

Optimal Phase Transitions in Compressed Sensing with Noisy Measurements

Yihong Wu and Sergio Verdú

Abstract—Compressed sensing deals with efficient recovery of analog signals from linear encodings. This paper presents a statistical study of compressed sensing by modeling the input signal as an i.i.d. random process. Three classes of encoders are considered, namely, optimal nonlinear, optimal linear and random linear encoders. Focusing on optimal decoders, we investigate the fundamental tradeoff between measurement rate and reconstruction fidelity gauged by the noise sensitivity. The optimal phase-transition threshold is determined as a functional of the input distribution and compared to suboptimal thresholds achieved by popular reconstruction algorithms. In particular, we show that Gaussian sensing matrices incur no penalty on the phase-transition threshold with respect to optimal nonlinear encoding. Our results also provide a rigorous justification of previous results based on replica heuristics in the weak-noise regime.

I. INTRODUCTION

Compressed sensing [?], [?] is a signal processing technique that compresses analog vectors by means of a linear transformation. By leveraging prior knowledge of the signal structure (e.g., sparsity) and by designing efficient nonlinear reconstruction algorithms, effective compression is achieved by taking a much smaller number of measurements than the dimension of the original signal.

Most of the compressed sensing literature focuses on the setup where

- a) performance is measured on a worst-case basis with respect to the n -dimensional input vector.
- b) the encoder is constrained to be a linear mapping characterized by a $k \times n$ matrix, called the *sensing* or *measurement matrix*, which is usually assumed to be random, and known at the decoder.
- c) the decoder is a low-complexity algorithm which is robust with respect to observation noise, for example, decoders based on convex optimizations such as ℓ_1 -minimization [?] and ℓ_1 -penalized least-squares (i.e. LASSO) [?], greedy algorithms such as matching pursuit [?], graph-based iterative decoders such as approximate message passing (AMP) [?].

In contrast, in this paper we formulate an information-theoretic fundamental limit in the following setup:

Y. Wu is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: yihongwu@wharton.upenn.edu. S. Verdú is with Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Email: verdu@princeton.edu

This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSol), an NSF Science and Technology Center, under Grant CCF-0939370.

- a) the input vector is random with a known distribution and performance is measured on an average basis. Similar Bayesian modeling is followed in some of the compressed sensing literature, for example, [?], [?], [?], [?].
- b) three different encoder classes are considered: *optimal nonlinear*, *optimal linear* and *random linear* encoders.
- c) the decoder is the *optimal* Bayes procedure.

The overarching goal is to investigate the fundamental tradeoff between reconstruction fidelity and measurement rate $\frac{k}{n}$ as $n \rightarrow \infty$, as a functional of the signal and noise statistics.

When the measurements are noiseless, the goal is to reconstruct the original signal in the sense of driving the error probability to zero as the ambient dimension, n , grows. For many input processes (e.g., i.i.d. ones), it turns out that there exists a threshold for the measurement rate, above which it is possible to achieve a vanishing error probability and below which the error probability will eventually approach one for any sequence of encoder-decoder pairs. Such a phenomenon is known as *phase transition* in statistical physics. In information-theoretic parlance, we say that the *strong converse* holds.

We introduced the framework of almost-lossless analog compression in [?] as a Shannon-theoretic formulation of noiseless compressed sensing. Under regularity conditions such as linearity of the encoder and Lipschitz continuity of the decoder, [?] derives various coding theorems for the minimal measurement rate involving the Rényi information dimension of the input distribution [?]. See also [?, Section III] for a non-asymptotic exposition.

The focus of this paper is the case where the measurements are corrupted by additive noise. Exact analog signal recovery is obviously impossible and we gauge reconstruction fidelity by the *noise sensitivity*, defined as the ratio between the mean-square reconstruction error and the noise variance. Similar to the behavior of error probability in the noiseless case, there exists a phase-transition threshold of measurement rate, which only depends on the input statistics, above which the noise sensitivity is bounded for all noise variances, and below which the noise sensitivity blows up as the noise variance tends to zero.

In Section II we consider three formulations of noise sensitivity: *optimal nonlinear*, *optimal linear* and *random linear* (with i.i.d. entries) encoder and the associated *optimal* decoder. For memoryless sources, we show that for any input distribution, the phase-transition threshold for optimal encoding is given by the input information dimension. Moreover, this result also holds for discrete-continuous mixtures with optimal

linear encoders and Gaussian random measurement matrices. The fact that randomly chosen sensing matrices turn out to incur no penalty in phase-transition threshold with respect to optimal nonlinear encoders lends further importance to the conventional compressed sensing setup. In addition, we compare the optimal phase-transition threshold to the sub-optimal threshold of several practical reconstruction algorithms under various input distributions. In particular, in Section III we demonstrate that the thresholds achieved by the LASSO decoder and the AMP decoder [?] lie far from the optimal boundary, especially in the highly sparse regime, which is most relevant to compressed sensing applications.

Invoking the results in [?], in Section II-F we show that the calculation of the reconstruction error with random measurement matrices based on *heuristic* replica methods in [?] predicts the correct phase-transition threshold. These results also serve as a rigorous verification of the replica calculations in [?] in the high-SNR regime (up to $o(\sigma^2)$ as the noise variance σ^2 vanishes).

Omitted proofs can be found in [?].

II. NOISY COMPRESSED SENSING

A. Setup

The basic setup of noisy compressed sensing is the joint source-channel coding problem shown in Fig. 1, where we

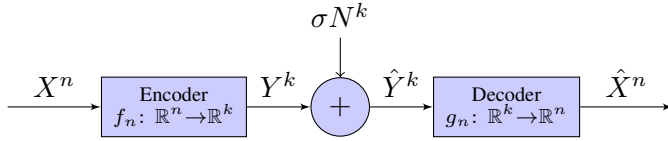


Fig. 1: Noisy compressed sensing setup.

assume that

- The source X^n consists of i.i.d. copies of a real-valued random variable X with unit variance.
- The channel is memoryless with additive Gaussian noise σN^k where $N^k \sim \mathcal{N}(0, \mathbf{I}_k)$.
- The *measurement rate*, i.e., the dimensionality compression ratio, is given by $R = \frac{k}{n}$.
- Unit average power constraint of on the encoded signal: $\frac{1}{k} \mathbb{E}[\|f_n(X^n)\|_2^2] \leq 1$.
- The reconstruction error is gauged by the per-symbol MSE distortion: $d(x^n, \hat{x}^n) = \frac{1}{n} \|\hat{x}^n - x^n\|_2^2$.

In this setup, the fundamental question is: For a given noise variance and measurement rate, what is the lowest reconstruction error? For a given encoder f , the corresponding optimal decoder g is the *MMSE estimator* of the input X^n given the channel output $\hat{Y}^k = f(X^n) + \sigma N^k$. Therefore the optimal distortion achieved by encoder f is $\text{mmse}(X^n | f(X^n) + \sigma N^k)$, where $\text{mmse}(U|V) \triangleq \mathbb{E}[\|U - \mathbb{E}[U|V]\|_2^2]$.

B. Signal model

Sparse vectors, supported on a subspace with dimension smaller than n , play an important role in signal processing and

statistical models. A stochastic model that captures sparsity is the following mixture distribution [?], [?], [?]:

$$P_X = (1 - \gamma)\delta_0 + \gamma P_c, \quad (1)$$

where δ_0 denotes a unit mass at 0, P_c is an absolutely continuous (with respect to Lebesgue measure) distribution from which the non-zero entries are drawn, and $0 \leq \gamma \leq 1$ parametrizes the signal sparsity. This model corresponds to the regime of *proportional (or linear) sparsity*, because X^n drawn independently from (1) has approximately γn non-zeros.

Generalizing (1), we henceforth consider *discrete-continuous mixed* distributions:

$$P_X = (1 - \gamma)P_d + \gamma P_c, \quad (2)$$

where P_d is a discrete probability measure and P_c is an absolutely continuous probability measure. In addition to sparsity, there are other signal structures that have been previously explored in the compressed sensing literature, which fit the model in (2). For example, the so-called *simple* signal in infrared absorption spectroscopy [?, Example 3, p. 914] is such that each entry of the signal vector is constrained to lie in the unit interval, with most of the entries saturated at the boundaries (0 or 1) (see Section III for its probabilistic model). Although most of the results in the present paper hold for *arbitrary* input distributions, with no practical loss of generality, we focus on discrete-continuous mixtures because of their relevance to compressed sensing applications.

C. Distortion-rate tradeoff

For a fixed noise variance, we define three distortion-rate functions that correspond to *optimal* encoding, *optimal linear* encoding and *random linear* encoding respectively.

1) *Optimal encoder*: The minimal distortion achieved by the optimal encoding scheme is given by:

$$D^*(X, R, \sigma^2) \triangleq \limsup_{n \rightarrow \infty} \frac{1}{n} \inf_f \{ \text{mmse}(X^n | f(X^n) + \sigma N^k) : \mathbb{E}[\|f(X^n)\|_2^2] \leq Rn \}. \quad (3)$$

The asymptotic optimization problem in (3) can be solved by applying Shannon's joint source-channel coding separation theorem for memoryless channels and sources [?, Section XI], which states that the lowest rate, R , that achieves distortion D is given by

$$R = \frac{R_X(D)}{C(\sigma^2)}, \quad (4)$$

where $R_X(\cdot)$ is the rate-distortion function of X under the mean-square error distortion and $C(\sigma^2) = \frac{1}{2} \log(1 + \sigma^{-2})$ is the AWGN channel capacity. By the monotonicity of the rate-distortion function, we have

$$D^*(X, R, \sigma^2) = R_X^{-1} \left(\frac{R}{2} \log(1 + \sigma^{-2}) \right). \quad (5)$$

In general, optimal joint source-channel encoders are nonlinear.

2) *Optimal linear encoder*: To analyze the fundamental limit of conventional noisy compressed sensing, we restrict the encoder f in (3) to be a linear mapping represented by a matrix $\mathbf{H} \in \mathbb{R}^{k \times n}$, and denote the left-hand side of (3) by $D_L^*(X, R, \sigma^2)$. Since X^n are i.i.d. with zero mean and unit variance, the input power constraint in (3) simplifies to $\|\mathbf{H}\|_F^2 \leq Rn$, where $\|\cdot\|_F$ denotes the Frobenius norm.

3) *Random linear encoder*: We consider the ensemble performance of a sequence of $k \times n$ random measurement matrices $\mathbf{A}_n$ with i.i.d. entries of zero mean and variance $\frac{1}{n}$ and define $D_L(X, R, \sigma^2)$ as in (3) with f replaced by $\mathbf{A}_n$. Note that the power constraint holds on average: $\mathbb{E}[\|\mathbf{A}_n\|_F^2] = k$. By definition, we have $0 \leq D^* \leq D_L^* \leq D_L \leq 1$.

D. Phase transition of noise sensitivity

A key objective of compressed sensing with noisy observations is to achieve robust reconstruction, obtaining a reconstruction error proportional to the noise variance. To quantify robustness, we analyze the *noise sensitivity*, namely, the ratio between the mean-square error and the noise variance, as a function of R and σ^2 . As a succinct characterization of robustness, we focus particular attention on the worst-case noise sensitivity:

Definition 1. The *worst-case noise sensitivity* of optimal encoding is defined as

$$\zeta^*(X, R) = \sup_{\sigma^2 > 0} \frac{D^*(X, R, \sigma^2)}{\sigma^2}. \quad (6)$$

For linear encoding, ζ_L^* and ζ_L are analogously defined with D^* replaced by D_L^* and D_L , respectively.

The phase-transition threshold of the noise sensitivity is defined as the minimal measurement rate R such that the noise sensitivity is bounded for all σ^2 :

Definition 2. For optimal encoding, define

$$\mathcal{R}^*(X) \triangleq \inf \{R > 0 : \zeta^*(X, R) < \infty\}. \quad (7)$$

For linear encoding, $\mathcal{R}_L^*(X)$ and $\mathcal{R}_L(X)$ are analogously defined with ζ^* in (7) replaced by ζ_L^* and ζ_L , respectively.

Alternatively, we can consider the *asymptotic noise sensitivity* by replacing the supremum in (6) with the limit as $\sigma^2 \rightarrow 0$, denoted by ξ^* , ξ_L^* and ξ_L respectively, which are finite if and only if the corresponding suprema are finite. Therefore, the phase-transition thresholds can also be defined as the minimum measurement rate for which the reconstruction error vanishes according to $O(\sigma^2)$ as $\sigma^2 \rightarrow 0$.

By definition, the phase-transition thresholds are ordered naturally as $0 \leq \mathcal{R}^*(X) \leq \mathcal{R}_L^*(X) \leq \mathcal{R}_L(X) \leq 1$. It can be shown that Gaussian input $X_G \sim \mathcal{N}(0, 1)$ maximizes D^* , D_L^* and D_L simultaneously under the variance constraint with explicit formulae given in [?], which yield the following phase-transition threshold:

$$\mathcal{R}^*(X_G) = \mathcal{R}_L^*(X_G) = \mathcal{R}_L(X_G) = 1. \quad (8)$$

The equality of the three phase-transition thresholds turns out to hold well beyond the Gaussian signal model. In the next subsection, we formulate and prove the existence of the phase thresholds for all three distortion-rate functions and discrete-continuous mixtures, which turn out to be equal to the information dimension of the input distribution.

E. Main results

In this subsection, we give bounds and exact characterizations of the three phase-transition thresholds introduced in Definition 2 in terms of the input information [?] and MMSE dimension [?], defined as follows.

Definition 3. Let X be a real-valued random variable. Let $m \in \mathbb{N}$. The *information dimension* of X is defined as

$$d(X) = \lim_{m \rightarrow \infty} \frac{H(\lfloor mX \rfloor)}{\log m}. \quad (9)$$

The *MMSE dimension* of X is defined as

$$\mathcal{D}(X) = \lim_{\text{snr} \rightarrow \infty} \text{snr} \cdot \text{mmse}(X, \text{snr}), \quad (10)$$

where $\text{mmse}(X, \text{snr}) \triangleq \text{mmse}(X | \sqrt{\text{snr}}X + N)$.

The $\liminf$ and $\limsup$ in (9) (resp. (10)) are called lower and upper information (resp. MMSE) dimensions of X , denoted by $\underline{d}(X)$ and $\bar{d}(X)$ (resp. $\underline{\mathcal{D}}(X)$ and $\bar{\mathcal{D}}(X)$).

It is shown in [?, Theorem 8] that the information dimensions are sandwiched between the MMSE dimensions:

$$0 \leq \underline{\mathcal{D}}(X) \leq \underline{d}(X) \leq \bar{d}(X) \leq \bar{\mathcal{D}}(X) \leq 1. \quad (11)$$

For discrete-continuous mixtures (2), the MMSE dimension coincides with the information dimension [?, Theorem 15]:

$$\mathcal{D}(X) = d(X) = \gamma. \quad (12)$$

It is possible that the MMSE dimension does not exist and the inequalities in (11) are strict (e.g., Cantor distributed X [?, Theorem 16]).

The phase-transition threshold for *optimal encoding* is given by the upper *information dimension* of the input:

Theorem 1. For any X , $\mathcal{R}^*(X) = \bar{d}(X)$.

The next result shows that random linear encoders with i.i.d. Gaussian coefficients also achieve information dimension for any discrete-continuous mixtures, which, in view of Theorem 1, implies that, at least asymptotically, (random) linear encoding suffices for robust reconstruction as long as the input distribution contains no singular component.

Theorem 2. Assume that X has a discrete-continuous mixed distribution as in (2) with $H(P_d) < \infty$. Then

$$\mathcal{R}^*(X) = \mathcal{R}_L^*(X) = \mathcal{R}_L(X) = d(X) = \mathcal{D}(X) = \gamma. \quad (13)$$

Moreover, (13) holds for any noise distribution with finite non-Gaussianity (defined as its relative entropy with respect to a Gaussian distribution with the same mean and variance).

F. Results relying on replica heuristics

Based on the statistical-physics approach in [?], [?], the decoupling principle results in [?] were imported into the compressed sensing setting in [?] to *postulate* the following formula for $D_L(X, R, \sigma^2)$. Note that this result is based on replica heuristics currently lacking a rigorous justification.

Replica Symmetry Postulate ([?, Corollary 1, p.5]).

$$D_L(X, R, \sigma^2) = \text{mmse}(X, \eta R \sigma^{-2}), \quad (14)$$

where $0 < \eta < 1$ satisfies $\eta^{-1} = 1 + \sigma^{-2} \text{mmse}(X, \eta R \sigma^{-2})$. In the case of multiple solutions, η is chosen to minimize the free energy $I(X; \sqrt{\eta R \sigma^{-2}} X + N) + \frac{R}{2}(\eta - 1 - \log \eta)$.

Assuming the validity of the replica symmetry postulate, it can be shown that the phase-transition threshold for random linear encoding is always sandwiched between the lower and the upper MMSE dimensions of the input. The relationship between the MMSE dimension and the information dimension in (11) plays a key role in analyzing the minimizer of the free energy.

Theorem 3. Assume that the replica symmetry postulate holds for X . Then for any i.i.d. random measurement matrix whose entries have zero mean and variance $\frac{1}{n}$,

$$\underline{\mathcal{D}}(X) \leq \mathcal{R}_L(X) \leq \overline{\mathcal{D}}(X). \quad (15)$$

Therefore if $\mathcal{D}(X)$ exists, we have

$$\mathcal{R}_L(X) = \mathcal{D}(X) = d(X), \quad (16)$$

and in addition, $D_L(X, R, \sigma^2) = \frac{d(X)}{R-d(X)} \sigma^2 (1 + o(1))$.

The general result in Theorem 3 holds for any input distribution but relies on the conjectured validity of the replica symmetry postulate. For the special case of discrete-continuous mixtures in (2), in view of (12), Theorem 3 predicts (with the caveat of the validity of the replica symmetry postulate) that the phase-transition threshold for random linear encoding is γ , which agrees with the rigorously proven result in Theorem 2. Therefore, the only added benefit of Theorem 3 is to allow singular components in the input distribution.

The achievability proof of $\mathcal{R}_L(X)$ in Theorem 3 is a direct application of the noiseless result in [?], where it is shown that there exists a sequence of Lipschitz decompressors with bounded Lipschitz constants and vanishing block error probability. The noiseless Lipschitz decompressor as a suboptimal estimator achieves finite noise sensitivity. This achievability strategy applies to any noise with finite variance, without requiring that the noise be additive, memoryless or that it have a density. In contrast, replica-based results rely crucially on the fact that the additive noise is memoryless Gaussian.

III. COMPARISONS TO LASSO AND AMP ALGORITHMS

In this section, we compare the phase transition thresholds of LASSO and AMP achieved in the Bayesian setting to the optimal thresholds for the following families of input distributions considered in [?, p. 18915], indexed by $\chi = \pm, +$ and $\square$ respectively, which all belong to the mixture form (2):

- $\pm$ sparse signals (1);
- $+$ sparse non-negative signals (1) with the continuous component P_c supported on $\mathbb{R}_+$.
- $\square$ simple signals (see Section II-B): $P_X = (1 - \gamma) \left(\frac{1}{2} \delta_0 + \frac{1}{2} \delta_1 \right) + \gamma P_c$, where P_c is some absolutely continuous distribution supported on the unit interval.

We consider the AMP decoder [?] and the LASSO decoder:

$$\tilde{g}_\lambda(y) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \|y - \mathbf{A}x\|_2^2 + \lambda \|x\|_1, \quad (17)$$

where $\lambda > 0$ is a regularization parameter. For Gaussian sensing matrices and Gaussian observation noise, the asymptotic mean-square error achieved by LASSO can be determined as a function of (P_X, λ, σ^2) by applying [?, Corollary 1.6]. For sparse X distributed according to (1), the following minimax expression for the worst-case (or asymptotic) noise sensitivity of LASSO under the least favorable prior with optimized λ is proposed in [?, Proposition 3.1(1.a)] and proved by [?]:

$$\tilde{\zeta}(X, R) = \tilde{\xi}(X, R) = \begin{cases} \frac{R_\pm(\gamma)}{R - R_\pm(\gamma)} & R > R_\pm(\gamma) \\ \infty & R \leq R_\pm(\gamma) \end{cases} \quad (18)$$

with

$$R_\pm(\gamma) = \min_{\tau \geq 0} \gamma(1 + \tau^2) + c(1 - \gamma)((1 + \tau^2)\Phi(-\tau) - \tau\varphi(\tau)) \quad (19)$$

and $c = 2$. Analogously, the LASSO decoder (17) can be adapted to other signal structures (see for example [?, Sec. VI-A]), resulting in the phase-transition threshold $R_+(\gamma)$ for sparse positive signals given by (19) with $c = 1$ and $R_\square(\gamma) = \frac{\gamma+1}{2}$ for simple signals. Furthermore, (18) also applies to the AMP algorithm [?]. The suboptimal thresholds are plotted in Fig. 2 along with the optimal threshold obtained from Theorem 2 which is γ . In the gray area below the diagonal in the (γ, R) -phase diagram, the noise sensitivity blows up for any sequence of sensing matrices and decoders. Moreover, we observe that the LASSO and AMP decoders are severely suboptimal unless γ is close to one. In the highly sparse regime which is most relevant to compressed sensing problems, it follows from [?, Theorem 3] that for sparse signals ($\chi = \pm$ or $+$), $R_\chi(\gamma) = 2\gamma \log_e \frac{1}{\gamma}(1 + o(1))$, as $\gamma \rightarrow 0$, which implies that R_χ has infinite slope at $\gamma = 0$. Therefore when $\gamma \ll 1$, the LASSO and AMP decoders require on the order of $2s \log_e \frac{n}{s}$ measurements to successfully recover the unknown vector with s nonzero components. In contrast, s measurements suffice when using an optimal decoder (or ℓ_0 -minimization decoder). The LASSO or AMP decoders are also highly suboptimal for simple signals, since $R_\square(\gamma)$ converges to $\frac{1}{2}$ instead of zero as $\gamma \rightarrow 0$. Interestingly, the phase-transition thresholds of block error probability in the noiseless case are identical to Fig. 2, as observed in [?, Sec. 5.1].

For sparse signals of the form (1) with $\gamma = 0.1$, Fig. 3 compares the asymptotic noise sensitivity of the LASSO (and AMP) decoder to the optimal noise sensitivity predicted by Theorem 3 based on replica heuristics. Note that the phase-transition threshold of LASSO is approximately 3.3 times the optimal.

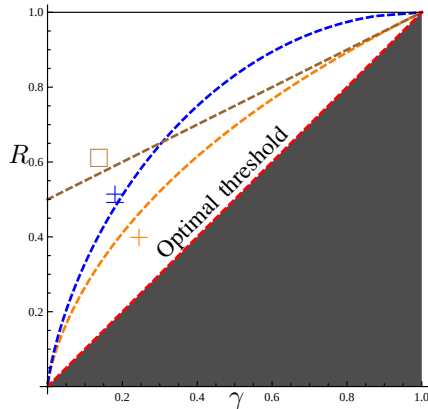


Fig. 2: Suboptimal thresholds obtained with LASSO and AMP v.s. optimal threshold for the three signal models in Section III.

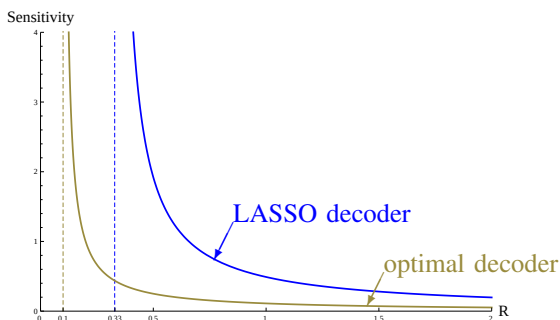


Fig. 3: Asymptotic noise sensitivity of the optimal decoder and the LASSO decoder exhibiting phase transitions: sparse signal model (1) with $\gamma = 0.1$.

IV. CONCLUDING REMARKS

As opposed to a worst-case (Hamming) approach, in this paper we adopt a statistical (Shannon) framework for compressed sensing by modeling input signals as random processes rather than individual sequences. As customary in information theory, it is advisable to initiate the study of fundamental limits assuming independent identically distributed information sources with known distributions. Naturally, this entails substantial loss of practical relevance, so generalization to sources with memory is left for future work. We have obtained the phase-transition thresholds (minimum measurement rate) of normalized MMSE with noisy observations achievable by optimal nonlinear, optimal linear, and random linear encoders combined with the corresponding optimal decoders. For discrete-continuous mixtures, the optimal phase-transition threshold is shown to be the information dimension of the input. The phase-transition thresholds of popular decoding algorithms (e.g., LASSO or AMP decoders) turn out to be far from the optimal boundary. In a recent preprint [?], it is shown that using sensing matrices constructed from spatially coupled error-correcting codes and the corresponding AMP decoder, information dimension and MMSE dimension are respectively achievable in both noiseless and noisy cases under mild conditions, which are optimal in view of the results in

[?].

One of our main findings is Theorem 2 which shows that *Gaussian* sensing matrices achieve the same phase-transition threshold as optimal nonlinear encoding, for any discrete-continuous mixture. This result is universal in the sense that it holds for arbitrary noise distributions with finite non-Gaussianness. Moreover, the fundamental limit depends on the input statistics only through the weight of the analog component, regardless of the specific discrete and continuous components. The universal optimality of random sensing matrices with non-Gaussian i.i.d. entries in terms of phase-transition thresholds remains open.

REFERENCES