COMPLEXITY-ADAPTIVE UNIVERSAL SIGNAL ESTIMATION FOR COMPRESSED SENSING

Junan Zhu,[n] *Dror Baron,*[n] *and Marco F. Duarte*[m*]

[n] North Carolina State University, ECE Department, Raleigh, NC 27695
[m] University of Massachusetts, ECE Department, Amherst, MA 01003

ABSTRACT

We study the compressed sensing (CS) signal estimation problem where a signal is measured via a linear matrix multiplication under additive noise. While this setup usually assumes sparsity or compressibility in the signal during estimation, additional signal structure that can be leveraged is often not known *a priori*. For signals with independent and identically distributed (i.i.d.) entries, existing CS algorithms achieve optimal or near optimal estimation error without knowing the statistics of the signal. This paper addresses estimating *stationary ergodic* non-i.i.d. signals with unknown statistics. We have previously proposed a *universal* CS approach to simultaneously estimate the statistics of a stationary ergodic signal as well as the signal itself. This paper significantly improves on our previous work, especially for continuous-valued signals, by offering a four-stage algorithm called *Complexity-Adaptive Universal Signal Estimation* (CAUSE), where the alphabet size of the estimate adaptively matches the coding complexity of the signal. Numerical results show that the new approach offers comparable and in some cases, especially for non-i.i.d. signals, lower mean square error than the prior art, despite not knowing the signal statistics.

***Index Terms*—** MAP estimation, Markov chain Monte Carlo, non-i.i.d. signals, signal estimation, universal algorithms.

1. INTRODUCTION

Many systems in science and engineering are approximately linear, and 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:

$$y = \Phi x + z,$$

where Φ is an $M \times N$ matrix and $z \in \mathbb{R}^M$ denotes the noise. By posing a sparsity or compressibility requirement on the signal and using this requirement as a prior during signal estimation, compressed sensing (CS) has shown that it is indeed possible to accurately estimate x from y, Φ, and the statistics of z even when $M \ll N$ [1, 2].

For independent and identically distributed (i.i.d.) signals, existing CS algorithms [3] achieve optimal or near optimal estimation error without knowing the statistics of the signal. When addressing *stationary ergodic* non-i.i.d. signals, however, one might not be certain about the structure or statistics of the signal prior to estimation. 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 [4, 5] that assign a probability to a sequence despite not knowing its statistics.

In prior work [6], we minimized the empirical entropy, which approximates the Kolmogorov complexity [7][1] of the signal, with a regularization term corresponding to a log likelihood for the noise. The concrete formulation proposed in [6] used Markov chain Monte Carlo (MCMC) [9] to minimize the objective function. Our approach resembles maximum a posteriori (MAP) estimation with a universal prior used for the signal x. Preliminary results were promising, but the estimation quality for continuous-valued signals fell short of that observed for discrete-valued signals.

This paper significantly improves on our earlier work by introducing a four-stage algorithm called *Complexity-Adaptive Universal Signal Estimation* (CAUSE), which allows the alphabet size of the estimate to vary dynamically in order to match the coding complexity of the signal. Our numerical experiments show that our improved algorithm offers comparable and in some cases, especially for non-i.i.d. signals, lower mean square error than the prior art, despite not knowing the signal statistics *a priori*.

2. BACKGROUND

Setting: Consider the noisy measurement setup via a linear operator under additive noise $y = \Phi x + z$, where the signal $x \in \mathbb{R}^N$ is generated by a stationary ergodic source X, and must be estimated from y, Φ, and the statistics of z. Note that we specifically focus on non-i.i.d. sources, and *the distribu-*

*This work was supported in part by the National Science Foundation under Grant CCF-1217749 and in part by the U.S. Army Research Office under Grant W911NF-04-D-0003.

[1] A recent paper [8], developed independently from and appearing simultaneously with our prior work [6], also considered the performance of Kolmogorov complexity minimization for CS estimation. The paper [8] offered theoretical results but no algorithmic approach.

Algorithm 1 : Merging two nearest adjacent levels

procedure $[\mathbb{Q}, w] = \mathbf{Merge}(\mathbb{Q}, w)$
  ▷ Find nearest adjacent levels $\mathbb{Q}(j), \mathbb{Q}(j+1)$
  ▷ Create a new level $\mathbb{Q}(\nu) = \frac{\mathbb{Q}(j)+\mathbb{Q}(j+1)}{2}$
  if $w_n = j$ or $j+1, \forall n \in \{1, .., N\}$ **then**
    $w_n = \nu$
  end if
  ▷ Remove levels j and $j+1$

tion f_X that generates x is unknown. Also, whenever we refer to a signal, we mean a signal generated by a certain source.

Our goal is to estimate x despite our lack of knowledge about f_X. Thus, we must search for an estimation mechanism that is agnostic to the specific distribution f_X. For concrete analysis, we assume the additive noise $z \in \mathbb{R}^M$ to be i.i.d. Gaussian, with mean zero and variance σ_z^2. Other noise distributions are readily supported.

Universal MAP estimation: The MAP estimator for x that we proposed in [6] has the form

$$x_{MAP} \triangleq \arg\max_w f_X(w) f_{Y|X}(y|w) = \arg\min_w \Psi^X(w), \quad (1)$$

where $\Psi^X(w) \triangleq -\ln(f_X(w)) + \frac{1}{2\sigma_z^2}\|y - \Phi w\|_2^2$ denotes the objective function (risk) and $\|\cdot\|_2$ denotes the ℓ_2 norm; our ideal risk would be $\Psi^X(x_{MAP})$.

The above optimization problem (1) is defined over a continuous-valued x. However, in order to reduce complexity, we process a discretized space instead. Let $\mathcal{Q}$ be a discretizer from the real numbers to a finite set of representation levels (levels for short) $\mathbb{Q} \subset \mathbb{R}$ comprised of two parts: (*i*) $\mathcal{Q}_1 : \mathbb{R} \to \{1, \ldots, Z\}$, where Z is the *alphabet size* of $\mathbb{Q}$, assigns representation indices to the continuous-valued signal; and (*ii*) $\mathcal{Q}_2 : \{1, \ldots, Z\} \to \mathbb{Q}$ maps to continuous-valued levels. We further denote $\mathbb{Q}(j)$ to refer to the mapping $\mathcal{Q}_2$ between a group of representation indices $j \in \{1, \ldots, Z\}^N$ and the corresponding levels in order to denote its dependence on $\mathbb{Q}$. To keep the presentation simple, we assume that the levels are monotone increasing with the representation index, i.e., $\mathbb{Q}(1) < \mathbb{Q}(2) < ... < \mathbb{Q}(Z)$.

The continuous-valued optimization (1) becomes an optimization over a finite space due to discretization. Let $w = [w_1, \ \ldots \ , w_N] \in \{1, \ldots, Z\}^N$ be the sequence of indices used to represent a possible discretized signal $\mathbb{Q}(w) \in \mathbb{Q}^N$. In contrast to conventional MAP estimation with prior $p(w)$, we propose a universal lossless compression formulation following the conventions of Weissman and co-authors [10, 11]. Let $p_U(w) = 2^{-H_q(w)}$ be a universal prior where $H_q(w)$ is the q-depth conditional empirical entropy [5]. Our objective function becomes $\Psi^{H_q}(w) \triangleq N H_q(w) + c_1\|y - \Phi\mathbb{Q}(w)\|_2^2$, where c_1 is a constant derived from σ_z^2; $\Psi^{H_q}(w)$ offers a trade-off between low entropy/complexity and the residual $\|y - \Phi\mathbb{Q}(w)\|_2^2$. Under a mild assumption that x was generated by a stationary ergodic source with low entropy, we hope to estimate x with low ℓ_2 error.

Optimization via MCMC: We use a stochastic MCMC relaxation [9] to achieve the globally minimum solution in the limit of infinite computation. MCMC performs Gibbs sampling from the Boltzmann probability mass function,

$$p_s(w) \triangleq \frac{1}{\zeta_s} \exp(-s\Psi^{H_q}(w)), \quad (2)$$

where $s > 0$ is inversely related to temperature in simulated annealing and ζ_s is a normalization constant. In each *iteration*, the n-th element w_n is sampled from its marginal distribution, while the rest of w, $w^{\setminus n} \triangleq \{w_i : i \in \{1, \ldots, N\} \setminus \{n\}\}$, remains unchanged. We call the operation of sampling all elements in w a *super-iteration*.

Algorithm 2 : Adding one level between most distant adjacent levels

procedure $[\mathbb{Q}, w] = \mathbf{Add}(\mathbb{Q}, w)$
  ▷ Find most distant adjacent levels $\mathbb{Q}(j), \mathbb{Q}(j+1)$
  ▷ Create a new level $\mathbb{Q}(\nu) = \frac{\mathbb{Q}(j)+\mathbb{Q}(j+1)}{2}$
  ▷ $\Pr\Big(w_n = j | w^{\setminus n} \text{fixed}\Big) = p_s(w_n = j)$ according to (2)
  if $w_n = j$ or $j+1, \forall n \in \{1, ..., N\}$ **then**
    $p_{\mathrm{M}} = \begin{cases} \frac{p_s(w_n=j+1)}{p_s(w_n=j)+p_s(w_n=j+1)}, & \text{if } w_n = j \\ \frac{p_s(w_n=j)}{p_s(w_n=j)+p_s(w_n=j+1)}, & \text{if } w_n = j+1 \end{cases}$
    Set $w_n = \nu$ with probability p_{M}
  end if

Optimal discretizer: In our previous work [6], we allowed the discretizer to be adaptive to the data; in particular, we used the optimal discretizer for w given by $\mathbb{Q}^* = \arg\min_{\overline{\mathbb{Q}}} \|y - \Phi\overline{\mathbb{Q}}(w)\|_2^2$. Our earlier work discussed how to accelerate updates to $\mathbb{Q}^*$ when a single element of w is modified in each iteration. For brevity, we refer the reader to Baron and Duarte [6] for further details about our previous algorithm, including pseudo-code and software.

3. COMPLEXITY-ADAPTIVE UNIVERSAL SIGNAL ESTIMATION

While promising, our previous work [6] did not estimate continuous-valued signals to the same degree of accuracy as that of discrete-valued ones. For example, despite the continuous-valued nature of a Markov-uniform signal (cf. Section 4), the optimal discretizer $\mathbb{Q}^*$ spent multiple levels on zero-valued entries of the signal and only had one or two nonzero levels, which magnified the noise, slowed down the convergence of the algorithm, and increased the mean square error of the nonzero components of the signal.

Ideally, we want to employ as many levels as the run time allows for continuous-valued signals, while employing the same number of levels as the alphabet size for discrete signals. Inspired by this observation, we propose to first use a fixed number of levels, and then add or remove levels depending on whether either option further minimizes the objective function. Additionally, the performance on discrete-valued signals could benefit from adjusting the alphabet size to the same number of discrete values of the signal. With these observations in mind, we design the CAUSE algorithm, which has four stages.

Algorithm 3 : Complexity-Adaptive Universal Signal Estimation

1: **procedure** $[\widetilde{x}, \mathbb{Q}_{\mathrm{O}}, w_{\mathrm{O}}, C_{\mathrm{O}}] = \mathbf{CAUSE}(\Phi, y, \sigma_z^2, K, r)$
2: ▷ ***Input:*** Discretize $\widehat{x} = \Phi^T y$ by a uniform discretizer
3: $\mathbb{Q} = \left[-\frac{Z-1}{2} : \sqrt{\frac{2}{Z-1}} : \frac{Z-1}{2}\right]$, obtaining w
4: ▷ ***Stage 1:*** $[\mathbb{Q}_{1_{\mathrm{O}}}, w_{1_{\mathrm{O}}}, C_1]$=MCMC$(\mathbb{Q}, w, \sigma_z^2, r)$
5: Set $T = \left(\max \mathbb{Q}_{1_{\mathrm{O}}} - \min \mathbb{Q}_{1_{\mathrm{O}}}\right) / \left(K \times (Z-1)\right)$
6: ▷ ***Stage 2:***
7: **while** $\exists j \in \{1, ..., Z-1\}$ s.t. $|\mathbb{Q}(j) - \mathbb{Q}(j+1)| < T$ **do**
8: $[\mathbb{Q}_2, w_2]$=Merge$(\mathbb{Q}_{1_{\mathrm{O}}}, w_{1_{\mathrm{O}}})$
9: **end while**
10: $[\mathbb{Q}_{2_{\mathrm{O}}}, w_{2_{\mathrm{O}}}, C_2]$=MCMC$(\mathbb{Q}_2, w_2, \sigma_z^2, r)$
11: ▷ ***Stage 3:*** Run Trials 1 to 3
12: Trial 1: $[\mathbb{Q}_{3/1}, w_{3/1}]$=Merge$(\mathbb{Q}_{2_{\mathrm{O}}}, w_{2_{\mathrm{O}}})$
$[\mathbb{Q}_{3/1}, w_{3/1}, C_{3/1}]$=MCMC$(\mathbb{Q}_{3/1}, w_{3/1}, \sigma_z^2, r)$
13: Trial 2: $[\mathbb{Q}_{3/2}, w_{3/2}]$=Add$(\mathbb{Q}_{2_{\mathrm{O}}}, w_{2_{\mathrm{O}}})$
$[\mathbb{Q}_{3/2}, w_{3/2}, C_{3/2}]$=MCMC$(\mathbb{Q}_{3/2}, w_{3/2}, \sigma_z^2, r)$
14: Trial 3: $[\mathbb{Q}_{3/3}, w_{3/3}, C_{3/3}]$=MCMC$(\mathbb{Q}_{2_{\mathrm{O}}}, w_{2_{\mathrm{O}}}, \sigma_z^2, r)$
15: ▷ ***Stage 4:*** D = the Trial minimizing $C_{3/\cdot}$
16: Set $\mathbb{Q}_4 = \mathbb{Q}_{3/D}, w_4 = w_{3/D}, C_4 = C_{\mathrm{O}} = C_{3/D}$
17: **while** $C_4 \leq C_{\mathrm{O}}$ and run time does not expire **do**
18: Set $C_{\mathrm{O}} = C_4, \mathbb{Q}_{\mathrm{O}} = \mathbb{Q}_4, w_{\mathrm{O}} = w_4, \widetilde{x} = \mathbb{Q}_{\mathrm{O}}(w_{\mathrm{O}})$
19: **switch** D **do**
20: **case** 1: $[\mathbb{Q}_4, w_4]$=Merge$(\mathbb{Q}_4, w_4)$
21: **case** 2: $[\mathbb{Q}_4, w_4]$=Add$(\mathbb{Q}_4, w_4)$
22: **case** 3: exit CAUSE
23: **end switch**
24: $[\mathbb{Q}_4, w_4, C_4]$=MCMC$(\mathbb{Q}_4, w_4, \sigma_z^2, r)$
25: **end while**

Before describing CAUSE in detail, we begin by introducing two procedures in Algorithms 1 and 2 that are used in CAUSE (Algorithm 3). The variables in round parentheses are inputs of each algorithm, and variables in square parentheses are outputs.

The CAUSE algorithm is shown in Algorithm 3, where we denote the MCMC-based algorithm in our previous work [6] by $[\mathbb{Q}_{\mathrm{O}}, w_{\mathrm{O}}, C] = \text{MCMC}(\mathbb{Q}, w, \sigma_z^2, r)$, where w are the indices of levels representing the starting point of the algorithm, σ_z^2 is the noise variance, and r is the number of super iterations to run; MCMC outputs the optimized representation levels $\mathbb{Q}_{\mathrm{O}}$, the corresponding sequence w_{O}, and the risk C. Note that one can use other discretizers in Line 2, and the colon marks in Line 3 refer to the step size. One can also change the number of super-iterations r in MCMC$(\cdot)$; larger r improves estimation quality while requiring longer runtime. The parameter K allows CAUSE to determine how close two levels can be before being merged in Stage 2.

4. NUMERICAL RESULTS

We implemented CAUSE in Matlab and tested it using several signal sources. For each source, signals x of length $N = 10000$ were generated. Each such x was multiplied by a Gaussian random matrix Φ, where the number of measurements M varied between 2000 and 7000. We then added measurement noise z whose variance was selected to ensure that the signal to noise ratio (SNR) was 5 or 10 dB. We initialized CAUSE with $Z = 7$ levels and set the parameter $K = 10$. Additionally, we set $r = 70$ in Stage 1 and $r = 10$ in latter stages. These choices of parameters provide a trade-off between run time and estimation quality. Next, we compared the estimation performance of CAUSE to that of CoSaMP (a greedy solver) [12] and turboGAMP (a message-passing algorithm) [13]. We also ran GPSR [14], which is an optimization based approach; its results were close to those of CoSaMP and are not included here. For each value of M and SNR, we computed the mean signal-to-distortion ratio (SDR) of each algorithm over 25 draws of x, Φ, and z. For brevity, we specifically include results for only two stationary ergodic non-i.i.d. sources.

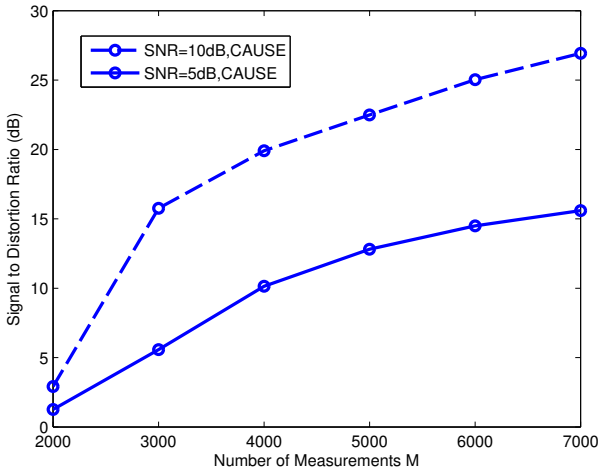

Fig. 1. *CAUSE estimation results for a Markov4 source as a function of the number of Gaussian random measurements M for different SNR values.*

Four-state Markov source (Markov4): To evaluate the performance of CAUSE for discrete-valued non-i.i.d. signals, we examined a four-state Markov source, featuring the pattern $+1,+1,-1,-1,+1,+1,-1,-1$,..., with 3% errors in state transitions, resulting in the Markov4 signal switching from -1 to $+1$ or vice-versa either too early or too late. The Markov4 signal is not sparse at all and we are not aware of any basis that can consistently transform it into a sparse representation. Because existing implementations of CS algorithms are not designed for the Markov4 signal and hence perform poorly (yielding SDR's below 5dB), we only include results for CAUSE in Fig. 1. CAUSE successfully estimated Markov4 with reasonable quality even when M was relatively small. It is worth noting that we also simulated a Bernoulli source with 3% nonzeros, which has the same entropy as the Markov4 source with 3% state transition errors. The performance of CAUSE and turboGAMP for this Bernoulli source are within 2dB of the minimum mean square error (MMSE) performance in the low SNR case, while CoSaMP lags behind in performance. Therefore, CAUSE succeeded in estimating the low-entropy Markov4 and Bernoulli signals by minimizing $\Psi^{H_q}(w)$, which promotes low complexity signals.

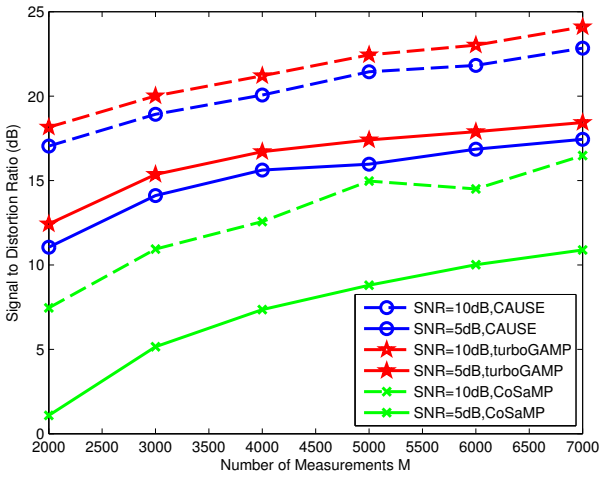

Fig. 2. *CAUSE, turboGAMP, and CoSaMP estimation results for a Markov-uniform source as a function of the number of Gaussian random measurements M for different SNR values.*

Markov-uniform source: We examined a source for sparse signals whose supports (the locations of the nonzero entries) are generated by a two-state Markov state machine (nonzero and zero states), and the nonzero values are uniformly distributed between 0 and 1. The transition from the zero to the nonzero state for adjacent entries has probability $\frac{3}{970}$; the transition from the nonzero to the zero state for adjacent entries has probability 10%. These transition probabilities yield 3% sparsity on average for the Markov-uniform signal. For turboGAMP, we use the MarkovChain1 (depth-1 Markov chain with one active state and one zero state) model to fit the support, Gaussian mixture model to fit the signal, and we provide the algorithm with the noise variance. CoSaMP is provided with Φ, y, and the sparsity rate; CAUSE is provided with Φ, y, and the noise variance.

Our previous work [6] performed poorly on this source because the optimal discretizer $\mathbb{Q}^*$ spent many levels for zero-valued entries of the signal, and only one or two levels for nonzeros. In contrast, CAUSE (*i*) merged multiple levels that coincided to zero in Stage 2; (*ii*) decided that it needed more nonzero levels in Stage 3; and (*iii*) kept adding levels in Stage 4, thus reducing the square error considerably. Fig. 2 shows that CAUSE achieves better performance than CoSaMP while still 1–2 dB below turboGAMP. This is an example of a source model that is well suited to the turboGAMP algorithm, which allows turboGAMP to outperform CAUSE. Nonetheless, turboGAMP requires the source distribution to be a good match to a parametric model fixed *a priori* to achieve such performance. This is in comparison to CAUSE, which effectively estimates the measured signal and its statistics simultaneously.

It is interesting to view the performance of CAUSE in light of a result by Donoho [15], who proved that MAP estimation with a Kolmogorov complexity prior can achieve twice the MMSE for a scalar channel. We leave the study of universal algorithms that achieve the MMSE for future work.

Acknowledgments

Preliminary conversations with Deanna Needell and Tsachy Weissman framed our thinking about universal CS. 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, Yanting Ma, and Nikhil Krishnan for commenting on our manuscript.

5. REFERENCES

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

[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] J. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

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

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

[6] D. Baron and M. F. Duarte, "Universal MAP estimation in compressed sensing," in *Proc. 49th Annual Allerton Conf. Comm., Control, Computing*, Sept. 2011, pp. 768–775.

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

[8] S. Jalali and A. Maleki, "Minimum complexity pursuit," in *Proc. 49th Annual Allerton Conf. Commun., Control, Computing*, Sept. 2011, pp. 1764–1770.

[9] 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.

[10] S. Jalali and T. Weissman, "Block and sliding-block lossy compression via MCMC," *IEEE Trans. Comm.*, vol. 60, no. 8, pp. 2187–2198, Aug. 2012.

[11] D. Baron and T. Weissman, "An MCMC approach to universal lossy compression of analog sources," *IEEE Trans. Sig. Process.*, vol. 60, pp. 5230–5240, Oct. 2012.

[12] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harm. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[13] J. Ziniel, S. Rangan, and P. Schniter, "A generalized framework for learning and recovery of structured sparse signals," in *Proc. IEEE Stat. Sig. Proc. Workshop (SSP)*, Aug. 2012, pp. 325–328.

[14] M. Figueiredo, R. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Top. Sign. Proces.*, vol. 1, pp. 586–597, Dec. 2007.

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