsymbol{\alpha}) = \prod_{n=1}^N \mathcal{N}(x_n; 0, \alpha_n^{-1})$ with the precisions $\{\alpha_n\}$ following a Gamma pdf, thus yielding an unconditional prior that is Student's-t. However, by plugging in an estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ (obtained, e.g., via the EM algorithm), one obtains the Gaussian *prior approximation* $\hat{p}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \mathrm{Diag}(\hat{\boldsymbol{\alpha}})^{-1})$. Then, assuming $\underline{\boldsymbol{w}}_{t-1} \sim \mathcal{N}(\boldsymbol{0}, v_w\boldsymbol{I})$ in the observation model $\underline{\boldsymbol{y}}_{t-1} = \underline{\boldsymbol{\Phi}}_{t-1}\boldsymbol{\Psi}\boldsymbol{x} + \underline{\boldsymbol{w}}_{t-1}$, and abbreviating $\underline{\boldsymbol{A}}_{t-1} \triangleq \underline{\boldsymbol{\Phi}}_{t-1}\boldsymbol{\Psi}$, we directly obtain [8]

$$\hat{p}(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-1}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \tag{18}$$

$$\boldsymbol{\Sigma}_x \triangleq \big(\tfrac{1}{v_w}\underline{\boldsymbol{A}}_{t-1}^{\mathsf{T}}\underline{\boldsymbol{A}}_{t-1} + \mathrm{Diag}(\hat{\boldsymbol{\alpha}})\big)^{-1} \tag{19}$$

$$\boldsymbol{\mu}_x \triangleq \tfrac{1}{v_w}\boldsymbol{\Sigma}_x\underline{\boldsymbol{A}}_{t-1}^{\mathsf{T}}\underline{\boldsymbol{y}}_{t-1}, \tag{20}$$



$\mathcal{N}(y_1; \boldsymbol{a}_1^\mathsf{T}\boldsymbol{x}, v_w)$
$\mathcal{N}(y_2; \boldsymbol{a}_2^\mathsf{T}\boldsymbol{x}, v_w)$
$\mathcal{N}(y_{\underline{M}_{t-1}}; \boldsymbol{a}_{\underline{M}_{t-1}}^\mathsf{T}\boldsymbol{x}, v_w)$
x_1
x_2
x_N
$p(x_1|s_1)$
$p(x_2|s_2)$
$p(x_N|s_N)$
s_1
s_2
s_N
$p(\boldsymbol{s})$
AMP
support decoding

Fig. 1. Factor graph just before measurement step t.

and thus the Gaussian posterior approximation $\hat{p}(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$ for $\boldsymbol{\mu}_u = \boldsymbol{\Psi}\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_u = \boldsymbol{\Psi}\boldsymbol{\Sigma}_x\boldsymbol{\Psi}^\mathsf{T}$.

Examples of other approaches include [7], which assumed a Laplace prior on $\boldsymbol{x}$ and applied expectation propagation to obtain a Gaussian posterior approximation, and [9], which used variational techniques.

III. Design under Structured Sparsity

A. Structured-Sparse Signal Model

We now propose a method to exploit structured sparsity in Bayesian experimental design. First, we assume a decoupled conditionally Gaussian prior on structured-sparse $\boldsymbol{x}$:

$$p(x_n\,|\,s_n) \;=\; s_n\mathcal{N}(x_n; \mu_x, v_x) + (1-s_n)\delta(x_n), \quad (21)$$

where $s_n \in \{0,1\}$ is a binary indicator, (μ_x, v_x) are the prior mean and variance of non-zero coefficients, and $\delta(\cdot)$ denotes the Dirac delta. In conjunction, we assume a joint prior pmf $p(\boldsymbol{s})$ on the support pattern $\boldsymbol{s} \triangleq [s_1, \ldots, s_N]^\mathsf{T}$. While we place no restrictions on $p(\boldsymbol{s})$, we note that Markov chain/field/tree priors typically lead to efficient algorithms [12]. The resulting factor graph is shown in Fig. 1 (where $\boldsymbol{a}_m^\mathsf{T}$ is used to denote the m^{th} row of $\underline{\boldsymbol{A}}_{t-1}$).

B. Marginal Inference via Turbo-AMP

Although exact evaluation of the joint posterior $p(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-1})$ is computationally impractical, the marginal posteriors $\{p(x_n\,|\,\underline{\boldsymbol{y}}_{t-1})\}_{n=1}^N$ can be closely approximated using loopy belief propagation. For this, we propose to use the following "turbo" inference procedure. To start, the generalized[2] [14] *approximate message passing* (AMP) algorithm [13] is used to infer $\boldsymbol{x}$ and $\boldsymbol{s}$ using the prior marginals $\{p(s_n)\}_{n=1}^N$. Then, treating the marginal likelihoods $\{p(\underline{\boldsymbol{y}}_{t-1}\,|\,s_n)\}_{n=1}^N$ returned by AMP as a (refined) prior on $\boldsymbol{s}$, we infer the support $\boldsymbol{s}$ using an appropriate soft decoding algorithm (e.g., Markov chain/tree/field inference [12]). Next, treating the marginal likelihoods on $\boldsymbol{s}$ returned by the decoder as a (further refined) prior on $\boldsymbol{s}$, we re-infer $\boldsymbol{x}$ and $\boldsymbol{s}$ using AMP. These alternations between AMP and support decoding are repeated until the likelihoods on $\boldsymbol{s}$ converge. Details on this *turbo-AMP* procedure can be found in [15].

C. Gaussian Posterior Approximation

For the purpose of Bayesian experimental design, the posteriors returned by AMP are problematic in two ways: they are i) marginal and ii) non-Gaussian. To remedy the situation, we propose to take an approach reminiscent of [8]: we convert AMP's marginal posteriors into a Gaussian *prior approximation* $\hat{p}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{m}}, \mathrm{Diag}(\hat{\boldsymbol{\alpha}})^{-1})$, which then leads directly to

$$\hat{p}(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-1}) \;=\; \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \quad (22)$$

$$\boldsymbol{\Sigma}_x \;\triangleq\; \big(\tfrac{1}{v_w}\underline{\boldsymbol{A}}_{t-1}^\mathsf{T}\underline{\boldsymbol{A}}_{t-1} + \mathrm{Diag}(\hat{\boldsymbol{\alpha}})\big)^{-1} \quad (23)$$

$$\boldsymbol{\mu}_x \;\triangleq\; \hat{\boldsymbol{m}} + \tfrac{1}{v_w}\boldsymbol{\Sigma}_x\underline{\boldsymbol{A}}_{t-1}^\mathsf{T}(\underline{\boldsymbol{y}}_{t-1} - \underline{\boldsymbol{A}}_{t-1}\hat{\boldsymbol{m}}), \quad (24)$$

and thus the Gaussian posterior approximation $\hat{p}(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$ for $\boldsymbol{\mu}_u = \boldsymbol{\Psi}\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_u = \boldsymbol{\Psi}\boldsymbol{\Sigma}_x\boldsymbol{\Psi}^\mathsf{T}$.

In deciding how AMP's marginal posteriors are mapped to the prior approximation $\mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{m}}, \mathrm{Diag}(\hat{\boldsymbol{\alpha}})^{-1})$, we recall from (15) that only the precisions $\hat{\boldsymbol{\alpha}} = [\alpha_1, \ldots, \alpha_N]^\mathsf{T}$ will affect the EIG; the means $\hat{\boldsymbol{m}}$ are inconsequential. Thus, using $(\hat{x}_n, \nu_n)$ to denote the posterior mean and variance on x_n returned by AMP, and using $\hat{s}_n$ to denote the posterior mean of s_n returned by AMP, we suggest several assignments for $\hat{\boldsymbol{\alpha}}$:

1) Var: $\hat{\alpha}_n^{-1} = \nu_n$.
2) Mean: $\hat{\alpha}_n^{-1} = |\hat{x}_n|^2$
3) Energy: $\hat{\alpha}_n^{-1} = |\hat{x}_n|^2 + \nu_n$
4) Support: $\hat{\alpha}_n^{-1} = \hat{s}_n\big(|\mu_x|^2 + v_x\big)$,

recalling that (μ_x, v_x) denotes the prior mean and variance on $\{x_n\}_{n=1}^N$ specified by the signal model (21).

We now make a few comments. First, we note that our Support-based approximation is perhaps closest in spirit to the BCS approach [8] detailed in Section II-C, since the *conditionally* Gaussian prior $p(x_n\,|\,s_n)$ is converted to the Gaussian prior approximation $\hat{p}(x_n)$ by simply "plugging in" an estimate of s_n. Second, we note that our Mean-based approximation requires only the point estimates $\{\hat{x}_n\}$ and thus facilitates kernel adaptation with non-Bayesian sparse recovery algorithms like LASSO. Third, we note from (23) that the most expensive step in kernel adaptation is the computation of the M_t smallest eigenvectors/values of the matrix $\underline{\boldsymbol{A}}_{t-1}^\mathsf{T}\underline{\boldsymbol{A}}_{t-1} + v_w\,\mathrm{Diag}(\hat{\boldsymbol{\alpha}})$, a task that is efficiently tackled by the Lanczos algorithm (e.g., "`eigs`" in Matlab).

IV. Numerical Experiments

We now report the results of numerical experiments where clustered-sparse signals were recovered under measurement kernel adaptation. The clustered-sparse signals $\boldsymbol{x}$ were generated according to the Bernoulli-Gaussian model $p(x_n|s_n) = s_n\mathcal{N}(x_n; 0, 1) + (1-s_n)\delta(x_n)$ with sparsity pattern $\{s_n\}$ generated by a Markov chain (MC) with transition probabilities $p_{01} \triangleq \Pr\{s_n\!=\!0\,|\,s_{n-1}\!=\!1\}$ and $p_{10} \triangleq \Pr\{s_n\!=\!1\,|\,s_{n-1}\!=\!0\}$. In all cases, we used signals of length $N = 500$ and set p_{01} and p_{10} so that the activity rate $\pi_s \triangleq \Pr\{s_n = 1\} = (1 + p_{01}/p_{10})^{-1} = 0.1$ and the clustering parameter $\gamma_s \triangleq p_{10}/\pi_s = 0.1$. In this case, the expected cluster length equals $1/p_{10} = 11.1$. Throughout, we assumed the canonical

[2] Since the columns of $\underline{\boldsymbol{A}}_{t-1}$ are not guaranteed to be equal energy, as required by Donoho/Maleki/Montanari's AMP [13], we need the generalized AMP proposed by Rangan in [14].

sparsity basis (i.e., $\boldsymbol{\Psi}=\boldsymbol{I}$ so that $\boldsymbol{u}=\boldsymbol{x}$ and $\underline{\boldsymbol{A}}_t=\underline{\boldsymbol{\Phi}}_t$) and the sensing energy constraint $\|\boldsymbol{\Phi}_t\|_F^2=M_tN$. We used $T=5$ measurement steps with $\underline{M}_0=100$ initial measurements (constructed using i.i.d Gaussian $\underline{\boldsymbol{\Phi}}_0$) and $M_t=25$ new measurements per adaptation step, Finally, the noise variance v_w was chosen so that SNR $\triangleq \mathrm{E}\{\|\boldsymbol{a}_m^{\mathsf{T}}\boldsymbol{x}\|_2^2\}/v_w=15\text{dB}$, and our performance metric is NMSE $\triangleq \mathrm{E}\{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2^2/\|\boldsymbol{x}\|_2^2\}$ averaged over 500 problem realizations.

In Fig. 2, we show average NMSE versus number of measurements $\underline{M}_t$, for both AMP (without support decoding) and turbo-AMP, under various adaptation schemes. For adaptation, we tested the four Gaussian approximation approaches to EIG-maximization proposed in Section III-C. As baselines, we also show the performance of AMP and turbo-AMP under non-adaptive (i.i.d Gaussian) kernels. From the figure, we see that the NMSE reduction from kernel adaptation is 5dB, from clustered-sparsity is 4dB, and from the combination is 12dB: more than the sum of the parts. Among the four approximation schemes used for EIG maximization, we see nearly identical performance for all but the Support approximation, which degrades by 1dB in the clustered-sparse context.

In Fig. 3, we show average NMSE versus number of measurements $\underline{M}_t$ under various recovery schemes, including BCS[3] [8] and LASSO[3] [1]. The figure shows that, under non-adaptive measurements, BCS and AMP both perform 5dB better than LASSO, while turbo-AMP performs 8dB better than LASSO. When the measurements are adapted to maximize EIG under our Mean-based approximation, we see significant (4dB-7dB) improvements over the corresponding non-adaptive cases. Furthemore, for adaptive BCS, our Mean-based approximation performs 1dB better than the Gaussian approximation suggested in [8]. Finally, our adaptive turbo-AMP comes within 2dB of the support-genie bound, where the support $\boldsymbol{s}$ of $\boldsymbol{x}$ is assumed to be known. In the latter case, the signal prior/posterior are Gaussian and thus EIG maximization can be implemented without approximation.

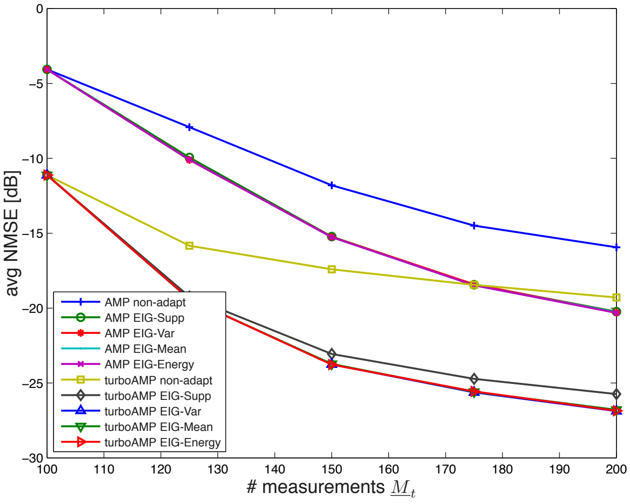


Fig. 2. Average NMSE versus # of measurements $\underline{M}_t$ for AMP and turbo-AMP under various kernel-adaptation schemes.

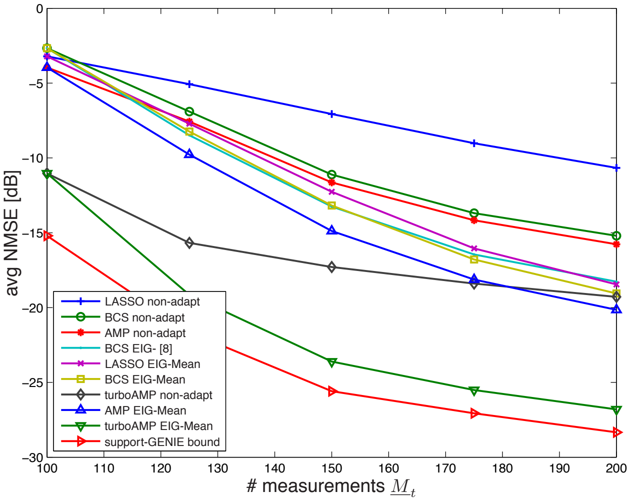


Fig. 3. Average NMSE versus # of measurements $\underline{M}_t$ for various recovery algorithms and kernel-adaptation schemes.

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] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, pp. 1982–2001, Apr. 2010.

[3] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 3rd ed., 2008.

[4] A. F. Molisch, "Ultrawideband propagation channels—Theory, measurement, and modeling," *IEEE Trans. Veh. Tech.*, vol. 54, pp. 1528–1545, Sep. 2005.

[5] E. J. Candès and M. A. Davenport, "How well can we estimate a sparse vector?," *arXiv:1104.5246*, June 2011.

[6] M. DeGroot, "Uncertainty, information, and sequential experiments," *Annals Math. Stats.*, vol. 33, no. 2, pp. 404–419, 1962.

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

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

[9] M. W. Seeger and H. Nickisch, "Large scale Bayesian inference and experimental design for sparse linear models," *SIAM J. Imaging Sciences*, vol. 4, no. 1, pp. 166–199, 2011.

[10] J. Haupt and R. Nowak, "Adaptive sensing for sparse recovery," in *Compressed Sensing: Theory and Applications* (Y. Eldar and G. Kutyniok, eds.), Cambridge Univ. Press, 2011.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[12] C. A. Bouman, "Markov random fields and stochastic image models," in *IEEE Int. Conf. Image Processing Tutorial*, Oct. 1995.

[13] 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.

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

[15] P. Schniter, "Turbo reconstruction of structured sparse signals," in *Proc. Conf. Inform. Science & Syst.*, (Princeton, NJ), Mar. 2010.

[3] To implement BCS, we used `BCS_fast_rvm.m` in `bcs_ver0.1.zip` from `http://people.ee.duke.edu/∼lcarin/BCS.html`, and to implement LASSO, we used `spg_bpdn.m` from `www.cs.ubc.ca/labs/scl/spgl1/`. For both algorithms, we performed, for each realization, a grid search over the noise variance parameter, and used the one that minimized NMSE. Since this tuning method is "genie aided," the BCS and LASSO performance that we report is better than what could be obtained in practice.