Universal MAP Estimation in Compressed Sensing

Dror Baron
Electrical and Computer Engineering
North Carolina State University
Raleigh, NC 27695
Email: barondror@ncsu.edu

Marco F. Duarte
Electrical and Computer Engineering
University of Massachusetts
Amherst, MA 01003
Email: mduarte@ecs.umass.edu

***Abstract*—We study the compressed sensing (CS) 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 provide initial theoretical, algorithmic, and experimental evidence based on maximum *a posteriori* (MAP) estimation that shows 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 it as a prior during recovery, it is indeed possible to accurately estimate x from y [1, 2].

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 (if not all) 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. The large majority of recovery algorithms pose an algebraic prior on the signal x. A second, separate class of Bayesian CS recovery algorithms poses a probabilistic prior on x, albeit still requiring a sparsity promoting model for the coefficients of x in a known transform domain [3–5]. 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 [6]. 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 [7–9].

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 *universal* to the particular statistics of the signal [10–12]. In this paper, we make an initial contribution in this direction by formulating an algorithm for recovery of arbitrary stationary ergodic sources of low complexity. In contrast to the existing CS recovery literature, our algorithm does not necessarily require the standard sparsity or compressibility prior. Instead, our approach is inspired by the Kolmogorov sampler (KS) [13, 14], a universal denoising algorithm. Both our approach and KS are based on the minimization of the Kolmogorov complexity [15–17] of a source, which can be accurately estimated for signals of interest via 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.

While our work is only an initial effort in the direction of universal estimation, we make several different contributions in this paper. First, we show that the maximum *a posteriori* (MAP) risk of an estimator based on a specific quantization grid converges asymptotically to the risk of the classical (known statistics) MAP estimator. Second, we propose a recovery algorithm based on Markov chain Monte Carlo (MCMC) to approximate this estimation procedure. We believe that for a sufficiently large number of randomized measurements and for well-behaved sources, the output of our MCMC recovery algorithm based on a universal prior converges in the limit of increased runtime to the correct MAP estimate; a proof of this result is part of ongoing work. Third, we identify computational bottlenecks in the implementation of our MCMC estimator and show approaches to solve them at low computational complexity. Fourth, we showcase encouraging experimental results that show recovery performance for a variety of types of signal structures (or statistics) that meets or exceeds that of the popular ℓ_1-norm minimization. For example, 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

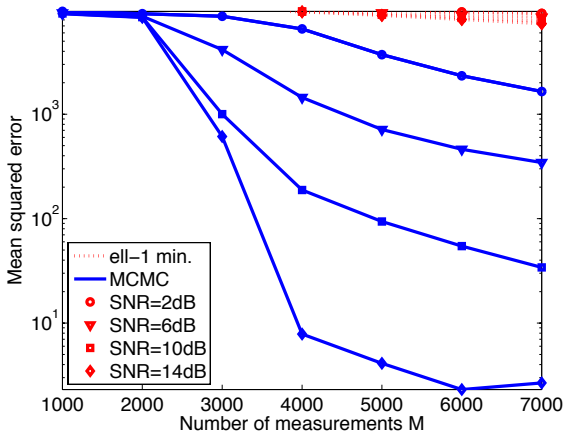

Fig. 1. *Universal MCMC and ℓ_1-norm minimization recovery results for the four-state Markov switching source of length $N = 10000$ as a function of the number of Gaussian random measurements M for different SNR values. MCMC significantly outperforms ℓ_1-norm minimization, which fails due to the signal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

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 also show some challenges faced by MCMC recovery of certain classes of sparse signals; we identify properties of the algorithm that cause these challenges, which remain to be addressed.

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 concrete MCMC algorithms for universal MAP estimation, and Section VI presents initial experimental results. We conclude with the proof of our main theoretical result 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. *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, $\Phi(m,n) \sim \mathcal{N}(0, \frac{1}{M})$.[1] 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 zero-mean 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., [18–22]. Since x was generated by an unknown source, we must search for an estimation mechanism that is agnostic to the specific distribution f_X.

[1]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.

B. Quantization

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 discrete, albeit still infinite. In Section IV, we describe an estimation approach that reduces the complexity of the problem from infinite to finite.

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],

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

where $K(x)$ denotes the Kolmogorov complexity of x, defined as the length of the shortest input to a Turing machine [23] 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 lossless compression algorithms [10, 11] 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 KS to the linear CS measurement setting $y = \Phi x$ but did not consider measurement noise [14]. Inspired by Donoho et al., we estimate x from noisy measurements y using the empirical entropy as a proxy for the KS.

III. MAP Estimation and Discretization

In this section, we assume for exposition 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 a technical condition on the input.

Condition 1: We require that the probability density has bounded derivatives

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

where $\frac{d}{dx_n}$ is the derivative with respect to (w.r.t.) entry n of x, $n \in \{1,\ldots,N\}$, and $\rho > 0$.

Let $\widetilde{x}_{MAP}$ be the quantization bin in $\mathcal{R}^N$ nearest to x_{MAP}. Condition 1 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 Condition 1 holds and the aspect ratio $\delta > 0$ in (2), and let the noise $z \in \mathbb{R}^M$ be i.i.d. zero-mean Gaussian. Then for all $\epsilon > 0$, the quantized $\widetilde{x}_{MAP}$ satisfies

$$\Psi_X(x_{MAP}) \le \Psi_X(\widetilde{x}_{MAP}) < \Psi_X(x_{MAP}) + N\epsilon$$

almost surely as $N \to \infty$.

Theorem 1 is proved in the Appendix; 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:

$$x_{MAP}(\mathcal{R}) \triangleq \arg\min_{w\in\mathcal{R}^N} \Psi_X(w). \tag{9}$$

From Theorem 1, we have

$$\Psi_X(x_{MAP}(\mathcal{R})) \le \Psi_X(\widetilde{x}_{MAP}) \le \Psi_X(x_{MAP}) + N\epsilon \tag{10}$$

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

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

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

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

$$p_X(w) \triangleq \frac{f_X(w)}{f_{\mathcal{R}}}. \tag{11}$$

We now have

$$\begin{aligned}&\min_{w\in\mathcal{R}^N}\left(-\ln(p_X(w)) + c_2\|y-\Phi w\|^2\right)\\ =\;& \Psi_X(x_{MAP}(\mathcal{R})) + \ln(f_{\mathcal{R}}).\end{aligned} \tag{12}$$

The additive constant $\ln(f_{\mathcal{R}})$ can be ignored during the MAP optimization over $\mathcal{R}^N$, so that (9) gives the MAP estimate of x over $\mathcal{R}^N$ due to (12).

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 $E_X[\|y-x_{KS}\|^2]$ 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 Kolmogorov-inspired estimator for CS over a quantized grid. Consider a universal prior p_U [10, 11] that might involve Kolmogorov complexity [15–17], e.g., $p_U(w) = 2^{-K(w)}$. 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

$$-\ln(p_U(w)) < -\ln(p_X(w)) + \epsilon N \tag{13}$$

for *all* $w \in \mathcal{R}^N$ and $N > N_0(X,\epsilon)$ [10, 11]. We optimize over an objective function (risk) that incorporates p_U:

$$\Psi_U(w) \triangleq -\ln(p_U(w)) + c_2\|y-\Phi w\|^2, \tag{14}$$

resulting in

$$x^U_{MAP} \triangleq \arg\min_{w\in\mathcal{R}^N} \Psi_U(w). \tag{15}$$

We now present a conjecture on the quality of the reconstruction x^U_{MAP}; experimental evidence to assess this claim is presented in Section VI.

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

$$E[(X - X^U_{MAP})^2] < 2E[(X - E[X|Y,\Phi])^2] + N\epsilon$$

for large N.

A limitation of our data-independent reproduction levels (3) is that $\mathcal{R}$ has infinite cardinality. One approach to circumvent this problem is to add a second technical condition that upper bounds $f_X(x)$ by an exponentially decaying function. Subject to this condition, there exists an integer $c_3 > 1$ such that a finite set of reproduction levels

$$\mathcal{R}_F \triangleq \left\{-\frac{c_3\gamma^2}{\gamma}, -\frac{c_3\gamma^2-1}{\gamma}, \ldots, \frac{c_3\gamma^2}{\gamma}\right\} \tag{16}$$

will quantize a broad range of values of x with the probability that any $|x_i| > c_3\gamma$ being sufficiently small. This finite quantization step reduces the complexity of the estimation problem from infinite to combinatorial.

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 [24]. Our approach is reminiscent of the framework by Weissman et al. and Yang et al. for lossy data compression [25–27].

A. Universal compressor

We propose a universal lossless compression formulation following the conventions of Weissman et al. [25, 26]. Our goal is to characterize $-\log(p_U(w))$, cf. (14). To do so, we use empirical entropy [28], which for stationary ergodic sources is identical to the per-symbol Kolmogorov complexity almost surely (cf. Section II-C).

To define the empirical entropy, let us first define the empirical symbol counts:

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

where q is the context depth [11, 29], $\beta \in \mathcal{R}_F$, $\alpha \in \mathcal{R}_F^q$, 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

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

and the order q conditional empirical entropy,

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

Allowing the context depth $q = o(\log(N))$ to grow slowly with N, various universal compression algorithms can achieve the empirical entropy $H_q(\cdot)$ asymptotically [11, 29]. 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 employ the MCMC approach [24], where 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. Pseudocode for our MCMC approach appears in Algorithm 1.

We define the energy $\varepsilon(w)$ in an analogous manner to $\Psi_U(w)$, using $H_q(w)$ as our universal coding length (14):

$$\varepsilon(w) \triangleq NH_q(w) + c_4\|y - \Phi w\|^2, \quad (20)$$

where $c_4 = c_2 \log_2(e)$. The minimization of energy by MCMC is analogous to minimizing $\Psi_U(w)$. The Boltzmann PMF is then defined as

$$p_s(w) \triangleq \frac{1}{\zeta_s} \exp(-s\varepsilon(w)), \quad (21)$$

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

$$x_{MAP}^U \triangleq \arg\min_{w \in \mathcal{R}_F^N} \varepsilon(w). \quad (22)$$

We use a stochastic *Markov chain Monte Carlo* (MCMC) relaxation [24] to approximate the globally minimum solution.

Algorithm 1 MCMC for Universal CS

Inputs: Initial point $x^* \in \mathbb{R}^n$, $\mathcal{R}_F$, σ_Z^2, r, c
Outputs: Approximation w of x_{MAP}^U
Initialize w by quantizing x^* to $\mathcal{R}_F^N$
Initialize $n_q(w,\alpha)[\beta],\ \forall\ \alpha \in \mathcal{R}_F^q,\ \beta \in \mathcal{R}_F$
for $t = 1$ to r **do** *// super-iteration*
  $s \leftarrow c\log(t)$ for some $c > 0$
  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

During the minimization process, we refer to the approximation as w.

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:

$$\begin{aligned} & p_s(w_n = a | w^{\setminus n}) \\ = \ & \frac{\exp\left(-s\varepsilon(w_1^{n-1} a w_{n+1}^N)\right)}{\sum_b \exp\left(-s\varepsilon(w_1^{n-1} b w_{n+1}^N)\right)} \\ = \ & \frac{1}{\sum_b \exp\left(-s\left[N\Delta H_q(w,n,b,a) + c_4 \Delta d(w,n,b,a)\right]\right)}, \end{aligned} \quad (23)$$

where

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

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

$$\begin{aligned} & \Delta d(w,n,b,a) \\ \triangleq \ & \|y - \Phi(w_1^{n-1} b w_{n+1}^N)\|^2 - \|y - \Phi(w_1^{n-1} a w_{n+1}^N)\|^2 \end{aligned} \quad (24)$$

is the change in $\|y - \Phi w\|^2$ when $w_n = a$ is replaced by b.

At low temperatures, i.e., large s, a small difference in energy $\varepsilon(w)$ 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^U_{MAP}.

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 simulated annealing, in super-iteration t we use inverse temperature $s = c\log(t)$ [24, 25]. The constant c plays a crucial role. If c is large, then the Boltzmann distribution (21) favors low-energy sequences too greedily, and the algorithm might get stuck in local minima. On the other hand, we conjecture that there exists a universal constant c_5 such that for $c < c_5$ the algorithm approaches the global minimum in the limit of infinitely many iterations. To argue why w will tend toward minimum energy, observe that Algorithm 1 optimizes over $|\mathcal{R}_F|^n$ possible outputs. As long as $c < c_5$, there is a sufficiently large probability to transition between any two outputs, and Algorithm 1 will likely not get bogged down in a local minimum. Based on related work [24, 25], we conjecture that as we process more super-iterations t, w converges in distribution to the set of minimal energy solutions, which includes x^U_{MAP} (22) since large s favors low-energy w.

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 [25] recompute the empirical conditional entropy in $O(q|\mathcal{R}_F|)$ time only for the $O(q)$ contexts whose corresponding counts are modified [25]. The same approach can be used in Line 16, 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 $O(rMN)$.

D. 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 (estimated) $\widehat{x}$ well. For finite N, estimation performance using the reproduction levels (16) could suffer.

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 $u \in \mathcal{Z}^N$. 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

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

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

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

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. The resulting coding length is our universal prior; it approximates the Kolmogorov complexity $K(\mathcal{A}_{opt}(u))$.

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

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

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

$$\begin{aligned}
\frac{d\Upsilon(\mathcal{A})}{d\mathcal{A}(\beta)} &= -2\sum_{m=1}^{M} \left(y_m - \sum_{n=1}^{N} \Phi_{mn}\mathcal{A}(u_n)\right)\left(\sum_{n=1}^{N} \Phi_{mn}1_{\{u_n=\beta\}}\right)\\
&= 0,\ \forall\ \beta \in \mathcal{Z}. \qquad (25)
\end{aligned}$$

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

$$\mu_{m\beta} \triangleq \sum_{n\in\mathcal{L}_\beta} \Phi_{mn}, \qquad (26)$$

for each $m \in \{1, \ldots, M\}$ and $\beta \in \mathcal{Z}$. We desire the derivatives to be zero, cf. (25):

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

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

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

for each $\beta \in \mathcal{Z}$. The system of equations can be described in matrix form (28). Note that by writing μ as a matrix with entries indexed by row m and column β given by (26), 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

$$\mathcal{A}_{opt} = \Omega^{-1} \Theta$$

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

$$\|y - \Phi \mathcal{A}(u)\|^2 = \sum_{m=1}^{M} \big(y_m - \sum_{\beta} \mu_{m\beta} \mathcal{A}_{opt}(\beta) \big)^2, \qquad (27)$$

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

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 in each iteration of the inner loop.

- In Line 13, the differences in empirical conditional entropy can be computed in $O(q|\mathcal{Z}|)$ time as demonstrated by Jalali and Weissman [25].
- In Line 14, we update $\mu_{m\beta}$ for $m \in \{1, \ldots, M\}$ in $O(M)$ time.
- Line 15 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 16 requires to invert Ω in $O(|\mathcal{Z}|^3)$ time.
- Line 17 requires $O(M|\mathcal{Z}|)$ time, cf. (27).
- Line 18 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

1: **Inputs**: Initial point $x^* \in \mathbb{R}^n$, $\mathcal{Z}$, σ_Z^2, r, c
2: **Outputs**: Approximation $\mathcal{A}(u)$ of x_{MAP}^U
3: Initialize u and $\mathcal{A}$ from x^* and $\mathcal{Z}$
4: Initialize $n_q(u, \alpha)[\beta]$, $\forall\ \alpha \in \mathcal{Z}^q$, $\beta \in \mathcal{Z}$
5: Initialize $\mu_{m\beta}$, $\forall\ m \in \{1, \ldots, M\}$, $\beta \in \mathcal{Z}$
6: Initialize Ω
7: **for** $t = 1$ to r **do** *// super-iteration*
8: $s \leftarrow c \log(t)$ for some $c > 0$
9: Draw permutation $\{1, \ldots, N\}$ at random
10: **for** $t' = 1$ to N **do** *// iteration*
11: Let n be component t' in permutation
12: **for** all β in $\mathcal{Z}$ **do** *// possible new u_n*
13: Compute $\Delta H_q(u, n, \beta, u_n)$
14: Compute $\mu_{m\beta}, \forall\ m \in \{1, \ldots, M\}$
15: Update Ω
16: Compute $\mathcal{A}_{opt}$ *// invert Ω*
17: Compute $\|y - \Phi \mathcal{A}(u_1^{n-1} \beta u_{n+1}^N)\|^2$
18: Compute $p_s(u_n = \beta | u^{\setminus n})$
19: **end for**
20: $\widetilde{u}_n \leftarrow u_n$ *// save previous value*
21: Generate u_n using $p_s(\cdot | u^{\setminus n})$ *// Gibbs*
22: Update $n_q(\cdot)[\cdot]$ at $O(q)$ relevant locations
23: Update $\mu_{m\beta}$, $\forall\ m$, $\beta \in \{u_n, \widetilde{u}_n\}$
24: Update Ω *// $O(1)$ rows and columns*
25: **end for**
26: **end for**
27: return $\mathcal{A}(u)$

VI. Numerical results

We implemented universal MCMC estimation (Algorithm 2) in Matlab and tested it using several stationary ergodic sources. Our code is available for download at http://people.engr.ncsu.edu/dzbaron/software/UCS_BaronDuarte/.[2] We chose an alphabet of size $|\mathcal{Z}| = 7$ for all sources tested. For each source, inputs x of length $N = 10000$ 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; for each value of M and SNR, we average the performance of each algorithm over 25 draws of x, Φ, and z. Although MCMC was slower than the ℓ_1-norm minimizer used in our simulations [30], 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 fails due to the input not being canonically sparse.

Next we examined three sources for sparse signals whose sparse support (the locations of the nonzero entries) was generated by a two state Markov state machine (nonzero and zero valued 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 differed in the distribution of the nonzero values; all three sources were 3% sparse on average, and ℓ_1-norm minimization offered reasonable recovery performance.

For our first two-state Markov source, the nonzeros were set

[2]Shortly before publication, we noticed a mismatch between the constant c_4 in (20) and that used in our implementation. We expect to update the code online accordingly in the near future.

$$\overbrace{\begin{bmatrix} \sum_{m=1}^{M} \mu_{m\beta_1}\mu_{m\beta_1} & \cdots & \sum_{m=1}^{M} \mu_{m\beta_{|\mathcal{Z}|}}\mu_{m\beta_1} \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{M} \mu_{m\beta_1}\mu_{m\beta_{|\mathcal{Z}|}} & \cdots & \sum_{m=1}^{M} \mu_{m\beta_{|\mathcal{Z}|}}\mu_{m\beta_{|\mathcal{Z}|}} \end{bmatrix}}^{\Omega} \begin{bmatrix} \mathcal{A}(\beta_1) \\ \vdots \\ \mathcal{A}(\beta_{|\mathcal{Z}|}) \end{bmatrix} = \overbrace{\begin{bmatrix} \sum_{m=1}^{M} y_m\mu_{m\beta_1} \\ \vdots \\ \sum_{m=1}^{M} y_m\mu_{m\beta_{|\mathcal{Z}|}} \end{bmatrix}}^{\theta}. \quad (28)$$

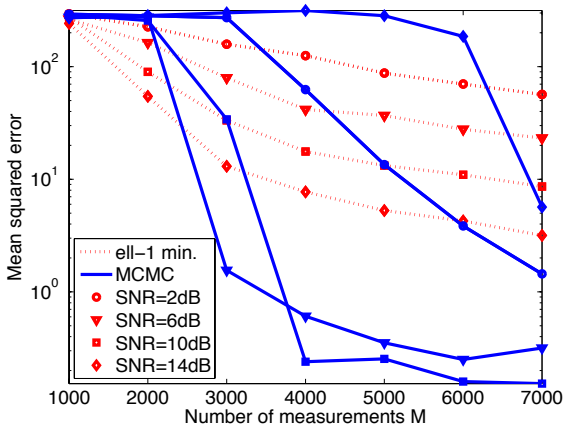

Fig. 2. *Universal MCMC and ℓ_1-norm minimization recovery results for a two-state Markov source with nonzero entries drawn from a Rademacher distribution as a function of the number of Gaussian random measurements M for different SNR values. MCMC outperforms ℓ_1-norm minimization for most cases shown, with the surprising exception of $SNR = 14dB$.*

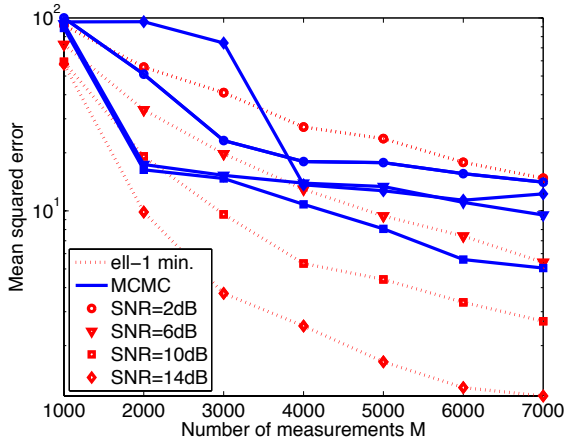

Fig. 3. *Universal MCMC and ℓ_1-norm minimization recovery results for a two-state Markov source with nonzero entries drawn from a uniform distribution $U[0,1]$ as a function of the number of Gaussian random measurements M for different SNR values. The performance of universal MCMC estimation is hampered likely due to the continuous-valued nature of the source studied.*

to a *constant* value +1; such signal structure has low entropy, and MCMC consistently outperformed ℓ_1 as expected; we do not include the results due to space limitations.

For our second two-state Markov source, the nonzeros were drawn from a *Rademacher* distribution, i.e., uniform over $\{-1,+1\}$ (Fig. 2). Here, universal MCMC estimation performed well for low to moderate SNR. Surprisingly, MCMC failed for high SNR. To study 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 is expected to appear for any particular input source 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. 3), MCMC performed 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 adaptive quantizer $\mathcal{A}$ 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%; else x_n was zero. For this source, MCMC outperforms ℓ_1-norm minimization, 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 [5]; similar computations were performed by Guo and Verdu [20]. Interestingly, the squared error achieved by MCMC is thrice the MMSE. One is left to wonder whether we could approach the mean squared error bound given in Conjecture 1 in the limit of infinitely many super-iterations.

Appendix. 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,

$$\begin{aligned} & \|y - \Phi\widetilde{x}_{MAP}\| \\ \leq\ & \|y - \Phi x_{MAP}\| + \|\Phi(x_{MAP} - \widetilde{x}_{MAP})\| \\ \leq\ & \|y - \Phi x_{MAP}\| + \|\Phi\|\|x_{MAP} - \widetilde{x}_{MAP}\| \\ \leq\ & \|y - \Phi x_{MAP}\| + \|\Phi\|\sqrt{N}\|x_{MAP} - \widetilde{x}_{MAP}\|_\infty \quad (29) \\ \leq\ & \|y - \Phi x_{MAP}\| + A\frac{\sqrt{N}}{2\lceil \ln(N)\rceil}, \quad (30) \end{aligned}$$

where the ℓ_∞ norm in (29) isolates the largest absolute difference between x_{MAP} and $\widetilde{x}_{MAP}$, which is $O(\log(N)^{-1})$ (3), and A is the spectral norm of Φ. For the distribution we assumed for the matrix Φ, $A = 1+\delta^{-0.5} = O(1)$ almost surely asymptotically [31], 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 finite. Substituting

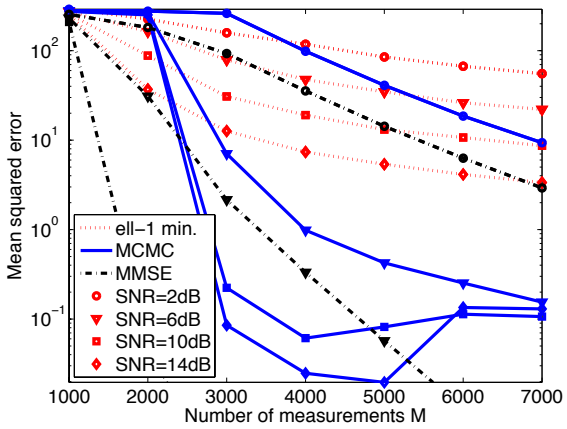

Fig. 4. *Universal MCMC and ℓ_1-norm minimization recovery results for a source with i.i.d. Bernoulli entries with nonzero probability of 3% as a function of the number of Gaussian random measurements M for different SNR values. MCMC outperforms ℓ_1-norm minimization for a sufficiently large number of measurements M.*

this observation into (30),

$$\begin{aligned} & \|y - \Phi \widetilde{x}_{MAP}\|^2 \\ \leq & \left(\|y - \Phi x_{MAP}\| + O\left(\frac{\sqrt{N}}{\log(N)} \right) \right)^2 \\ = & \|y - \Phi x_{MAP}\|^2 + O\left(\frac{N}{\log(N)} \right) \\ = & \|y - \Phi x_{MAP}\|^2 + o(N) \end{aligned} \tag{31}$$

almost surely asymptotically. That is, discretization doesn't 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),

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

Because the log derivatives of f_X are bounded, cf. (8),

$$\begin{aligned} & |\ln(f_X(\widetilde{x}_{MAP})) - \ln(f_X(x_{MAP}))| \\ < & \rho \|x_{MAP} - \widetilde{x}_{MAP}\|_1 = o(N). \end{aligned} \tag{32}$$

Combining (7), (31), and (32), we see that $\Psi_X(\widetilde{x}_{MAP}) - \Psi_X(x_{MAP}) = o(N)$ almost surely asymptotically. □

Acknowledgments

Preliminary conversations with Deanna Needell and Tsachy Weissman framed our thinking about universal compressed sensing. Phil Schniter provided unparalleled assistance in detailed conversations, feedback on our drafts, and probing questions. MFD was partially supported by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch.

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.*, 2011, To appear.

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

[11] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, Jul. 1984.

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

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

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

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

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

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

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

[19] D.L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[20] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983, 2005.

[21] 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.

[22] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, 2002.

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

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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