Sparse Regression Codes: Recent Results and Future Directions

(Invited Paper)

Ramji Venkataramanan
University of Cambridge
ramji.v@eng.cam.ac.uk

Sekhar Tatikonda
Yale University
sekhar.tatikonda@yale.edu

***Abstract*—Sparse Superposition or Sparse Regression codes were recently introduced by Barron and Joseph for communication over the AWGN channel. The code is defined in terms of a design matrix; codewords are linear combinations of subsets of columns of the matrix. These codes achieve the AWGN channel capacity with computationally feasible decoding. We have shown that they also achieve the optimal rate-distortion function for Gaussian sources. Further, the sparse regression codebook has a partitioned structure that facilitates random binning and superposition. In this paper, we review existing results concerning Sparse Regression codes and discuss directions for future research.**

I. Introduction

Developing computationally efficient codes that approach the Shannon-theoretic limits of communication and compression has long been one of the important goals of information theory. Starting with the introduction of turbo and LDPC codes [1] in the '90s, there have been significant advances towards this goal. Polar codes were the first codes with feasible encoding and decoding that were shown to provably attain the information-theoretic limit for discrete-alphabet symmetric sources and channels [2]–[4]. Spatially coupled ensembles have recently been shown to achieve the capacity of binary-input symmetric-output channels with belief propagation decoding [5].

There are many communication settings where the source or channel alphabet is inherently continuous, notably AWGN channels and Gaussian sources. In this paper, we discuss a new class of codes called Sparse Superposition Codes or Sparse Regression Codes (SPARC) that approach the Shannon limits for these problems with computationally efficient encoding and decoding. These codes were introduced by Barron and Joseph [6]–[8] for communication over the AWGN channel. In [8], it was shown that SPARCs achieve the AWGN capacity with feasible decoding. SPARCs for lossy compression were first considered in [9]. In [10], [11], we showed that SPARCs also attain the optimal rate-distortion function of Gaussian sources with feasible algorithms. Further, we showed in [12] that the source and channel coding modules can be combined to implement superposition and binning, which are key ingredients of several multi-terminal source and channel coding problems. Thus SPARCs offer a promising framework to build fast, rate-optimal codes for a variety of problems in network information theory.

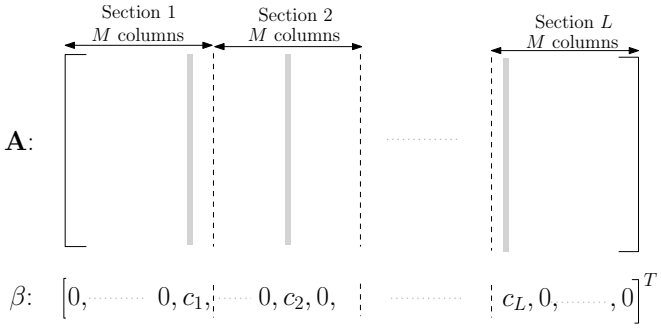

Fig. 1. $\mathbf{A}$ is an $n \times ML$ matrix and β is a $ML \times 1$ vector. The positions of the non-zeros in β correspond to the gray columns of $\mathbf{A}$ which combine to form the codeword $\mathbf{A}\beta$.

The goal of this paper is to give an overview of these results, and discuss directions for future research, including questions to be addressed in order for these codes to be used in practical applications.

Notation: Upper-case letters are used to denote random variables, lower-case for their realizations, and bold-face letters to denote random vectors and matrices. All vectors have length n. $\|\mathbf{X}\|$ denotes the ℓ_2-norm of vector $\mathbf{X}$, and $|\mathbf{X}| = \|\mathbf{X}\|/\sqrt{n}$ is the normalized version. We use natural logarithms, so entropy is measured in nats. $\mathcal{N}(\mu, \sigma^2)$ refers to the Gaussian distribution with mean μ and variance σ^2. To limit the number of symbols introduced, we reuse notation across sections. The model description at the beginning of each section explains all the variables used in it.

II. The Sparse Regression Codebook

A sparse regression code is defined in terms of a design matrix $\mathbf{A}$ of dimension $n \times ML$ whose entries are i.i.d. $\mathcal{N}(0, 1)$. Here n is the block length and M and L are integers whose values will be specified shortly. As shown in Fig. 1, one can think of the matrix $\mathbf{A}$ as composed of L sections with M columns each. Each codeword is a linear combination of L columns, with one column from each section. Formally, a codeword can be expressed as $\mathbf{A}\beta$, where β is a $ML \times 1$ vector $(\beta_1, \ldots, \beta_{ML})$ with the following property: there is exactly one non-zero β_i for $i \in \{1, \ldots, M\}$, one non-zero β_j for $j \in \{M+1, \ldots, 2M\}$, and so forth. The non-zero value of β in section i is set to c_i. The values of $\{c_i\}_{i=1}^{L}$ are

fixed a priori depending on the problem at hand. Denote the set of all β's that satisfy this property by $\mathcal{B}_{M,L}$.

Since there are M columns in each of the L sections, the total number of codewords is M^L. To obtain a compression rate of R nats/sample, we need

$$M^L = e^{nR} \quad \text{or} \quad L \log M = nR \tag{1}$$

There are several choices for the pair (M, L) which satisfy this. For example, $L = 1$ and $M = e^{nR}$ recovers the Shannon-style random codebook; here the number of columns in the dictionary $\mathbf{A}$ is e^{nR}, i.e., exponential in n. For our constructions, we choose $M = L^b$ for some $b > 1$ so that (1) implies

$$L \log L = nR/b. \tag{2}$$

Thus L is now $\Theta(n/\log n)$, and the number of columns ML in the dictionary $\mathbf{A}$ is now $\Theta\left((n/\log n)^{b+1}\right)$, a *polynomial* in n. This reduction in dictionary complexity can be harnessed to develop efficient encoders and decoders for the SPARC.

Since each codeword in a SPARC is a linear combination of L columns of $\mathbf{A}$ (one from each section), codewords sharing one or more common columns in the sum will be dependent. Also, SPARCs are not linear codes since the sum of two codewords does not equal another codeword in general.

III. Point-to-point Channel Coding

Sparse regression codes were first introduced for communication over the AWGN channel. Consider a channel with input X and output Y defined by

$$Y = X + Z$$

where $Z \sim \mathcal{N}(0, N)$ is a noise variable independent of X. There is an average power constraint P on the input X. Denote the signal-to-noise ratio P/N by by v.

Each message in the set $\{1, \ldots, M^L = e^{nR}\}$ is indexed by a unique $\beta \in \mathcal{B}_{M,L}$. To transmit the message corresponding to β, the encoder produces the channel input $\mathbf{X} = \mathbf{A}\beta$. To design an efficient decoder, [7], [8] considered the following choice of $\{c_i\}_{i=1}^L$:

$$c_i = \left[\sigma^2 e^{2\mathcal{C}} (e^{\frac{2\mathcal{C}}{L}} - 1) \cdot e^{-\frac{2\mathcal{C}i}{L}}\right]^{1/2}, \qquad i = 1, \ldots, L \tag{3}$$

where

$$\mathcal{C} = \tfrac{1}{2} \log(1 + v)$$

is the channel capacity. This choice of coefficients corresponds to treating each section as a code of rate $\frac{\mathcal{C}}{L}$, with the total power P divided among the L sections so that they can be successively decoded (in each step, treating the codeword contributions from undecoded sections as noise). However, successive decoding will perform poorly since the number of sections is large (recall that $L = \Theta(n/\log n)$). An error in decoding one section affects the decoding of future sections, leading to a large number of section errors after L steps.

The decoder proposed in [8] controls the accumulation of section errors by performing *adaptive* successive decoding. The main ingredients of this algorithm are described below. In the first step, the decoder computes the inner product of each column with the received sequence $\mathbf{Y}/\|\mathbf{Y}\|$, and picks those columns for which this test statistic exceeds a pre-specified threshold τ. In the second step, a residue $\mathbf{R}_1$ is formed by subtracting the contribution of the selected columns from $\mathbf{Y}$. The inner product of $\mathbf{R}_1/\|\mathbf{R}_1\|$ with each column from the undecoded sections gives the test statistics for the second step. The decoder again selects the columns for which this statistic crosses the threshold. The algorithm continues in this fashion until at most a pre-specified number of steps, arranged to be of the order of $\log M$. The algorithm terminates earlier if at least one column has been selected from each section, or the test-statistics in any step are all below the threshold.

Ideally, the decoder selects the column from each section that is part of the transmitted codeword. For a particular section, there are three possible ways an error could occur when the algorithm is completed. The first is by selecting exactly one column in that section, but the wrong one; the second case is when two or more terms are selected, and the third is when no term is selected. The section error rate is the fraction of sections in which errors occur.

The analysis of the algorithm is challenging due to the dependence between the test statistics used in each step. A variation of the above algorithm which is more tractable is analyzed in [7]. The essence of the result stated below states that rates up to

$$\mathcal{C}^* := \frac{\mathcal{C}}{1 + \delta_M} \tag{4}$$

can be achieved with $O(\delta_M)$ fraction of section errors, where

$$\delta_M = \frac{1}{\sqrt{\pi \log M}}. \tag{5}$$

Theorem 1. *[7] Let the rate $R < \mathcal{C}^*$ be expressed in the form*

$$\frac{\mathcal{C}^*}{1 + \frac{\kappa}{\log M}} \tag{6}$$

with $\kappa > 0$. Then with probability at least $1 - P_e$ the adaptive successive decoder has section error rate less than

$$\delta_{err} := \frac{\delta_M}{2\mathcal{C}} + \frac{3\kappa + 5}{8\mathcal{C} \log M} \tag{7}$$

where

$$P_e = \kappa_1 e^{-\kappa_2 L \min\{\kappa_3 \Delta^2,\ \kappa_4 \Delta\}}. \tag{8}$$

In (8)*, $\Delta = \frac{\mathcal{C}^* - R}{\mathcal{C}^*}$, κ_1 is a polynomial in M, and κ_2, κ_3 and κ_4 are constants which depend on the snr v.*

Remark: When combined with an outer Reed-Solomon code of rate $(1 - 2\delta_{err})$, the decoder achieves a block error probability of P_e with the overall rate being $(1 - \delta_{err})R$.

Theorem 1 tells us that the adaptive successive decoder can achieve rates of the order of $1/\sqrt{\log M}$ below capacity. This gap is due to the decoding algorithm rather than the structure of the sparse regression ensemble. Indeed, it was shown in [6] that with optimal (maximum-likelihood) decoding, SPARC achieve an error exponent of the same order as the random coding exponent [13].

It is shown in [14] that gap from capacity can be improved to $O(\log\log M/\log M)$ using a power allocation slightly modified from (3). The coefficients c_i^2 are now chosen proportional to

$$\max\{e^{-\frac{2\mathcal{C}i}{L}},\ e^{-2\mathcal{C}}(1+\tfrac{\kappa}{\sqrt{2\log M}})\}$$

which slightly boosts the power of the coefficients c_i for i close to L. This helps ensure that, even towards the end of the algorithm, there will be sections for which the true terms are expected to have inner product above threshold.

IV. Point-to-Point Source Coding

Consider an ergodic source X with mean 0 and variance σ^2, to be compressed at rate R. The quality of reconstruction is measured through the mean-squared distortion criterion

$$d_n(\mathbf{X},\hat{\mathbf{X}}) = |\mathbf{X}-\hat{\mathbf{X}}|^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{X}_i)^2.$$

It was shown in [11] that any ergodic source can be compressed to the Gaussian distortion-rate function

$$D^*(R) = \sigma^2 e^{-2R}$$

using a SPARC with a simple successive approximation encoder. The scheme is briefly described below.

The non-zero coefficients in each section are chosen to be

$$c_i = \sqrt{\frac{2R\sigma^2}{L}\left(1-\frac{2R}{L}\right)^{i-1}}, \quad i=1,\ldots,L. \tag{9}$$

Given a source sequence $\mathbf{X}$, the encoder determines the codeword as follows. Set the initial 'residue' $\mathbf{R}_0 = \mathbf{X}$. In Step $i,\ i=1,\ldots,L$, pick

$$m_i = \underset{j:\ (i-1)M+1\le\, j\,\le iM}{\operatorname{argmax}} \left\langle \mathbf{A}_j,\ \frac{\mathbf{R}_{i-1}}{\|\mathbf{R}_{i-1}\|}\right\rangle \tag{10}$$

and set

$$\mathbf{R}_i = \mathbf{R}_{i-1} - c_i\mathbf{A}_{m_i}. \tag{11}$$

The codeword $\hat{\beta}$ has non-zero values in positions $m_i,\ 1\le i\le L$.

The encoder described above may be interpreted in terms of successive refinement [15]. We can think of each section of the design matrix $\mathbf{A}$ as a codebook of rate R/L. For step $i,\ i=1,\ldots,L$, the residue $\mathbf{R}_{i-1}$ acts as the 'source' sequence and the variance of the new residue $\mathbf{R}_i$ is the resulting distortion. The minimum mean-squared distortion with a Gaussian codebook [16] at rate R/L is

$$D_i^* = |R_{i-1}|^2\exp(-2R/L) \approx |R_{i-1}|^2(1-2R/L) \tag{12}$$

for $R/L \ll 1$. The typical value of the distortion in Section i is close to D_i^* since the algorithm is equivalent to maximum-likelihood encoding within each section. Since the rate R/L is infinitesimal, the deviations from D_i^* in each section can be large. However, the final distortion $|\mathbf{R}_L^2|$ is close to the typical value σ^2e^{-2R} with excess distortion probability that falls *exponentially* in L. The result is stated below.

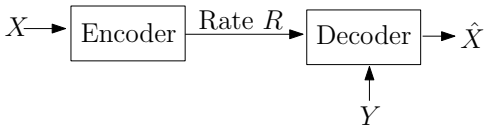

Fig. 2. Compressing X with decoder side-information Y

Theorem 2. *[11] Consider a length n source sequence $\mathbf{S}$ generated by an ergodic source having mean 0 and variance σ^2. Let $\delta_0,\delta_1,\delta_2$ be any positive constants such that*

$$\Delta \triangleq \delta_0 + 5R(\delta_1+\delta_2) < 0.5. \tag{13}$$

The SPARC with the proposed encoding algorithm produces a codeword $\mathbf{A}\hat{\beta}$ that satisfies

$$P\left(\ |\mathbf{S}-\mathbf{A}\hat{\beta}|^2\ >\ \sigma^2e^{-2R}(1+e^R\Delta)^2\right) < p_0+p_1+p_2 \tag{14}$$

for sufficiently large M,L, where

$$p_0 = P\left(\ \left|\frac{|\mathbf{S}|}{\sigma}-1\right| > \delta_0\ \right),\quad p_1 = 2ML\exp\left(-n\delta_1^2/8\right),$$
$$p_2 = \left(\frac{M^{2\delta_2}}{8\log M}\right)^{-L}. \tag{15}$$

Gap from $D^*(R)$: To achieve distortions close to the Gaussian $D^*(R)$ with high probability, we need p_0,p_1,p_2 to all go to 0. In particular, for $p_2\to 0$ with growing L, from (15) we require that

$$\delta_2 > \frac{\log(8\log M)}{2\log M}. \tag{16}$$

To approach $D^*(R)$, note that we need n,L,M to all go to ∞ while satisfying (1). When n,L,M are sufficiently large, (16) dictates how small Δ can be: the distortion is of order $\frac{\log\log M}{\log M}$ higher than the optimal value $D^*(R)=\sigma^2e^{-2R}$.

We note that this gap from $D^*(R)$ is of a smaller order than the gap from capacity in channel coding (given in Theorem 1), despite the adaptive successive channel decoding being computationally more complex than successive approximation. This is due to the fact that in channel coding there is exactly one transmitted codeword that has to be recovered, so the number of section errors needs to be kept small. In contrast, for compression there may be a number of codewords that are good representations of the source sequence, and the successive approximation encoder finds one of them. This encoder is not optimal, and finding better encoders with reasonable computational complexity is an important direction for future work.

V. Binning with SPARCs

In this section, we describe how random binning can be implemented using SPARCs. Binning is a key ingredient of several multi-terminal source and channel coding problems [17]. We explain the main ideas in the context of lossy compression with decoder side-information. The rate-distortion limit for this problem was obtained by Wyner and Ziv [18].

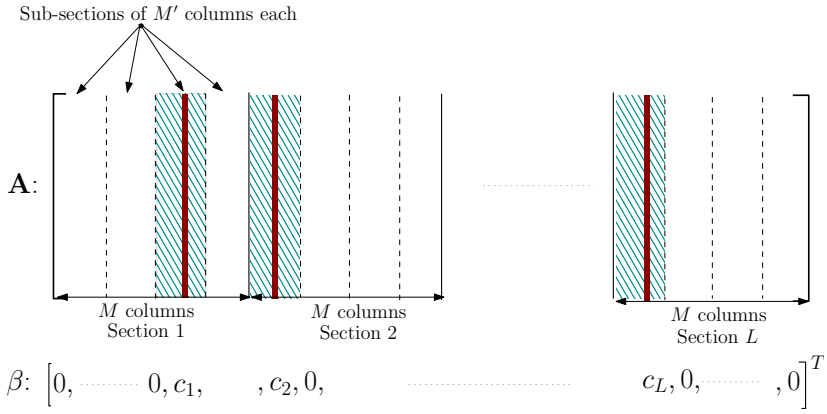

Fig. 3. Each section is divided into subsections of M' columns. A bin is formed by specifying a subsection in each of the L sections as shown by the shaded regions.

The model is depicted in Fig. 2. Consider an i.i.d Gaussian source $X \sim \mathcal{N}(0, \sigma^2)$ to be compressed with mean-squared distortion D. The decoder side-information Y is noisy version of X and is related to X by $Y = X + Z$, where $Z \sim \mathcal{N}(0, N)$ is independent of X. The sequence $\mathbf{Y}$ is available at the decoder non-causally. If $\mathbf{Y}$ were available at the encoder as well, the optimal strategy is to compress $\mathbf{Z} = \mathbf{Y} - \mathbf{X}$ to within distortion D; the minimum rate required for this is $\frac{1}{2}\log\frac{\text{Var}(X|Y)}{D}$ nats/sample.

Wyner and Ziv showed [18] that this rate is achievable even when $\mathbf{Y}$ is available at only the decoder. The coding scheme consists of a high-rate compression codebook partitioned into bins, each of which serves as a lower-rate channel code. The codebook is defined over an auxiliary random variable U jointly distributed with X according to

$$U = X + V \tag{17}$$

where $V \sim \mathcal{N}(0, Q)$ is independent of X. (17) can be equivalently written in terms of the reverse test channel as

$$X = aU + V' \tag{18}$$

where $a = \frac{\sigma^2}{\sigma^2+Q}$, and $V' \sim \mathcal{N}(0, \frac{\sigma^2 Q}{\sigma^2+Q})$ is independent of U. The first step in the coding scheme is to find a codeword $\mathbf{U}$ such that $a\mathbf{U} - \mathbf{X}$ has normalized squared norm less than $\frac{\sigma^2 Q}{\sigma^2+Q}$. From Theorem 2, we know that a SPARC can be used to perform this quantization if the codebook size satisfies

$$M^L > \exp(nI(X;U)). \tag{19}$$

The codebook is partitioned into bins by dividing each section of the design matrix $\mathbf{A}$ into subsections of M' columns each as shown in Figure 3. The encoder only indicates to the decoder which subsection in each of the L sections of β contains a non-zero. More precisely, it sends a tuple $(p_1, \ldots, p_L)$ where $p_i \in \{1, \ldots, \frac{M}{M'}\}$ indicates a subsection in the ith section of $\mathbf{A}$. Thus a bin is a subset of the codebook consisting of codewords corresponding to β's with ones in the sections specified by $(p_1, \ldots, p_L)$. There are $(M/M')^L$ bins and the rate R required to send the bin index to the decoder is determined as

$$e^{nR} = (M/M')^L. \tag{20}$$

The decoder's task is to recover the codeword $\mathbf{U}$ using the bin index and the side-information $\mathbf{Y}$. This is equivalent to a *channel decoding* problem with the side-information Y related to U as

$$Y = X + Z = aU + V' + Z \tag{21}$$

where U, V' and Z are mutually independent.

Observe that each bin is itself a smaller sparse regression codebook with M'^L codewords, defined by a $n \times M'L$ sub-matrix of $\mathbf{A}$. Since Theorem 1 shows that SPARCs can achieve the AWGN channel capacity, the decoder can perfectly recover $\mathbf{U}$ if the number of codewords in each bin satisfies

$$M'^L < \exp\big(n\ \tfrac{1}{2}\log(1 + \mathsf{snr})\big) \tag{22}$$

where

$$\mathsf{snr} = \frac{a^2\text{Var}(U)}{\text{Var}(V') + \text{Var}(Z)} = \frac{\sigma^4}{\sigma^2 Q + (\sigma^2+Q)N}. \tag{23}$$

Combining (19), (20) and (22), we conclude that a distortion D can be achieved when the number of bins e^{nR} satisfies

$$R > I(U;X) - I(U;Y) = \frac{1}{2}\log\frac{\text{Var}(X|Y)}{D}$$

where the last inequality is obtained by setting $Q = \frac{\text{Var}(X|Y)D}{\text{Var}(X|Y)-D}$. After decoding $\mathbf{U}$, the decoder reconstructs $\hat{\mathbf{X}}$ as the MMSE estimate of $\mathbf{X}$ given $(\mathbf{U}, \mathbf{Y})$. It can be verified that the expected squared-error distortion is D.

Using the binning strategy described above, it can also be shown that SPARCs achieve the capacity of the AWGN channel with state known non-causally at the encoder. This is the channel coding dual of the Gaussian Wyner-Ziv problem, often referred to as 'writing on dirty paper' [19]. The optimal coding scheme involves partitioning a high-rate channel code into bins, each of which is a good compression code.

VI. Future Directions

Improved coding algorithms: The channel decoder and source encoder described in Sections III and IV have relatively slow convergence to the Shannon limit – the gap is $O\left(\frac{\log\log n}{\log n}\right)$. Developing feasible channel coding decoders and compression encoders with faster convergence to the optimal rates is an important direction for further work.

The results in [6], [10] show that SPARCs under optimal (minimum-distance) decoding/encoding are essentially as good as random coding ensembles, both in terms of rate and error exponent performance. This suggests that it is possible to design efficient decoders/encoders with smaller gap from the Shannon limits. From finite block-length results, e.g. [20], the best possible gap is of order $1/\sqrt{n}$.

Barron and Cho [21] recently proposed a channel decoder that uses soft decisions based on Bayes optimal statistics in each iteration, followed by thresholding only at the final step. Their analysis and numerical simulations indicate substantial performance improvement over the original adaptive successive decoder. A complete analysis of the soft-decision decoder is an important outstanding problem. Message-passing ideas inspired by turbo and LDPC decoders may also prove useful in constructing better decoders.

Coupled with the codebook structure that makes it simple to implement binning and superposition, improved algorithms for channel decoding and source encoding will pave the way for fast, rate-optimal SPARCs for network problems such as interference channels, broadcast channels, distributed source coding and multiple descriptions.

Compression performance with optimal encoding: The rate-distortion performance of SPARCs ensemble under optimal (minimum-distance) encoding was analyzed in [10] where it was shown that the code achieves the Shannon rate-distortion function with the optimal error exponent for all distortions $D < \frac{\sigma^2}{4.91}$. The exact rate-distortion performance of the minimum-distance encoder outside this region is an open question. The performance of the successive approximation encoder leads us to believe that the result in [10] can be extended to all distortions. Recent advances in measure concentration for dependent random variables may prove useful in resolving this question. The analysis of the optimal encoder is equivalent to analyzing the large-deviation behavior of a sum of Bernoulli random variables with a specific dependence structure. A complete characterization of SPARC compression with optimal encoding will also yield the SPARC error exponent for the Gaussian Wyner-Ziv and Gelfand-Pinsker problems.

Low-complexity dictionaries: Choosing the entries of the dictionary $\mathbf{A}$ to be i.i.d $\mathcal{N}(0,1)$ leads to high storage complexity. A simpler alternative is to construct the dictionary with binary entries picked uniformly at random from $\{+1,-1\}$. The performance of such a dictionary under maximum-likelihood decoding was recently analyzed in [22]. Empirically, ± 1 dictionaries appear to have the same performance as Gaussian dictionaries when used with the encoding and decoding algorithms described in Sections III and IV. Formally establishing such a result remains an open problem.

Extension to Discrete Alphabets: The basic idea behind the sparse regression ensemble is to define each codeword as the sum of L columns of a dictionary, where $L \sim \frac{n}{\log n}$. In principle, this idea can be applied to channels/sources with discrete alphabets where addition is defined over the appropriate finite field. Designing efficient encoders and decoders is a challenging problem because the current techniques (especially for channel coding) rely on properties specific to Gaussian sequences.

References

[1] T. Richardson and R. L. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[2] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3051 –3073, July 2009.

[3] S. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1751 –1768, April 2010.

[4] S. Korada and R. Urbanke, "Polar codes for Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker," in *IEEE Information Theory Workshop*, Jan. 2010.

[5] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," 2012. http://arxiv.org/abs/1201.2999.

[6] A. Barron and A. Joseph, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *IEEE Trans. on Inf. Theory*, vol. 58, pp. 2541–2557, Feb 2012.

[7] A. Barron and A. Joseph, "Toward fast reliable communication at rates near capacity with Gaussian noise," in *Proc. 2010 IEEE ISIT*.

[8] A. Joseph and A. Barron, "Fast sparse superposition codes have exponentially small error probability for $R < C$," *Submitted to IEEE Trans. Inf. Theory*, 2012. http://arxiv.org/abs/1207.2406.

[9] I. Kontoyiannis, K. Rad, and S. Gitzenis, "Sparse superposition codes for Gaussian vector quantization," in *2010 IEEE Inf. Theory Workshop*, p. 1, Jan. 2010.

[10] R. Venkataramanan, A. Joseph, and S. Tatikonda, "Gaussian rate-distortion via sparse linear regression over compact dictionaries," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*. http://arxiv.org/abs/1202.0840.

[11] R. Venkataramanan, T. Sarkar, and S. Tatikonda, "Lossy compression via sparse linear regression: Computationally efficient encoding and decoding," in *Proc. 2013 IEEE Int. Symp. Inf. Theory*. http://arxiv.org/abs/1212.1707.

[12] R. Venkataramanan and S. Tatikonda, "Sparse regression codes for multi-terminal source and channel coding," in *50th Allerton Conf. on Commun., Control, and Computing*, 2012.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.

[14] A. Joseph, *Achieving information-theoretic limits with high-dimensional regression.* PhD thesis, Department of Statistics, Yale University, 2012.

[15] W. Equitz and T. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, pp. 269 –275, Mar 1991.

[16] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inf. Theory*, vol. 43, pp. 38 –47, Jan 1997.

[17] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2012.

[18] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1 – 10, January 1976.

[19] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, pp. 439 – 441, May 1983.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[21] A. R. Barron and S. Cho, "High-rate sparse superposition codes with iteratively optimal estimates," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*.

[22] Y. Takeishi, M. Kawakita, and J. Takeuchi, "Least squares superposition codes with bernoulli dictionary are still reliable at rates up to capacity," in *Proc. 2013 IEEE Int. Symp. Inf. Theory*.