Signal Recovery in Compressed Sensing via Universal Priors

Dror Baron, *Senior Member, IEEE,* and Marco F. Duarte, *Member, IEEE*

Abstract

We study the compressed sensing (CS) signal estimation problem where an input is measured via a linear matrix multiplication under additive noise. While this setup usually assumes sparsity or compressibility in the observed signal during recovery, the signal structure that can be leveraged is often not known *a priori*. In this paper, we consider *universal* CS recovery, where the statistics of a stationary ergodic signal source are estimated simultaneously with the signal itself. We focus on a maximum *a posteriori* (MAP) estimation framework that leverages universal priors such as Kolmogorov complexity and minimum description length. We provide theoretical results that support the algorithmic feasibility of universal MAP estimation through a Markov Chain Monte Carlo implementation. We also include simulation results that showcase the promise of universality in CS, particularly for low-complexity sources that do not exhibit standard sparsity or compressibility.

I. Introduction

Since many systems in science and engineering are approximately linear, linear inverse problems have attracted great attention in the signal processing community. A signal $x \in \mathbb{R}^N$ is recorded via a linear operator under additive noise:

$$y = \Phi x + z, \tag{1}$$

where Φ is an $M \times N$ matrix and $z \in \mathbb{R}^N$ denotes the noise. The goal is to estimate x from the measurements y given knowledge of Φ and a model for the noise z. When $M \ll N$, the setup is known as compressed sensing (CS) and the estimation problem is commonly referred to as recovery or reconstruction; by posing a sparsity or compressibility requirement on the signal and using this requirement as a prior during recovery, it is indeed possible to accurately estimate x from y [2, 3].

An early version of this work appeared at the 49th Allerton Conference on Communications, Control, and Computing, Monticello, IL, September 2011 [1].

M. F. Duarte was partially supported by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch.

D. Baron is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695. E-mail: dzbaron@ncsu.edu

M. F. Duarte is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003. E-mail: mduarte@ecs.umass.edu

While in CS the acquisition can be designed independently of the particular signal prior through the use of randomized measurement matrices Φ, the majority of existing recovery algorithms require knowledge of the sparsity structure of x, i.e., the choice of transformation that renders a sparse coefficient vector for the signal. A second, separate class of Bayesian CS recovery algorithms poses a probabilistic prior for the coefficients of x in a known transform domain [4–6]. In contrast, complexity-based regularization methods can use arbitrary prior information on the signal model and come with analytical guarantees, but are only computationally efficient for specific signal models, such as the independent-entry Laplacian model [7]. As a fourth alternative, there exist algorithms that can formulate dictionaries that yield sparse representations for the signals of interest when a large amount of training data is available [8–10].

In certain cases, one might not be certain about the structure or statistics of the source prior to recovery. It would nonetheless be desirable to formulate algorithms to estimate x that are agnostic to the particular statistics of the signal. Therefore, we shift our focus from the standard sparsity or compressibility priors to *universal* priors [11, 12]. Such concepts have been previously leveraged in the Kolmogorov sampler universal denoising algorithm [13], which minimizes Kolmogorov complexity [14–17].[1] Related approaches based on minimum description length (MDL) [20–22] minimize the complexity of the estimated signal with respect to some class of parametric sources.

Unfortunately, while MDL may provide a suitable algorithmic recovery framework for parametric sources, alternative approaches for non-parametric sources based on Kolmogorov complexity are not computable in practice. To address this computational problem, we confine our attention to stationary ergodic sources and develop an algorithmic framework for universal signal estimation in CS systems. Our framework leverages the fact that for stationary ergodic sources, both the per-symbol empirical entropy and Kolmogorov complexity converge asymptotically almost surely to the entropy rate of the source [23]. We aim to minimize the empirical entropy; our minimization is regularized by introducing a log likelihood for the noise model, which is equivalent to the standard least squares under additive white Gaussian noise. Other noise distributions are readily supported.

We make several contributions toward our universal CS framework. First, we show that for well-behaved sources the maximum *a posteriori* (MAP) risk over a specific quantization grid converges asymptotically in the signal length to the MAP risk of the non-quantized estimator. Second, we apply this quantization equivalence result to a MAP estimator driven by a universal prior, providing a finite-computation universal estimation scheme. Third, we propose a recovery algorithm based on Markov chain Monte Carlo (MCMC) to approximate this estimation procedure. Fourth, we prove that for a sufficiently large number of iterations the output of our MCMC recovery algorithm converges to the correct MAP estimate. Fifth, we identify computational bottlenecks in the implementation of our MCMC estimator and show approaches to reduce their complexity. Sixth, we develop an adaptive quantization scheme that tailors a set of representation levels to minimize the quantization error within the MCMC iterations and that provides an accelerated implementation. Finally, we showcase encouraging experimental results that show

[1] A recent paper by Jalali and Maleki [18], developed independently from and appearing simultaneously with our work [1, 19], considered the performance of Kolmogorov complexity minimization for CS recovery from measurements corrupted by noise of bounded magnitude.

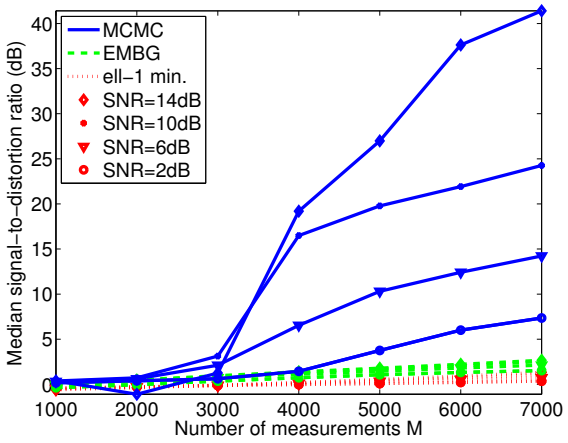

Fig. 1. *Universal MCMC, EMBG [24], and ℓ_1-norm minimization recovery results for the four-state Markov switching source of length $N = 10000$ as a function of the number of Gaussian random measurements M for different SNR values. For $M \geq 4000$, MCMC significantly outperforms ℓ_1-norm minimization and EMBG, which fail due to the signal not being sparse in a fixed basis. Each point in the graph represents median performance over 25 signal and random measurement matrix draws.*

recovery performance for a variety of types of signal structures (or statistics) that meets or exceeds that of popular state-of-the-art algorithms.

To showcase the potential of our universal estimation approach, Fig. 1 illustrates recovery results from Gaussian measurement matrices for a four-state Markov source of length $N = 10000$ that generates the pattern $+1, +1, -1, -1, +1, +1, -1, -1 \ldots$ with 3% errors in state transitions, resulting in the signal switching from -1 to +1 or vice-versa either too early or too late. While it is well known that sparsity-promoting recovery algorithms such as ℓ_1-norm minimization can recover sparse sources from linear measurements, the aforementioned switching source is not sparse in a foreknown basis, rendering such algorithms not applicable. In contrast, our MCMC recovery algorithm estimates this source with high fidelity when the signal to noise ratio (SNR) is sufficiently large and a moderate number of measurements M is available. Our experimental results in Section VII also show some challenges faced by MCMC recovery of certain classes of sparse signals; we identify properties of the algorithm that may cause these challenges, which remain to be addressed. While the practical value of our MCMC algorithm may be reduced due to its high computational cost, our approach provides a starting point toward further performance gains of more practical algorithms for computing the universal MAP estimator. Furthermore, our experiments will show that the performance of MCMC is comparable to and sometimes significantly better than existing state-of-the-art algorithms.

This paper is organized as follows. Section II provides background content. Section III overviews MAP estimation and quantization, and Section IV introduces universal MAP estimation. Section V formulates a concrete MCMC algorithm for universal MAP estimation, Section VI describes our proposed adaptive quantization scheme for MCMC, and Section VII presents initial experimental results. We conclude in Section VIII. The proofs of our main theoretical results appear in the appendix.

II. Background and related work

A. Compressed Sensing

Consider the noisy measurement setup via a linear operator (1). The input vector $x \in \mathbb{R}^N$ is generated by a stationary and ergodic source X, and must be estimated from y and Φ. *The distribution f_X that generates X is unknown.* The matrix $\Phi \in \mathbb{R}^{M \times N}$ has independent and identically distributed (i.i.d.) Gaussian entries, $\Phi(m,n) \sim \mathcal{N}(0, \frac{1}{M})$.[2] These moments ensure that columns of the matrix have unit norm on average. For concrete analysis, we assume the noise $z \in \mathbb{R}^M$ to be i.i.d. Gaussian, with mean zero and known variance σ_Z^2 for simplicity. Other noise distributions are readily supported.

We focus on the setting where $M, N \to \infty$ and the aspect ratio is positive,

$$\delta \triangleq \lim_{N \to \infty} \frac{M}{N} > 0. \tag{2}$$

Similar settings have been discussed in the literature, e.g., [25, 26]. Since x was generated by an unknown source, we must search for an estimation mechanism that is agnostic to the specific distribution f_X.

B. Quantization

Following the approach of Baron and Weissman [27], we define the set of data-independent reproduction levels for quantizing x as

$$\mathcal{R} \triangleq \left\{ \ldots, -\frac{1}{\gamma}, 0, \frac{1}{\gamma}, \ldots \right\}, \tag{3}$$

where $\gamma = \lceil \ln(N) \rceil$. As N increases, $\mathcal{R}$ will quantize x to a greater resolution. In Section III, we will show that under suitable conditions on f_X, performing maximum *a posteriori* (MAP) estimation over the discrete alphabet $\mathcal{R}$ asymptotically converges to the MAP estimate over the continuous distribution f_X. This reduces the complexity of the estimation problem from continuous to combinatorial.

C. Related work

For a scalar channel, e.g., $\Phi = I$ and $y = x + z$, Donoho proposed the the Kolmogorov sampler (KS) for denoising [13],

$$x_{KS} \triangleq \arg\min_w K(w) \ s.t. \ \|w - y\|^2 < \tau, \tag{4}$$

where $K(x)$ denotes the Kolmogorov complexity of x, defined as the length of the shortest input to a Turing machine [28] that generates the output x and then halts, and $\tau = N\sigma_Z^2$ controls for the presence of noise. It can be shown that $K(x)$ asymptotically captures the statistics of the stationary ergodic source X, and the per-symbol complexity achieves the entropy rate $H \triangleq H(X)$, i.e., $\lim_{N \to \infty} \frac{1}{N} K(x) = H$ almost surely. Noting that universal

[2] In contrast to our analytical and numerical results, the algorithm presented in Section V is not dependent on a particular choice for the matrix Φ.

lossless compression algorithms [11, 12] achieve the entropy rate for any discrete-valued finite state machine source X, we see that these algorithms achieve the per-symbol Kolmogorov complexity almost surely.

Donoho et al. expanded the KS to the linear CS measurement setting $y = \Phi x$ but did not consider measurement noise [29]. A recent paper by Jalali and Maleki [18], which appeared simultaneously with our work [1, 19], provides an analysis of a modified KS suitable for measurements corrupted by noise of bounded magnitude. Inspired by [29], we estimate x from noisy measurements y using the empirical entropy as a proxy for the Kolmogorov complexity.

Separate notions of complexity-based regularization have also been shown to be well suited for denoising and CS recovery [7, 30–32]. For example, minimum description length (MDL) [20–22, 32] provides a universal framework composed of classes of parametric signal models for which the signal complexity can be defined sharply. In principle, complexity-based regularization approaches can yield MDL-flavored CS recovery algorithms that are universal for parametric classes of sources [7, 30, 31]. An alternative universal denoising approach computes the universal conditional expectation of the signal [19]; we leave this for future work.

III. MAP ESTIMATION AND DISCRETIZATION

In this section, we assume for exposition purposes that we know the input statistics f_X. Given the measurements y, the MAP estimator for x has the form

$$x_{MAP} \triangleq \arg\max_w f_X(w) f_{Y|X}(y|w). \tag{5}$$

Because z is i.i.d. Gaussian with mean zero and known variance σ_Z^2,

$$f_{Y|X}(y|w) = c_1 e^{-c_2 \|y - \Phi w\|^2}, \tag{6}$$

where $c_1 = (2\pi\sigma_Z^2)^{-M/2}$ and $c_2 = \frac{1}{2\sigma_Z^2}$ are constants and $\|\cdot\|$ denotes the Euclidean norm. Plugging into (5) and taking log likelihoods, we obtain

$$x_{MAP} = \arg\min_w \Psi^X(w),$$

where $\Psi^X(\cdot)$ denotes the objective function (risk)

$$\Psi^X(w) \triangleq -\ln(f_X(w)) + c_2 \|y - \Phi w\|^2; \tag{7}$$

our ideal risk would be $\Psi^X(x_{MAP})$.

Instead of performing continuous-valued MAP estimation, we optimize for the MAP in the discretized domain $\mathcal{R}^N$. We begin with two technical conditions on the input.

Condition 1: We require that the probability density f_X has bounded support, i.e., there exists $\Lambda = [x_{\min}, x_{\max}]$ such that $f_X(x) = 0$ for $x \notin \Lambda^N$.

Condition 2: We require that the probability density has bounded derivatives

$$\left| \frac{d}{dx_n} \ln(f_X(x)) \right| < \rho \tag{8}$$

for $x \in \Lambda^N$, where $\frac{d}{dx_n}$ is the derivative with respect to entry n of x, $n \in \{1, \ldots, N\}$, and $\rho > 0$ is a constant.

A limitation of the data-independent reproduction levels (3) is that $\mathcal{R}$ has infinite cardinality. Thanks to Condition 1, for each value of γ there exists a constant $c_3 > 0$ such that a finite set of reproduction levels

$$\mathcal{R}_F \triangleq \left\{ -\frac{c_3\gamma^2}{\gamma}, -\frac{c_3\gamma^2 - 1}{\gamma}, \ldots, \frac{c_3\gamma^2}{\gamma} \right\} \tag{9}$$

will quantize the range of values Λ to the desired accuracy. This finite quantization step reduces the complexity of the estimation problem from infinite to combinatorial. In fact, if we fix $c_3 = 1$, then the range $\mathcal{R}_F$ covers all symbols x_i with overwhelming probability for sufficiently large N. The resulting reduction in complexity is due to the structure in $\mathcal{R}_F$ and independent of the particular statistics of the source X.

Let $\widetilde{x}_{MAP}$ be the quantization bin in $(\mathcal{R}_F)^N$ nearest to x_{MAP}. Condition 2 ensures that a small perturbation from x_{MAP} to $\widetilde{x}_{MAP}$ does not change $\ln(f_X(\cdot))$ by much. We use this fact to prove that $\Psi^X(\widetilde{x}_{MAP})$ is sufficiently close to $\Psi^X(x_{MAP})$ asymptotically.

Theorem 1: Let $\Phi \in \mathbb{R}^{M\times N}$ be an i.i.d. Gaussian measurement matrix where each entry has mean zero and variance $\frac{1}{M}$. Suppose that Conditions 1 and 2 hold, the aspect ratio $\delta > 0$ in (2), and the noise $z \in \mathbb{R}^M$ is i.i.d. zero-mean Gaussian with finite variance σ_Z^2. Then for all $\epsilon > 0$, the quantized estimator $\widetilde{x}_{MAP}$ satisfies

$$\Psi^X(x_{MAP}) \leq \Psi^X(\widetilde{x}_{MAP}) < \Psi^X(x_{MAP}) + N\epsilon$$

almost surely as $N \to \infty$.

Theorem 1 is proved in Appendix A; it shows that in terms of the MAP objective function, $\widetilde{x}_{MAP}$ is near-optimal almost surely asymptotically. Thus, it is natural to perform the MAP optimization directly in the quantized domain:

$$x_{MAP}(\mathcal{R}_F) \triangleq \arg \min_{w\in(\mathcal{R}_F)^N} \Psi^X(w). \tag{10}$$

From Theorem 1, we have

$$\Psi^X(x_{MAP}) \leq \Psi^X(x_{MAP}(\mathcal{R}_F)) \leq \Psi^X(\widetilde{x}_{MAP}) \leq \Psi^X(x_{MAP}) + N\epsilon \tag{11}$$

almost surely asymptotically for any $\epsilon > 0$.

Discrete probability space: Now that we have set up a quantization grid $(\mathcal{R}_F)^N$ for x, we convert the distribution f_X to a probability mass function (PMF) p_X over $(\mathcal{R}_F)^N$. Let

$$f_{\mathcal{R}_F} \triangleq \sum_{w\in(\mathcal{R}_F)^N} f_X(w),$$

and define the PMF $p_X(\cdot)$ as

$$p_X(w) \triangleq \frac{f_X(w)}{f_{\mathcal{R}_F}}. \tag{12}$$

We now have

$$\begin{aligned} &\min_{w\in(\mathcal{R}_F)^N} \left(-\ln(p_X(w)) + c_2\|y - \Phi w\|^2\right) \\ = \quad &\Psi^X(x_{MAP}(\mathcal{R}_F)) + \ln(f_{\mathcal{R}_F}). \end{aligned} \tag{13}$$

The additive constant $\ln(f_{\mathcal{R}_F})$ can be ignored during the MAP optimization over $(\mathcal{R}_F)^N$, so that (10) gives the MAP estimate of x over $(\mathcal{R}_F)^N$ due to (13).

IV. UNIVERSAL MAP ESTIMATION

In [13], Donoho showed that for the scalar channel $y = x+z$: (i) the Kolmogorov sampler x_{KS} (4) is drawn from the posterior distribution $p_{X|Y}(x|y)$; and (ii) the mean square error (MSE) of this estimate $E_{X,Z,\Phi}[\|y - x_{KS}\|^2]$ is equal to twice the minimum mean squared error (MMSE).

Given that Theorem 1 shows that the risk penalty due to quantization vanishes asymptotically in N, we now describe a universal estimator for CS over a quantized grid. Consider a universal prior p_U [11, 12] that might involve Kolmogorov complexity [14–16], e.g., $p_U(w) = 2^{-K(w)}$, or MDL complexity with respect to some class of parametric sources [20–22]. The universal prior has the fortuitous property that for every stationary ergodic source X and fixed $\epsilon > 0$, there exists some minimal $N_0(X, \epsilon)$ such that

$$-\ln(p_U(w)) < -\ln(p_X(w)) + \epsilon N \tag{14}$$

for *all* $w \in (\mathcal{R}_F)^N$ and $N > N_0(X, \epsilon)$ [11, 12]. We optimize over an objective function (risk) that incorporates p_U and the presence of additive white Gaussian noise in the measurements:[3]

$$\Psi^U(w) \triangleq -\ln(p_U(w)) + c_2\|y - \Phi w\|^2, \tag{15}$$

resulting in

$$x^U_{MAP} \triangleq \arg \min_{w\in(\mathcal{R}_F)^N} \Psi^U(w). \tag{16}$$

Based on the results by Donoho for the scalar channel [13], we now present a conjecture on the quality of the reconstruction x^U_{MAP}; experimental evidence to assess this claim is presented in Section VII.

Conjecture 1: Assume that the conditions of Theorem 1 hold. Then for all $\epsilon > 0$, the mean squared error of the universal MAP estimator x^U_{MAP} satisfies

$$E_{X,Z,\Phi}\left[\|x - x^U_{MAP}\|^2\right] < 2E_{X,Z,\Phi}\left[\|x - E_X[x|y, \Phi]\|^2\right] + N\epsilon$$

for sufficiently large N.

We note in passing that when the SNR is low, the MMSE could be almost as large as the energy of the signal, $\|x\|_2^2$. In these problems, achieving twice the MMSE means achieving a mean squared error that could be greater than $\|x\|_2^2$, which is not a useful guarantee. In contrast, the mean square error achieved by $\widehat{x}_{MMSE}$ will be smaller than $\|x\|_2^2$, which is a *much* better signal estimate than $\widehat{x}_{MAP}$. This excellent performance at low SNR is particularly useful in finance (cf. [33] and references therein).

V. ALGORITHMIC APPROACH

Although the results of the previous section are theoretically appealing, a brute force optimization of x^U_{MAP} is computationally intractable. Instead, we propose an algorithmic approach based on Markov chain Monte Carlo (MCMC) methods [34]. Our approach is reminiscent of the framework by Weissman et al. and Yang et al. for lossy data compression [27, 35, 36].

[3]The formulation in (16) corresponds to a Lagrangian relaxation of the approach studied in [18].

A. Universal compressor

We propose a universal lossless compression formulation following the conventions of Weissman et al. [27, 35]. Our goal is to characterize $-\log(p_U(w))$, cf. (15). To do so, we use empirical entropy, which for stationary ergodic sources converges to the per-symbol entropy rate almost surely [23].

To define the empirical entropy, we first define the empirical symbol counts:

$$n_q(w,\alpha)[\beta] \triangleq |\{i \in [q+1, N] : w_{i-q}^{i-1} = \alpha, w^i = \beta\}|, \tag{17}$$

where q is the context depth [12, 37], $\beta \in \mathcal{R}_F$, $\alpha \in (\mathcal{R}_F)^q$, and w_i^j is the string comprising symbols i through j within w. We now define the order q conditional empirical probability for the context α as

$$p_q(w,\alpha)[\beta] \triangleq \frac{n_q(w,\alpha)[\beta]}{\sum_{\beta' \in \mathcal{R}_F} n_q(w,\alpha)[\beta']}, \tag{18}$$

and the order q conditional empirical entropy,

$$H_q(w) \triangleq -\frac{1}{N} \sum_{\alpha \in (\mathcal{R}_F)^q, \beta \in \mathcal{R}_F} n_q(w,\alpha)[\beta] \log_2 \left(p_q(w,\alpha)[\beta]\right), \tag{19}$$

where the sum is only over nonzero counts and probabilities.

Allowing the context depth $q \triangleq q_N = o(\log(N))$ to grow slowly with N, various universal compression algorithms can achieve the empirical entropy $H_q(\cdot)$ asymptotically [11, 12, 37]. On the other hand, no compressor can outperform the entropy rate. Additionally, for large N, the empirical symbol counts with context depth q provide a sufficiently precise characterization of the source statistics. Therefore, H_q provides a concise approximation to the per-symbol coding length of a universal compressor.

B. Markov chain Monte Carlo

Having approximated the coding length, we now describe how to optimize our objective function. We define the energy $\Psi^{H_q}(w)$ in an analogous manner to $\Psi^U(w)$, using $H_q(w)$ as our universal coding length (15):

$$\Psi^{H_q}(w) \triangleq N H_q(w) + c_4 \|y - \Phi w\|^2, \tag{20}$$

where $c_4 = c_2 \log_2(e)$. The minimization of this energy is analogous to minimizing $\Psi^U(w)$. The Boltzmann PMF is then defined as

$$p_s(w) \triangleq \frac{1}{\zeta_s} \exp(-s\Psi^{H_q}(w)), \tag{21}$$

where $s > 0$ is inversely related to temperature in simulated annealing and ζ_s is a normalization constant.

Ideally, our goal is to compute the globally minimum energy solution

$$x_{MAP}^{H_q} \triangleq \arg \min_{w \in (\mathcal{R}_F)^N} \Psi^{H_q}(w). \tag{22}$$

We use a stochastic MCMC relaxation [34] to achieve the globally minimum solution in the limit of infinite computation. In MCMC, the space $w \in (\mathcal{R}_F)^N$ is analogous to a statistical mechanical system, and at low temperatures the system tends toward low energies.

MCMC samples from the Boltzmann PMF (21) using a *Gibbs sampler*: in each iteration, a single element w_n is generated while the rest of w, $w^{\setminus n} \triangleq \{w_i : \ n \neq i\}$, remains unchanged. We denote by $w_1^{n-1}\beta w_{n+1}^N$ the concatenation of the initial portion of the output vector w_1^{n-1}, the symbol $\beta \in \mathcal{R}_F$, and the latter portion of the output w_{n+1}^N. The Gibbs sampler updates w_n by resampling from the PMF:

$$\begin{aligned} p_s(w_n = a|w^{\setminus n}) &= \frac{\exp\left(-s\Psi^{H_q}(w_1^{n-1}aw_{n+1}^N)\right)}{\sum_{b\in\mathcal{R}_F}\exp\left(-s\Psi^{H_q}(w_1^{n-1}bw_{n+1}^N)\right)} \\ &= \frac{1}{\sum_{b\in\mathcal{R}_F}\exp\left(-s\left[N\Delta H_q(w,n,b,a) + c_4\Delta d(w,n,b,a)\right]\right)}, \end{aligned} \tag{23}$$

where

$$\Delta H_q(w,n,b,a) \triangleq H_q(w_1^{n-1}bw_{n+1}^N) - H_q(w_1^{n-1}aw_{n+1}^N)$$

is the change in empirical entropy $H_q(w)$ (19) when $w_n = a$ is replaced by b, and

$$\Delta d(w,n,b,a) \triangleq \|y - \Phi(w_1^{n-1}bw_{n+1}^N)\|^2 - \|y - \Phi(w_1^{n-1}aw_{n+1}^N)\|^2 \tag{24}$$

is the change in $\|y - \Phi w\|^2$ when $w_n = a$ is replaced by b. The maximum change in the energy within an iteration of Algorithm 1 is then bounded by

$$\Delta_q = \max_{1\le n\le N}\max_{w\in(\mathcal{R}_F)^N}\max_{a,b\in\mathcal{R}_F}|N\Delta H_q(w,n,b,a) + c_4\Delta d(w,n,b,a)|. \tag{25}$$

Note that X is assumed bounded (cf. Condition 1) so that (24–25) are bounded as well. During the execution of the algorithm, we set a sequence of decreasing temperatures that takes into account the maximum change given in (25):

$$s_t \triangleq \ln(t)/(cN\Delta_q) \text{ for some } c > 1. \tag{26}$$

At low temperatures, i.e., large s_t, a small difference in energy $\Psi^{H_q}(w)$ drives a big difference in probability. Therefore, we begin at a high temperature where the Gibbs sampler can freely move around $(\mathcal{R}_F)^N$. As the temperature is reduced, the PMF becomes more sensitive to changes in energy (21), and the trend toward w with lower energy grows stronger. In each iteration, the Gibbs sampler modifies w_n in a random manner that resembles heat bath concepts in statistical physics. Although MCMC could sink into a local minimum, we decrease the temperature slowly enough that the randomness of Gibbs sampling eventually drives MCMC out of the local minimum toward the globally optimal x_{MAP}^U.

Pseudocode for our MCMC approach appears in Algorithm 1. We refer to the processing of a single location as an iteration and group the processing of the N different entries of w, randomly permuted, into super-iterations. During the minimization process, we refer to the approximation as w.

The following theorem is proven in Appendix B, following the framework espoused in [35].

Theorem 2: Let X be a stationary ergodic source that obeys Condition 1. Then the outcome w^r of Algorithm 1 after r iterations obeys

$$\lim_{r\to\infty}\Psi^{H_q}(w^r) = \min_{v\in(\mathcal{R}_F)^N}\Psi^{H_q}(v) = \Psi^{H_q}\left(x_{MAP}^{H_q}\right).$$

Algorithm 1 MCMC for Universal CS

Inputs: Initial point $x^* \in \mathbb{R}^n$, reproduction alphabet $\mathcal{R}_F$, noise variance σ_Z^2, number of super-iterations r, temperature constant $c > 1$

Outputs: Approximation w of x_{MAP}^U

Initialize w by quantizing x^* to $(\mathcal{R}_F)^N$

Compute $n_q(w, \alpha)[\beta],\ \forall\ \alpha \in (\mathcal{R}_F)^q,\ \beta \in \mathcal{R}_F$

for $t = 1$ to r **do** *// super-iteration*

  $s \leftarrow \ln(t)/(cN\Delta_q)$ *// $s = s_t$, cf. (26)*

  Draw permutation $\{1, \ldots, N\}$ at random

  for $t' = 1$ to N **do** *// iteration*

    Let n be component t' in permutation

    for all β in $\mathcal{R}_F$ **do** *// possible new w_n*

      Compute $\Delta H_q(w, n, \beta, w_n)$

      Compute $\Delta d(w, n, \beta, w_n)$

      Compute $p_s(w_n = \beta | w^{\setminus n})$

    end for

    Generate w_n using $p_s(\cdot | w^{\setminus n})$ *// Gibbs*

    Update $n_q(w, \alpha)[\beta],\ \forall\ \alpha \in (\mathcal{R}_F)^q,\ \beta \in \mathcal{R}_F$

  end for

end for

return w

Theorem 2 shows that Algorithm 1 matches the best-possible performance of the universal MAP estimator, which we believe to be close to the performance of the MMSE estimator (Conjecture 1 in Section IV). To gain some insight about the convergence process of MCMC, suppose that at iteration t the energy of the algorithm output $\Psi^{H_q}(w)$ has converged to a steady state (see Appendix B for details on convergence). We can then focus on the probabilistic ratio

$$\frac{p_{s_t}(w)}{p_{s_t}(x_{MAP}^{H_q})},$$

i.e., the ratio by which $p_{s_t}(w)$ is smaller than $p_{s_t}(x_{MAP}^{H_q})$ for arbitrary $w \in (\mathcal{R}_F)^N$. When we *square* the number of super-iterations from t to t^2, the inverse temperature is *doubled* from s_t to $2s_t$, and the corresponding ratio at time t^2 is

$$\frac{p_{2s_t}(w)}{p_{2s_t}(x_{MAP}^{H_q})} = \frac{\exp(-2s_t\Psi^{H_q}(w))}{\exp(-2s_t\Psi^{H_q}(x_{MAP}^{H_q}))} = \left(\frac{\exp(-s_t\Psi^{H_q}(w))}{\exp(-s_t\Psi^{H_q}(x_{MAP}^{H_q}))}\right)^2 = \left(\frac{p_{s_t}(w)}{p_{s_t}(x_{MAP}^{H_q})}\right)^2.$$

That is, between super-iterations t and t^2 the probability ratio between a higher energy $w \in (\mathcal{R}_F)^N$ to some minimal energy x^U_{MAP} is also squared.

The practical value of our proposed MCMC approach may be reduced due to its high computational cost, dictated by the number of iterations r required for convergence to the universal MAP estimator. Nonetheless, this approach provides a starting point toward further performance gains of more practical algorithms for computing the universal MAP estimator; furthermore, our experiments in Section VII will show that the performance of MCMC is comparable to and sometimes significantly better than existing state-of-the-art algorithms.

C. Computational challenges

Studying the pseudocode of Algorithm 1, we recognize that Lines 11–13 must be implemented efficiently, as they run $rN|\mathcal{R}_F|$ times. Lines 11 and 12 are especially challenging.

For Line 11, a naive update of $\Delta H_q(w, n, b, a)$ has complexity $O(|\mathcal{R}_F|^{q+1})$, cf. (19). To address this problem, Jalali and Weissman [35] recompute the empirical conditional entropy in $O(q|\mathcal{R}_F|)$ time only for the $O(q)$ contexts whose corresponding counts are modified [35]. The same approach can be used in Line 16, again reducing computation from $O(|\mathcal{R}_F|^{q+1})$ to $O(q|\mathcal{R}_F|)$.

We now focus on computation of $\Delta d(w, n, b, w_n)$ in Line 12. Define $v = y - \Phi w$. From (24) we get

$$\begin{aligned}\Delta d(w, n, b, w_n) &= \sum_{m=1}^{M} \left[(v_m - \Phi_{mn}(b - w_n))^2 - (v_m)^2\right] \\ &= \sum_{m=1}^{M} \left[2v_m\Phi_{mn}(w_n - b) + (\Phi_{mn}(w_n - b))^2\right] \\ &= 2(w_n - b)\langle v, \Phi_n\rangle + (w_n - b)^2\|\Phi_n\|^2,\end{aligned}$$

where Φ_n is column n of Φ. By pre-computing the inner product $\langle v, \Phi_n\rangle$ and squared ℓ_2 norm $\|\Phi_n\|^2$, Line 12 can be implemented in constant time. Seeing that the inner product and squared ℓ_2 norm require $O(M)$ time, which is aggregated over $|\mathcal{R}_F|$ calls per iteration to Line 12, $\Delta d(w, n, b, a)$ requires $O(Nr(M + |\mathcal{R}_F|))$ time in total. Combined with the computation for Line 11, and utilizing that $M \gg q|\mathcal{R}_F|^2$ in practice, the entire runtime of our algorithm is $O(rMN)$.

VI. Adaptive reproduction levels

While Algorithm 1 is a first step toward universal CS, it suffers from a large number of reproduction levels $|\mathcal{R}_F|$. In order to meet a target performance level, N must be large enough to ensure that $\mathcal{R}_F$ quantizes a broad enough range of values of $\mathbb{R}$ finely enough to represent the estimate $\widehat{x}$ well. For finite N, estimation performance using the reproduction levels (9) could suffer from high computational complexity.

To estimate better with finite N, we utilize reproduction levels that are *adaptive* instead of the fixed levels in $\mathcal{R}_F$. To do so, instead of $w \in (\mathcal{R}_F)^N$, we optimize over $u \in \mathcal{Z}^N$, where $|\mathcal{Z}| < |\mathcal{R}_F|$. The new alphabet $\mathcal{Z}$ does not directly correspond to real numbers. Instead, there is an adaptive mapping $\mathcal{A} : \mathcal{Z} \to \mathbb{R}$. Considering the energy

function (20), we now compute the empirical symbol counts $n_q(u,\alpha)[\beta]$, order q conditional empirical probabilities $p_q(u,\alpha)[\beta]$, and order q conditional empirical entropy $H_q(u)$ using $u \in \mathcal{Z}^N$, $\alpha \in \mathcal{Z}^q$, and $\beta \in \mathcal{Z}$, cf. (17), (18), and (19). Similarly, we use $\|y-\Phi\mathcal{A}(u)\|^2$ instead of $\|y-\Phi w\|^2$, where $\mathcal{A}(u)$ is the straightforward vector extension of $\mathcal{A}$. These modifications yield an adaptive energy function

$$\Psi_a^{H_q}(u) \triangleq NH_q(u) + c_4\|y-\Phi\mathcal{A}(u)\|^2.$$

We choose $\mathcal{A}_{opt}$ to optimize for squared ℓ_2 error,

$$\mathcal{A}_{opt} \triangleq \arg\min_{\mathcal{A}} \|y-\Phi\mathcal{A}(u)\|_2^2 = \arg\min_{\mathcal{A}} \left[\sum_{m=1}^{M}(y_m - [\Phi\mathcal{A}(u)]_m)^2\right],$$

where $[\Phi\mathcal{A}(u)]_m$ denotes the m^{th} entry of the vector $\Phi\mathcal{A}(u)$. The optimal mapping depends entirely on y, Φ, and u. From a coding perspective, describing $\mathcal{A}_{opt}(u)$ requires $H_q(u)$ bits for u and $|\mathcal{Z}|b\log\log(N)$ bits for $\mathcal{A}_{opt}$ to match the resolution of the nonadaptive alphabet $\mathcal{R}_F$, with $b>1$ an arbitrary constant [27]. The resulting coding length defines our universal prior.

A. *Optimization of reproduction levels*

We now describe the optimization procedure for $\mathcal{A}_{opt}$, which must be computationally efficient. Write

$$\Upsilon(\mathcal{A}) \triangleq \|y-\Phi\mathcal{A}(u)\|_2^2 = \sum_{m=1}^{M}\left(y_m - \sum_{n=1}^{N}\Phi_{mn}\mathcal{A}(u_n)\right)^2.$$

For $\Upsilon(\mathcal{A})$ to be minimal, we need zero-valued derivatives.

$$\begin{aligned}\frac{d\Upsilon(\mathcal{A})}{d\mathcal{A}(\beta)} &= -2\sum_{m=1}^{M}\left(y_m - \sum_{n=1}^{N}\Phi_{mn}\mathcal{A}(u_n)\right)\left(\sum_{n=1}^{N}\Phi_{mn}1_{\{u_n=\beta\}}\right)\\ &= 0,\ \forall\ \beta\in\mathcal{Z}.\end{aligned} \tag{27}$$

Define the location sets

$$\mathcal{L}_\beta \triangleq \{n : 1\le n\le N, u_n=\beta\}$$

for each $\beta\in\mathcal{Z}$, and rewrite the derivatives of $\Upsilon(\mathcal{A})$,

$$\frac{d\Upsilon(\mathcal{A})}{d\mathcal{A}(\beta)} = -2\sum_{m=1}^{M}\left(y_m - \sum_{\lambda\in\mathcal{Z}}\sum_{n\in\mathcal{L}_\lambda}\Phi_{mn}\mathcal{A}(\lambda)\right)\left(\sum_{n\in\mathcal{L}_\beta}\Phi_{mn}\right).$$

Let the per-character averaged column values be

$$\mu_{m\beta} \triangleq \sum_{n\in\mathcal{L}_\beta}\Phi_{mn}, \tag{28}$$

for each $m\in\{1,\dots,M\}$ and $\beta\in\mathcal{Z}$. We desire the derivatives to be zero, cf. (27):

$$0 = \sum_{m=1}^{M}\left(y_m - \sum_{\lambda\in\mathcal{Z}}\mathcal{A}(\lambda)\mu_{m\lambda}\right)\mu_{m\beta}.$$

Thus, we must satisfy the system of equations,

$$\sum_{m=1}^{M} y_m \mu_{m\beta} = \sum_{m=1}^{M} \left(\sum_{\lambda \in \mathcal{Z}} \mathcal{A}(\lambda) \mu_{m\lambda} \right) \mu_{m\beta}$$

for each $\beta \in \mathcal{Z}$. We can write the right hand side of each of these equations as

$$\begin{aligned} &\sum_{m=1}^{M} \left(\sum_{\lambda \in \mathcal{Z}} \mathcal{A}(\lambda) \mu_{m\lambda} \right) \mu_{m\beta} \\ &= \sum_{\lambda \in \mathcal{Z}} \mathcal{A}(\lambda) \sum_{m=1}^{M} \mu_{m\lambda} \mu_{m\beta}, \end{aligned}$$

for each $\beta \in \mathcal{Z}$. The system of equations can be described in matrix form as

$$\overbrace{\begin{bmatrix} \sum_{m=1}^{M} \mu_{m\beta_1} \mu_{m\beta_1} & \cdots & \sum_{m=1}^{M} \mu_{m\beta_{|\mathcal{Z}|}} \mu_{m\beta_1} \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{M} \mu_{m\beta_1} \mu_{m\beta_{|\mathcal{Z}|}} & \cdots & \sum_{m=1}^{M} \mu_{m\beta_{|\mathcal{Z}|}} \mu_{m\beta_{|\mathcal{Z}|}} \end{bmatrix}}^{\Omega} \begin{bmatrix} \mathcal{A}(\beta_1) \\ \vdots \\ \mathcal{A}(\beta_{|\mathcal{Z}|}) \end{bmatrix} = \overbrace{\begin{bmatrix} \sum_{m=1}^{M} y_m \mu_{m\beta_1} \\ \vdots \\ \sum_{m=1}^{M} y_m \mu_{m\beta_{|\mathcal{Z}|}} \end{bmatrix}}^{\Theta}. \tag{29}$$

Note that by writing μ as a matrix with entries indexed by row m and column β given by (28), we can write Ω as a Gram matrix, $\Omega = \mu^T \mu$, and we also have $\Theta = \mu^T y$. The optimal $\mathcal{A}$ can be computed as a $|\mathcal{Z}| \times 1$ vector

$$\mathcal{A}_{opt} = \Omega^{-1} \Theta = (\mu^T \mu)^{-1} \mu^T y$$

if the $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix Ω is invertible. We note in passing that numerical stability is improved by regularizing Ω. Note also that

$$\|y - \Phi \mathcal{A}(u)\|^2 = \sum_{m=1}^{M} \left(y_m - \sum_{\beta \in \mathcal{Z}} \mu_{m\beta} \mathcal{A}_{opt}(\beta) \right)^2, \tag{30}$$

which can be computed in $O(M|\mathcal{Z}|)$ time instead of $O(MN)$.

B. Computational complexity

Pseudocode for the adaptive reproduction level estimation appears as Algorithm 2. We discuss computational requirements for each line of the pseudocode that is run within the inner loop.

- In Line 12, the differences in empirical conditional entropy can be computed in $O(q|\mathcal{Z}|)$ time as demonstrated by Jalali and Weissman [35].
- In Line 13, we update $\mu_{m\beta}$ for $m \in \{1, \ldots, M\}$ in $O(M)$ time.
- Line 14 updates Ω. Because we only need to update $O(1)$ columns and $O(1)$ rows, each such column and row contains $O(|\mathcal{Z}|)$ entries, and each entry is a sum over $O(M)$ terms, we need $O(M|\mathcal{Z}|)$ time.
- Line 15 requires to invert Ω in $O(|\mathcal{Z}|^3)$ time.
- Line 16 requires $O(M|\mathcal{Z}|)$ time, cf. (30).
- Line 17 requires $O(|\mathcal{Z}|)$ time.

In practice we typically have $M \gg |\mathcal{Z}|^2$, and so the aggregate complexity is $O(rMN|\mathcal{Z}|)$, which is greater than the computational complexity of the fixed reproduction level Algorithm 1 by a factor of $O(|\mathcal{Z}|)$.

Algorithm 2 MCMC with Adaptive Levels

Inputs: Initial point $x^* \in \mathbb{R}^n$, adaptive alphabet $\mathcal{Z}$, noise variance σ_Z^2, number of super-iterations r, temperature constant $c > 1$

Outputs: Approximation $\mathcal{A}(u)$ of x_{MAP}^U

Initialize u and $\mathcal{A}$ from x^* and $\mathcal{Z}$

Compute $n_q(u, \alpha)[\beta],\ \forall\ \alpha \in \mathcal{Z}^q,\ \beta \in \mathcal{Z}$

Initialize Ω

for $t = 1$ to r **do** *// super-iteration*

  $s \leftarrow \ln(t)/(cN\Delta_q)$ *// $s = s_t$, cf. (26)*

  Draw permutation $\{1, \ldots, N\}$ at random

  for $t' = 1$ to N **do** *// iteration*

    Let n be component t' in permutation

    for all β in $\mathcal{Z}$ **do** *// possible new u_n*

      Compute $\Delta H_q(u, n, \beta, u_n)$

      Compute $\mu_{m\beta}, \forall\ m \in \{1, \ldots, M\}$

      Update Ω *// $O(1)$ rows and columns*

      Compute $\mathcal{A}_{opt}$ *// invert Ω*

      Compute $\|y - \Phi\mathcal{A}(u_1^{n-1}\beta u_{n+1}^N)\|^2$

      Compute $p_s(u_n = \beta | u^{\setminus n})$

    end for

    $\widetilde{u}_n \leftarrow u_n$ *// save previous value*

    Generate u_n using $p_s(\cdot | u^{\setminus n})$ *// Gibbs*

    Update $n_q(\cdot)[\cdot]$ at $O(q)$ relevant locations

    Update $\mu_{m\beta},\ \forall\ m,\ \beta \in \{u_n, \widetilde{u}_n\}$

    Update Ω *// $O(1)$ rows and columns*

  end for

end for

return $\mathcal{A}(u)$

VII. Numerical results

We implemented universal MCMC estimation (Algorithm 2) in Matlab and tested it using several stationary ergodic sources. Although not all of our sources obey Condition 2, they nonetheless illustrate the performance of the algorithm. Our code is available for download at http://people.engr.ncsu.edu/dzbaron/software/UCS_BaronDuarte/. We chose an alphabet of size $|\mathcal{Z}| = 7$ for all sources tested. For each source, inputs x of length $N = 10000$

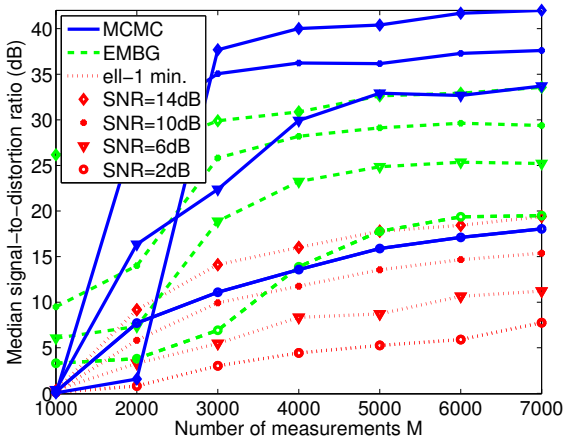

Fig. 2. *Universal MCMC, EMBG, and ℓ_1-norm minimization recovery results for a two-state Markov source with nonzero entries of value $+1$ as a function of the number of Gaussian random measurements M for different SNR values. MCMC outperforms both ℓ_1-norm minimization and EMBG due to the two quantization levels required to encode the observed distribution.*

were generated. Each such x was multiplied by a random Gaussian matrix Φ. We then added measurement noise z whose variance was selected to ensure that the signal to noise ratio (SNR) was 2, 6, 10, or 14 dB. We compared the performance of universal MCMC estimation starting from an initial estimate $x^* = \Phi^T y$ to that of ℓ_1-norm minimization recovery and the expectation-maximization Bernoulli-Gaussian approximate message passing algorithm (EMBG) of [24]; for each value of M and SNR, we plot the median signal-to-distortion ratio of each algorithm over 25 draws of x, Φ, and z; the median yields somewhat smoother plots than the mean signal-to-distortion ratio, which has the same flavor. Although MCMC was slower than the ℓ_1-norm minimizer [38] and the EMBG algorithm used in our simulations, its runtime of several minutes was nonetheless reasonable.

Results for the switching source (Fig. 1) were highlighted in Section I, where it is seen that MCMC performs well while ℓ_1-norm minimization and EMBG fail due to the input not being canonically sparse.

Next, we examine three sources for sparse signals whose sparse supports (the locations of the nonzero entries) are generated by a two-state Markov state machine (nonzero and zero states). The transition from zero to nonzero state for adjacent entries has probability $\frac{3}{970}$, while the transition from nonzero to zero state for adjacent entries has probability 10%; these parameters yield 3% sparsity on average. The three sources considered differ in the distribution of the nonzero values, and ℓ_1-norm minimization offers reasonable recovery performance.

For our first two-state Markov source, the nonzeros are set to a *constant* value +1; such signal structure has low entropy. Figure 2 shows that MCMC consistently outperforms ℓ_1-norm minimization as expected; MCMC also outperforms EMBG due to only two quantization levels being necessary to accurately model the observed distribution.

For our second two-state Markov source, the nonzeros are drawn from a *Rademacher* distribution, i.e., uniform over $\{-1, +1\}$ (Fig. 3). Here, universal MCMC estimation performs well for low to moderate SNR, outperforming both ℓ_1-norm minimization and EMBG for sufficiently large values of M. Surprisingly, MCMC fails for high SNR.

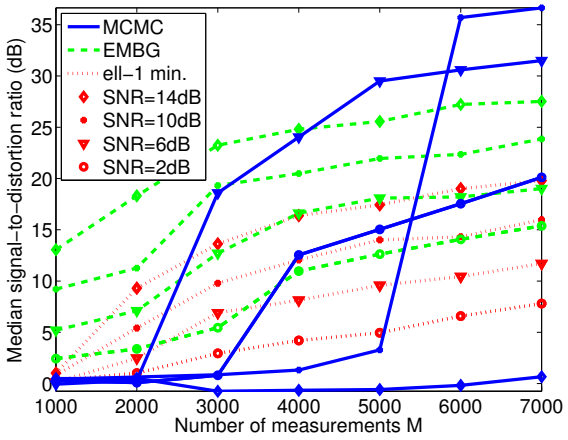

Fig. 3. *Universal MCMC, EMBG, and ℓ_1-norm minimization recovery results for a two-state Markov source with nonzero entries drawn from a Rademacher (± 1) distribution as a function of the number of Gaussian random measurements M for different SNR values. MCMC outperforms ℓ_1-norm minimization and EMBG for most cases shown, with the surprising exception of $SNR = 14dB$.*

To explain this behavior, we note that for high SNR the constant c_2 in (6) is large, and the Gibbs sampler (23) is strongly motivated to minimize the quadratic term $\|y - \Phi\mathcal{A}(u)\|^2$ while accommodating larger values for the empirical entropies H_q, which in turn allows for more complex sources to be used in the estimate. Such behavior within the universal MCMC algorithm will push its estimates away from the low-complexity priors that we seek to promote, and may also appear for other sources once the SNR is sufficiently high. We plan to further study this behavior in light of the parameters of the Gibbs sampler, including the choice of initial point x^*.

For our third two-state Markov source featuring nonzero values following a *uniform* distribution $U[0, 1]$ (Fig. 4), MCMC performs poorly. The problem we saw during execution is that the adaptive alphabet $\mathcal{A}$ spends many of the representation levels in $\mathcal{Z}$ on zero-valued entries of the signal, and only one level for the nonzeros, which leads to poor quantization. We noticed that the quantization step for the initial estimate x^* in Step 3 of Algorithm 2 pushes us away from convergence. We conjecture that a key hurdle is the calculation of an initial quantizer that is suitable for a variety of sources.

Our last source generates the entries of the signal by drawing them independently from a Bernoulli distribution, where each x_n was +1 with probability 3% and zero otherwise. For this source, MCMC outperforms ℓ_1-norm minimization and EMBG, except when the number of measurements M is low. We also compared the performance of the two algorithms to the MMSE achievable in the Bayesian regime with known statistics [39]; similar computations were performed by Guo and Wang [26]. Interestingly, the squared error achieved by MCMC is thrice the MMSE. We are left to wonder whether one could approach the mean squared error bound given in Conjecture 1 in the limit of infinitely many super-iterations.

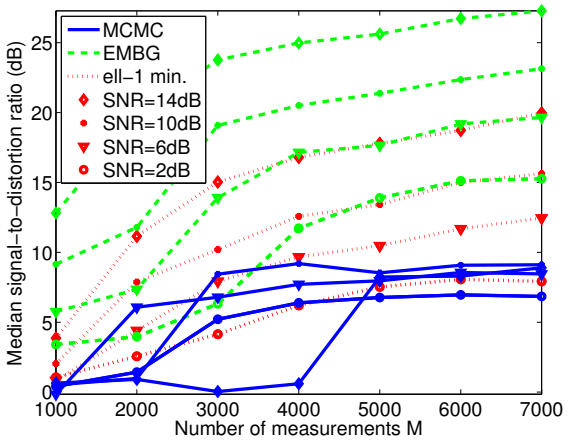

Fig. 4. *Universal MCMC, EMBG, and ℓ_1-norm minimization recovery results for a two-state Markov source with nonzero entries drawn from a uniform distribution $U[0,1]$ as a function of the number of Gaussian random measurements M for different SNR values. The performance of universal MCMC estimation is hampered likely due to the continuous-valued nature of the source studied.*

VIII. Conclusions

This paper provides an initial step towards the formulation of computationally feasible universal algorithms for signal estimation from linear measurements. Here, universality denotes the property that the algorithm need not be informed of the probability distribution for the recorded signal prior to acquisition; rather, the algorithm simultaneously builds estimates both of the observed signal and its distribution. Inspired by the Kolmogorov sampler [13] and motivated by the need for a computationally tractable framework, our contribution focuses on stationary ergodic signal sources and relies on a maximum *a posteriori* (MAP) estimation algorithm, for which it is possible to establish asymptotic near-optimality with high probability. The algorithm is then implemented via a Markov Chain Monte-Carlo (MCMC) formulation that is proven to be convergent in the limit of infinite computation. While the practical impact of an MCMC-based approach is mitigated due to its high computational cost, our experiments have shown that its performance is comparable to and in some cases significantly better than existing state-of-the-art algorithms, with a significant advantage observed for sources that are non-sparse.

Our expectation is that these initial results will spur additional work to improve the computational cost of implementing universal MAP estimation from linear measurements, including techniques that accelerate the convergence of MCMC. An alternative approach to implement universal MAP estimation could be obtained by moving from MCMC to other optimization algorithm such as belief propagation.

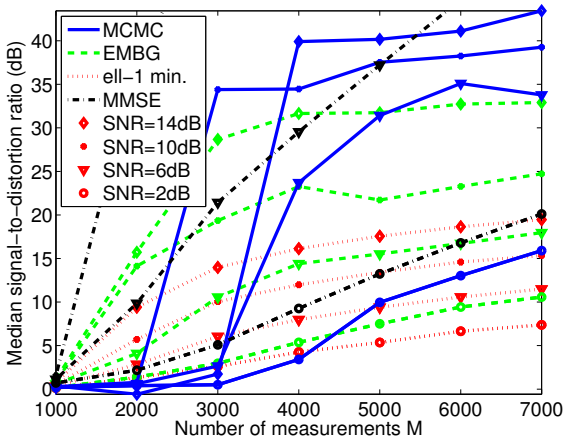

Fig. 5. *Universal MCMC, EMBG, and ℓ_1-norm minimization recovery results for a source with i.i.d. Bernoulli entries with nonzero probability of 3% as a function of the number of Gaussian random measurements M for different SNR values. MCMC outperforms ℓ_1-norm minimization and EMBG for a sufficiently large number of measurements M.*

Appendix A. Proof of Theorem 1

The lower bound $\Psi^X(x_{MAP}) \leq \Psi^X(\widetilde{x}_{MAP})$ is trivial, because x_{MAP} is the MAP solution. For the upper bound,

$$\begin{aligned}
& \|y - \Phi\widetilde{x}_{MAP}\| \\
\leq\ & \|y - \Phi x_{MAP}\| + \|\Phi(x_{MAP} - \widetilde{x}_{MAP})\| \\
\leq\ & \|y - \Phi x_{MAP}\| + \|\Phi\|\|x_{MAP} - \widetilde{x}_{MAP}\| \\
\leq\ & \|y - \Phi x_{MAP}\| + \|\Phi\|\sqrt{N}\|x_{MAP} - \widetilde{x}_{MAP}\|_\infty \qquad (31) \\
\leq\ & \|y - \Phi x_{MAP}\| + O\left(\frac{\sqrt{N}}{\log(N)}\right), \qquad (32)
\end{aligned}$$

where the ℓ_∞ norm in (31) isolates the largest absolute difference between x_{MAP} and $\widetilde{x}_{MAP}$, which is $O(\log(N)^{-1})$ (3); and for the distribution we assumed for the matrix Φ we obtain $\|\Phi\| = 1 + \delta^{-0.5} = O(1)$ almost surely asymptotically [40], where δ is the aspect ratio (2). Observe that $\|y - \Phi x_{MAP}\| = O(\sqrt{N})$, because the per-element signal to noise ratio of the measurement process (1) is a function of σ_Z and therefore finite. Substituting this observation into (32),

$$\begin{aligned}
& \|y - \Phi\widetilde{x}_{MAP}\|^2 \\
\leq\ & \left(\|y - \Phi x_{MAP}\| + O\left(\frac{\sqrt{N}}{\log(N)}\right)\right)^2 \\
=\ & \|y - \Phi x_{MAP}\|^2 + O\left(\frac{N}{\log(N)}\right) \\
=\ & \|y - \Phi x_{MAP}\|^2 + o(N) \qquad (33)
\end{aligned}$$

almost surely asymptotically. That is, discretization does not change the noise hypothesized by the MAP estimator by much.

We now show that $\ln(f_X(\widetilde{x}_{MAP}))$ is similar to $\ln(f_X(x_{MAP}))$. Owing to our reproduction levels (3),

$$\|x_{MAP} - \widetilde{x}_{MAP}\|_1 \leq O\left(\frac{N}{\log(N)}\right) = o(N).$$

Because the log derivatives of f_X are bounded, cf. Condition 2,

$$\begin{aligned} & |\ln(f_X(\widetilde{x}_{MAP})) - \ln(f_X(x_{MAP}))| \\ < \quad & \rho\|x_{MAP} - \widetilde{x}_{MAP}\|_1 = o(N). \end{aligned} \tag{34}$$

Combining (7), (33), and (34), we see that $\Psi^X(\widetilde{x}_{MAP}) - \Psi^X(x_{MAP}) = o(N)$ almost surely asymptotically. □

APPENDIX B. PROOF OF THEOREM 2

Our proof mimics a very similar proof presented in [35] for lossy source coding; we include all details for completeness. The proof technique relies on mathematical properties of non-homogeneous (e.g., time-varying) Markov Chains (MCs) [41]. Through the proof, $\mathcal{S} \triangleq (\mathcal{R}_F)^N$ denotes the state space of the MC of codewords generated by Algorithm 1, with size $|\mathcal{S}| = |\mathcal{R}_F|^N$. We define a stochastic transition matrix $P_{(t)}$ from $\mathcal{S}$ to itself given by the Boltzmann distribution for super-iteration t in Algorithm 1. Similarly, $\pi_{(t)}$ defines the stable-state distribution on $\mathcal{S}$ for $P_{(t)}$, satisfying $\pi_{(t)}P_{(t)} = \pi_{(t)}$.

Definition 1: [41] *Dobrushin's ergodic coefficient* of a MC transition matrix P is denoted by $\delta(P)$ and defined as

$$\delta(P) \triangleq \max_{1 \leq i,j \leq N} \frac{1}{2}\|p_i - p_j\|_1,$$

where p_i denotes the i^{th} row of P, $1 \leq n \leq N$.

From the definition, $0 \leq \delta(P) \leq 1$. Moreover, the ergodic coefficient can be rewritten as

$$\delta(P) = 1 - \min_{1 \leq i,j \leq N} \sum_{k=1}^{N} \min(p_{ik}, p_{jk}), \tag{35}$$

where p_{ij} denotes the entry of P at the i^{th} row and j^{th} column.

We group the product of transition matrices across super-iterations as $P_{(t_1 \to t_2)} = \prod_{t=t_1}^{t_2} P_{(t)}$. There are two common characterizations for the stable-state behavior of a non-homogeneous MC.

Definition 2: [41] A non-homogeneous MC is called *weakly ergodic* if for any distributions μ and ν over the state space $\mathcal{S}$, and any $t_1 \in \mathbb{N}$,

$$\limsup_{t_2 \to \infty} \|\mu P_{(t_1 \to t_2)} - \nu P_{(t_1 \to t_2)}\|_1 = 0.$$

Similarly, a non-homogeneous MC is called *strongly ergodic* if there exists a distribution π over the state space $\mathcal{S}$ such that for any distribution μ over $\mathcal{S}$, and any $t_1 \in \mathbb{N}$,

$$\limsup_{t_2 \to \infty} \|\mu P_{(t_1 \to t_2)} - \pi\|_1 = 0.$$

We will use the following two theorems from [41] in our proof.

Theorem 3: [41] A MC is weakly ergodic if and only if there exists a sequence of integers $0 \leq t_1 \leq t_2 \leq \ldots$ such that

$$\sum_{i=1}^{\infty} \left(1 - \delta\left(P_{(t_i \to t_{i+1})}\right)\right) = \infty.$$

Theorem 4: [41] Let a MC be weakly ergodic. Assume that there exists a sequence of probability distributions $\{\pi_{(t)}\}_{i=1}^{\infty}$ on the state space $\mathcal{S}$ such that $\pi_{(t)}P_{(i)} = \pi_{(t)}$. Then the MC is strongly ergodic if

$$\sum_{t=1}^{\infty} \|\pi_{(t)} - \pi_{(t+1)}\|_1 < \infty.$$

The rest of proof is structured as follows. First, we show that the sequence of stable-state distributions for the MC used by Algorithm 1 converges to a uniform distribution over the set of sequences that minimize the energy function as the iteration count t increases. Then, we show using Theorems 3 and 4 that the non-homogeneous MC used in Algorithm 1 is strongly ergodic, which by the definition of strong ergodicity implies that Algorithm 1 always converges to the stable distribution found above. This implies that the outcome of Algorithm 1 converges to a minimum-energy solution as $t \to \infty$, completing the proof of Theorem 2.

We therefore begin by finding the stable-state distribution for the non-homogeneous MC used by Algorithm 1. At each super-iteration t, the distribution defined as

$$\pi_{(t)}(w) \triangleq \frac{\exp(-s_t \Psi^{H_q}(w))}{\sum_{z \in \mathcal{S}} \exp(-s_t \Psi^{H_q}(z))} = \frac{1}{\sum_{z \in \mathcal{S}} \exp(-s_t(\Psi^{H_q}(z) - \Psi^{H_q}(w)))}. \tag{36}$$

satisfies $\pi_{(t)}P_{(t)} = \pi_{(t)}$, cf. (23). We can show that the distribution $\pi_{(t)}$ converges to a uniform distribution over the set of sequences that minimize the energy function, i.e.,

$$\lim_{t \to \infty} \pi_{(t)}(w) = \begin{cases} 0 & y \notin \mathcal{H}, \\ \frac{1}{|\mathcal{H}|} & w \in \mathcal{H}, \end{cases} \tag{37}$$

where $\mathcal{H} = \{w \in \mathcal{S} \text{ s.t. } \Psi^{H_q}(w) = \min_{z \in \mathcal{S}} \Psi^{H_q}(z)\}$. To show (37), we will show that $\pi_{(t)}(w)$ is increasing for $w \in \mathcal{H}$ and eventually decreasing for $w \in \mathcal{H}^C$. Since for $w \in \mathcal{H}$ and $z \in \mathcal{S}$ we have $\Psi^{H_q}(z) - \Psi^{H_q}(w) \geq 0$, for $t_1 < t_2$ we have

$$\sum_{z \in \mathcal{S}} \exp(-s_{t_1}(\Psi^{H_q}(z) - \Psi^{H_q}(w))) \geq \sum_{z \in \mathcal{S}} \exp(-s_{t_2}(\Psi^{H_q}(z) - \Psi^{H_q}(w))),$$

which together with (36) implies $\pi_{(t_1)}(w) \leq \pi_{(t_2)}(w)$. On the other hand, if $w \in \mathcal{H}^C$, then

$$\pi_{(t)}(w) = \frac{1}{\sum_{z:\Psi^{H_q}(z) \geq \Psi^{H_q}(w)} \exp(-s_t(\Psi^{H_q}(z) - \Psi^{H_q}(w))) + \sum_{z:\Psi^{H_q}(z) < \Psi^{H_q}(w)} \exp(-s_t(\Psi^{H_q}(z) - \Psi^{H_q}(w)))}. \tag{38}$$

For sufficiently large s_t, the denominator of (38) is dominated by the second term, which increases when s_t increases, and therefore $\pi_{(t)}(w)$ decreases for $w \in \mathcal{H}^C$ as t increases. Finally, since all sequences $w \in \mathcal{H}$ have the same energy $\Psi^{H_q}(w)$, it follows that the distribution is uniform over the symbols in $\mathcal{H}$.

Having shown convergence of the non-homogenous MC's stable-state distributions, we now show that the non-homogeneous MC is strongly ergodic. The transition matrix $P_{(t)}$ of the MC at iteration t depends on the temperature s_t in (26) used within Algorithm 1. We first show that the MC used in Algorithm 1 is weakly ergodic via Theorem 3; the proof of the following Lemma is given in Appendix C.

Lemma 1: The ergodic coefficient of $P_{(t)}$ for any $t \geq 0$ is upper bounded by

$$\delta\left(P_{(t)}\right) \leq 1 - \exp(-s_t N \Delta_q),$$

where Δ_q is defined in (25).

We note in passing that Condition 1 ensures that Δ_q is finite. Using Lemma 1 and (26), we can evaluate the sum given in Theorem 3 as

$$\sum_{j=1}^{\infty}\left(1-\delta\left(P_{(j)}\right)\right) \geq \sum_{j=1}^{\infty}\exp(-s_j N\Delta_q) = \sum_{j=1}^{\infty}\frac{1}{j^{1/c}} = \infty,$$

and therefore the non-homogeneous MC defined by $\{P_{(t)}\}_{t=1}^{\infty}$ is weakly ergodic. Now we can use Theorem 4 to show that the MC is strongly ergodic by proving that

$$\sum_{t=1}^{\infty}\|\pi_{(t)} - \pi_{(t+1)}\|_1 < \infty. \tag{39}$$

Since we know from earlier in the proof that $\pi_{(t)}(w)$ is increasing for $w \in \mathcal{H}$ and eventually decreasing for $w \in \mathcal{H}^C$, there exists a $t_0 \in \mathbb{N}$ such that for any $t_1 > t_0$,

$$\begin{aligned}
\sum_{t=t_0}^{t_1}\|\pi_{(t)} - \pi_{(t+1)}\|_1 &= \sum_{w\in\mathcal{H}}\sum_{t=t_0}^{t_1}\left(\pi_{(t+1)}(w) - \pi_{(t)}(w)\right) + \sum_{w\notin\mathcal{H}}\sum_{t=t_0}^{t_1}\left(\pi_{(t)}(w) - \pi_{(t+1)}(w)\right) \\
&= \sum_{w\in\mathcal{H}}\left(\pi_{(t_1+1)}(w) - \pi_{(t_0)}(w)\right) + \sum_{w\notin\mathcal{H}}\left(\pi_{(t_0)}(w) - \pi_{(t_1+1)}(w)\right) \\
&= \|\pi_{(t_1+1)} - \pi_{(t_0)}\|_1 \leq \|\pi_{(t_1+1)}\|_1 + \|\pi_{(t_0)}\|_1 = 2.
\end{aligned}$$

Since the right hand side does not depend on t_1, we have that $\sum_{t=1}^{\infty}\|\pi_{(t)} - \pi_{(t+1)}\|_1 < \infty$. This implies that the non-homogeneous MC used by Algorithm 1 is strongly ergodic, and thus completes the proof of Theorem 2.

APPENDIX C. PROOF OF LEMMA 1

Let y_1, y_2 be two arbitrary sequences in $\mathcal{S}$. The probability of transitioning from a given state to a neighboring state in an iteration within iteration t' of super-iteration t of Algorithm 1 is given by (23), and can be rewritten as

$$\begin{aligned}
P_{(t,t')}(w_1^{t'-1}aw_{t'+1}^N | w_1^{t'-1}bw_{t'+1}^N) = p_{s_t}(w_{t'} = a|w^{\setminus t'}) &= \frac{\exp\left(-s_t\Psi^{H_q}(w_1^{t'-1}aw_{t'+1}^N)\right)}{\sum_{b\in\mathcal{R}_F}\exp\left(-s_t\Psi^{H_q}(w_1^{t'-1}bw_{t'+1}^N)\right)} \\
&= \frac{\exp\left(-s_t\left(\Psi^{H_q}(w_1^{t'-1}aw_{t'+1}^N) - \Psi^{H_q}_{\min,t'}(w_1^{t'-1}, w_{t'+1}^N)\right)\right)}{\sum_{b\in\mathcal{R}_F}\exp\left(-s_t\left(\Psi^{H_q}(w_1^{t'-1}bw_{t'+1}^N) - \Psi^{H_q}_{\min,t'}(w_1^{t'-1}, w_{t'+1}^N)\right)\right)} \\
&\geq \frac{\exp(-s_t\Delta_q)}{|\mathcal{R}_F|},
\end{aligned}$$

where $\Psi^{H_q}_{\min,t'}(w_1^{t'-1}, w_{t'+1}^N) = \min_{\alpha \in \mathcal{R}_F} \Psi^{H_q}(w_1^{t'-1} \alpha w_{t'+1}^N)$. Therefore, the smallest probability of transition from y_1 to y_2 within super-iteration t of Algorithm 1 is bounded by

$$\min_{y_1, y_2 \in \mathcal{R}_F} P_{(t)}(y_2|y_1) \geq \prod_{t'=1}^{N} \frac{\exp(-s_t \Delta_q)}{|\mathcal{R}_F|} = \frac{\exp(-s_t N \Delta_q)}{|\mathcal{R}_F|^N} = \frac{\exp(-s_t N \Delta_q)}{|\mathcal{S}|}.$$

Using the alternative definition of the ergodic coefficient given in (35),

$$\delta\left(P_{(t)}\right) = 1 - \min_{y_1, y_2 \in \mathcal{S}} \sum_{z \in \mathcal{S}} \min(P_{(t)}(z|y_1), P_{(t)}(z|y_2)) \leq 1 - |\mathcal{S}| \frac{\exp(-s_t N \Delta_q)}{|\mathcal{S}|} = 1 - \exp(-s_t N \Delta_q),$$

proving the lemma.

ACKNOWLEDGMENTS

Preliminary conversations with Deanna Needell and Tsachy Weissman framed our thinking about universal compressed sensing. Phil Schniter was instrumental in formulating the proposed framework and shepherding our progress through detailed conversations, feedback on our drafts, and probing questions. Final thanks to Jin Tan for thoroughly proofreading our manuscript.

REFERENCES

[1] D. Baron and M. F. Duarte, "Universal MAP estimation in compressed sensing," in *Proc. 49th Annual Allerton Conf. Comm., Control, Computing*, Sep. 2011.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

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

[4] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[5] M. W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *ICML '08: Proc. 25th Int. Conf. Machine learning*, 2008, pp. 912–919.

[6] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, pp. 269–280, Jan. 2010.

[7] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4036–4048, Sep. 2006.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Workshop Neural Info. Proc. Sys. (NIPS)*, Vancouver, Canada, Dec. 2008.

[10] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 130–144, Jan. 2012.

[11] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[12] J. Rissanen, "A universal data compression system," *IEEE Trans. Inf. Theory*, vol. 29, no. 5, pp. 656–664, 1983.

[13] D. L. Donoho, "The Kolmogorov sampler," Department of Statistics Technical Report 2002-4, Stanford University, Stanford, CA, Jan. 2002.

[14] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *J. ACM*, vol. 13, no. 4, pp. 547–569, 1966.

[15] R. J. Solomonoff, "A formal theory of inductive inference. Part I," *Inf. and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[16] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Inf. Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[17] M. Li and P. M. B. Vitanyi, *An introduction to Kolmogorov complexity and its applications*, Springer-Verlag, New York, 2008.

[18] S. Jalali and A. Maleki, "Minimum complexity pursuit," in *Proc. 49th Annual Allerton Conf. Comm., Control, Computing*, Sep. 2011.

[19] D. Baron, "Information complexity and estimation," in *Fourth Workshop Inf. Theoretic Methods Science Eng. (WITMSE 2011)*, Aug. 2011.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[22] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Comput J*, vol. 11, no. 2, pp. 185–194, 1968.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2006.

[24] J. P. Vila and P. Schniter, "Expectation-maximization bernoulli-gaussian approximate message passing," in *Proc. 45th Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2011.

[25] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," *CoRR*, vol. arXiv:1001.2228v1, Jan. 2010.

[26] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, 2008.

[27] D. Baron and T. Weissman, "An MCMC approach to lossy compression of continuous sources," in *Proc. Data Compression Conf. (DCC)*, Mar. 2010, pp. 40–48.

[28] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[29] D. Donoho, H. Kakavand, and J. Mammen, "The simplest solution to an underdetermined system of linear equations," in *Int. Symp. Inf. Theory (ISIT2006)*, Jul. 2006, pp. 1924–1928.

[30] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.

[31] J. D. Haupt and R. Nowak, *Compressed Sensing: Theory and Applications*, chapter Adaptive sensing for sparse recovery, Cambridge Univ. Press, 2012.

[32] I. Ramírez and G. Sapiro, "An MDL framework for sparse coding and dictionary learning," *IEEE Trans. Signal Proc.*, 2011, To appear.

[33] R. C. Grinold and R. N. Kahn, *Active portfolio management: a quantitative approach for providing superior returns and controlling risk*, McGraw-Hill Companies, 2000.

[34] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, pp. 721–741, Nov. 1984.

[35] S. Jalali and T. Weissman, "Rate-distortion via Markov chain Monte Carlo," in *Proc. Int. Symp. Inf. Theory (ISIT2008)*, Jul. 2008, pp. 852–856.

[36] E. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1465–1476, Sep. 1997.

[37] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

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

[39] D. Guo, D. Baron, and S. Shamai, "A single-letter characterization of optimal noisy compressed sensing," in *Proc. 47th Allerton Conf. Commun., Control, and Comput.*, Sep. 2009.

[40] M. Ledoux, *The concentration of measure phenomenon*, vol. 89, Amer. Math. Society, 2001.

[41] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, vol. 31, Springer Verlag, 1999.