On the rate distortion function of Bernoulli Gaussian sequences

Cheng Chang
D. E. Shaw & Co, New York.
Email: cchang@eecs.berkeley.edu

***Abstract*—We study the rate distortion function of the Bernoulli-Gaussian random variable which can be used to model sparse signals. Both lower and upper bounds on the rate distortion function are given. We show that the bounds are almost tight in the low distortion regime for sparse signals. Interestingly, a naive coding scheme is near-optimal in this scenario.**

I. Introduction

Consider a sequence real numbers [1] $\mathsf{x}_1, \mathsf{x}_2, ...\mathsf{x}_n$, where x_i is either *exactly* zero or an arbitrary non-zero real number. In the signal processing literature, the sequence x^n is called sparse if most of the entries are zero. In their seminal work on compressive sensing [3] and [6], Candès, Tao and Donoho show that, to *losslessly* reconstruct the sparse sequnce x^n, only a fraction of the n measurements are needed. Here a measurement is a linear projection of x^n on the real line $\mathcal{R}$. Furthermore, the reconstruction can be done by a linear programming based efficient algorithm. In the compressed sensing literature, the non-zero part of the sparse signals are, in general, arbitrary real numbers without any statistical distribution assigned to them. Furthermore the compressed sensing system is to recover the signals x^n *losslessly*. A natural question to ask is what if the source statistics are known to the coding system? More significantly, what if the goal of the sensing system is only to recover the original sequence within a certain distortion? In the recent work by Fletcher etc. [8], [7], [9], the authors studied the compressive sensing problem for sparse Gaussian signals. What is lacking in these papers is an information theoretic study of the bounds on the rate distortion functions of the sparse signals. In this paper, we attempt to answer these questions.

A. Bernoulli-Gaussian random variable $\Xi(p, \sigma^2)$

The information theoretic model of the "sparse Gaussian" signals is captured in the following what we call a Bernoulli-Gaussian random variable.

Definition 1: A random variable x is Bernoulli-Gaussian, denoted by $\Xi(p, \sigma^2)$, if $\mathsf{x} = \mathsf{b} \times \mathsf{s}$, where s is a Gaussian random variable with mean 0 and variance σ^2, $\mathsf{s} \sim N(0, \sigma^2)$, and b is a Bernoulli p random variable, $\Pr(\mathsf{b} = 0) = 1 - p$ and $\Pr(\mathsf{b} = 1) = p$, $p \in [0, 1]$.

This random variable is a mixture of a continuous and a discrete random variable. A continuous pdf is not well defined as the point probability at 0 is not zero. This adds to the difficulties to the study of the rate distortion functions as the famous Shannon lower bound does not apply. The main result of this paper is a lower bound and an upper bound on the rate distortion functions of a sequence of independent random variables with distribution $\Xi(p, \sigma^2)$. First, we review some of the results in rate distortion theory in both the average sense and more importantly, the strong sense.

B. Review of rate distortion theory

In the standard setup of rate distortion theory, the encoder f_n maps a length-n sequence $\mathsf{x}^n \in \mathcal{X}^n$, $\mathsf{x} \sim p_{\mathsf{x}}$, into nR bits and the decoder g_n reconstruct a lossy version of the original sequence:

$$f_n : \mathcal{X}^n \to \{0,1\}^{nR} \quad and \; g_n : \{0,1\}^{nR} \to \hat{\mathcal{X}}^n,$$

and the distortion is defined as[2] $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i)$. The rate distortion function in the average sense, defined in [5], is the infimum of rate R, such that a lossy source coding system (f_n, g_n) of that rate exist with

$$\lim_{n\to\infty} E\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n)))\right) \leq D \tag{1}$$

The rate distortion function in the strong sense is defined similarly with the following criteria for the coding system: for all $\delta > 0$

$$\lim_{n\to\infty} \Pr\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n)) \geq D + \delta\right) = 0 \tag{2}$$

It turns out that the rate distortions function for both the average distortion and the strong distortion are the same for discrete random variables as detailed in Chapter 13.6 [5]. This result can be generalized to continuous and mixed random variables, like Gaussian Bernoulli, whose variance is finite and whose pdf satisfies some regularity conditions [4]. The proof can be carried out by quantizing the probability density

This work was done when the author was with Hewlett-Packard Labs, Palo Alto.

[1] We use x, y, u for random variables and x, y, u for the realization of the random variables. We denote by x^n the sequence $\mathsf{x}_1, \mathsf{x}_2, ...\mathsf{x}_n$. We use bit and $\log_2$ in this paper.

[2] In this paper, the distortion is the squared error distortion, i.e. $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$.

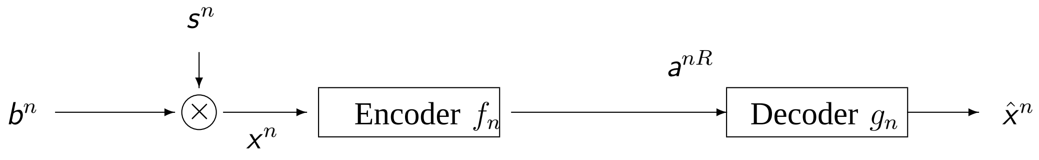

Fig. 1. A lossy source coding system for Bernoulli-Gaussian sequence $x^n = b^n \times s^n$, $a^{nR} \in \{0,1\}^{nR}$

function and then by using the proof for discrete random variables in Chapter 13.6 [5]. A good lossy coding system in the strong sense is not *necessarily* good in the average distortion sense. However, given a good lossy coding system in the strong sense, we can modify it in a way such that it is good in both senses if the mean and variance of the random variable are finite [4]. The following lemma characterizes the rate distortion function $R(D, p_x)$, for $x \sim p_x$, in both the average sense and the strong sense.

Lemma 1: Rate distortion theorem [10]:

$$R(D, p_x) = \min_{p_{\hat{x}|x}: E(d(x,\hat{x})) \le D} I(x; \hat{x}). \quad (3)$$

where the expectation is taken over the joint distribution $p_x(x)p_{\hat{x}|x}(\hat{x}|x)$.

It is usually difficult to evaluate (3), however, for Gaussian random variables the rate distortion functions can be easily evaluated:

$$R(D, N(0,\sigma^2)) = \{ \begin{array}{cc} \frac{1}{2}\log_2 \frac{\sigma^2}{D}, & 0 \le D \le \sigma^2, \\ 0, & D > \sigma^2. \end{array} \quad (4)$$

It is known that for the squared error distortion, with the same variance, Gaussian random variable has the highest rate distortion function [5] [2]. And for the squared error distortion, the rate distortion function is lower bounded by the Shannon lower bound [5]. Hence the rate distortion function $R(D, p_x)$ for continuous distribution p_x with variance σ^2 can be bounded by:

$$h(p_x) - \frac{D}{2}\log(2\pi e) \le R(D, p_x) \le R(D, N(0,\sigma^2)) \quad (5)$$

where $R(D, N(0,\sigma^2))$ is given in (4) and $h(p_x)$ is the continuous entropy of p_x.

The obvious limitation of the above Shannon lower bound is that the continuous entropy is not well defined for non-continuous random variables. For the Bernoulli-Gaussian random variable $\Xi(p,\sigma^2)$, the conditional entropy is, roughly speaking, $-\infty$, due to the fact that the $P(x=0) > 0$. And it is not known how the Shannon lower bound can be generalized to non-continuous random variables such as Bernoulli-Gaussian.

C. Rate distortion function for $\Xi(p,\sigma^2)$

A lossy source coding system for Bernoulli-Gaussian sequences is shown in Fig. 1. We aim to derive an upper and a lower bound on the rate distortion function $R(D, \Xi(p,\sigma^2))$. We summarize some properties and bounds of $R(D, \Xi(p,\sigma^2))$ in the following four propositions.

First we explain why we only need to study $R(D, \Xi(p,1))$. We will simply write $R(D, \Xi(p,1))$ as $R(D,p)$ in the rest of the paper and investigate $R(D,p)$.

Proposition 1: $R(D, \Xi(p,\sigma^2)) = R(\frac{D}{\sigma^2}, \Xi(p,1))$.

From this point on, we only investigate $R(D,p)$. Now we give three somewhat obvious bounds.

Proposition 2: Upper bound I on $R(D,p)$:

$$\begin{aligned} R(D,p) &\le H(p) + pR(\frac{D}{p}, N(0,1)) \\ &= pR(D, N(0,p)) + H(p) \quad (6) \end{aligned}$$

where $R(D, N(0,1))$ is the Gaussian rate distortion function for $N(0,1)$, defined in (4).

Proposition 3: Upper bound II on $R(D,p)$:

$$R(D,p) \le R(D, N(0,p)) \quad (7)$$

Proposition 4: A lower bound on $R(D,p)$:

$$R(D,p) \ge pR(\frac{D}{p}, N(0,1)) = pR(D, N(0,p)) \quad (8)$$

The above propositions are fairly straightforward. Proposition 1 is due to the squared error distortion. To prove Proposition 2, we construct a very simple coding system that first losslessly describe the locations of the non-zero elements of $x^n \sim \Xi(p,1)$ with roughly $nH(p)$ bits, then lossily describe the non zero part of the sequence, of roughly length $nH(p)$, using a Gaussian lossy coder with roughly $pnR(D, N(0,p))$ bits. We prove Proposition 3 by using the well known fact that under squared error distortion, for continuous random variables with the same variance, Gaussian sequences require the highest rate. The difficulty is that $\Xi(p,1)$ is not a continuous random variable. We approximate $\Xi(p,1)$ by a sequence of continuous random variables whose rate distortion functions converge to that of $\Xi(p,1)$. In the proof of Proposition 4, we use a genie-based proof by by letting the decoder know the non-zero locations (b^n) for free and derive *a* lower bound of $R(D,p)$ by the Gaussian rate distortion function for the non-zero part of x^n. Due to the page limit, we leave the details of the proofs in the tech report [4].

Among the three bounds described in Proposition 2, 3 and 4, we find the lower bound in Proposition 4 the most unsatisfactory. Shannon lower bound (5) does not apply to the Bernoulli-Gaussian random $\Xi(p,1)$ variables because the differential entropy of $\Xi(p,1)$ is, roughly speaking, negative infinity. In next several sections, we aim to improve the lower bound in Proposition 4. As a simple corollary of this new lower bound, we give a closed form lower bound on the rate distortion function that improves the previous bound by $p\log_2\frac{1}{p}$ in the low distortion regime ($D \ll 1$). Notice that the gap between the lower bound in Proposition 4 the upper bound in Proposition 2 is $H(p)$, and $\lim\limits_{p\to 0}\frac{p\log_2\frac{1}{p}}{H(p)} = 1$. Hence we close the gap between the upper and lower bound for sparse signals ($p \ll 1$) in low distortion regime ($D \ll 1$).

II. Main result: A new lower bound on $R(D,p)$

The main result of this paper is an improved lower bound on the rate distortion function for Bernoulli-Gaussian random variable $\Xi(p,1)$, summarized in the following theorem.

Theorem 1: A improved lower bound on the rate distortion function $R(D,p)$ for Bernoulli-Gaussian random variable $\Xi(p,1)$ under distortion constraint D.

$$R(D,p) \geq pR(D, N(0,p)) + R_I(D,p). \tag{9}$$

where $R_I(D,p) = \max\limits_{L\geq 0}\{\min\limits_{U\geq L, r\in[0,1-p]:T(L,U,r)\leq D} h(L,U,r)\}$

$$\text{and } T(L,U,r) = rL^2 + 2p\int_L^U (s-L)^2\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds, \tag{10}$$

$$h(L,U,r) = \begin{cases} (p\times\Pr(|\mathsf{s}|>U)+r)\mathcal{D}(\frac{p\times\Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r}\|p), \\ \qquad\qquad \text{if } \frac{p\times\Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r} \geq p \\ 0, \quad \text{if } \frac{p\times\Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r} < p \end{cases}$$

where $\mathsf{s} \sim N(0,1)$ and $\mathcal{D}(p_1\|p_2)$ is the KL divergence between the Bernoulli-p_1 and Bernoulli-p_2 distributions.

We defer the sketch of the proof to the next section. It is hard to see the significance of Theorem 1 in its current form. Comparing (9) with the lower bound in (6) in Proposition 4, we see that $R_I(D,p)$ is the improvement over the known lower bound in Proposition 4. So how much is the improvement? The following proposition tells us that the improvement is bigger for small distortion D, i.e. the high resolution regime.

Proposition 5: $R_I(D,p)$ is monotonically decreasing with D, i.e. for $D_1 > D_2$, $R_I(D_1,p) \leq R_i(D_2,p)$.

The improvement cannot exceed $H(p)$ because the gap between the upper bound in Proposition 2 and the lower bound in Proposition 4 is $H(p)$. In the low distortion regime, i.e. $D \ll 1$, the next proposition tells us that the improvement $R_I(D,p)$ is close to $p\log_2\frac{1}{p}$.

Proposition 6: Asymptotic behavior of $R_I(D,p)$ in the low distortion regime: for any $p > 0$

$$\lim_{D\to 0} R_I(D,p) \geq p\log_2\frac{1}{p}$$

This is a very interesting result as we know that for $p \to 0$

$$H(p) = p\log_2(\frac{1}{p}) + (1-p)\log_2(\frac{1}{1-p}) = p\log_2(\frac{1}{p}) + O(p)$$

As $\log_2(\frac{1}{p}) \gg 1$ for $p \ll 1$, so the ratio of the improvement, $R_I(p,D)$, over $H(p)$ converges to 1 as $p\to 0$. Where $H(p)$ is the gap between the lower bound in Proposition 4 and the upper bound in Proposition 2, This tells us that in this regime ($p \ll 1$ and $D \ll 1$), the upper bound in Proposition 2 and the lower bound in Theorem 1 together with Proposition 6 are quite tight.

In summary, we have the following lower and upper bounds that are almost identical in the low distortion ($D \ll 1$) regime:
Upper bound(Proposition 2): $pR(D, N(0,p)) + H(p)$.
Lower bound(Proposition 6): $pR(D, N(0,p)) + p\log_2\frac{1}{p}$.

The gap between these two bounds is at most $O(p)$ which is $o(H(p))$ for small $p \ll 1$.

An example

We plot the bounds in Propositions 2, 3, 4 and Theorem 1 in Figure 2 for $p = 0.1$, i.e. bounds for the rate distortion function for random variable $\Xi(0.1,1)$. As shown in Figure 2. As predicted in Proposition 5, the improvement $R_I(D,p)$ on the lower bound is bigger for smaller D.

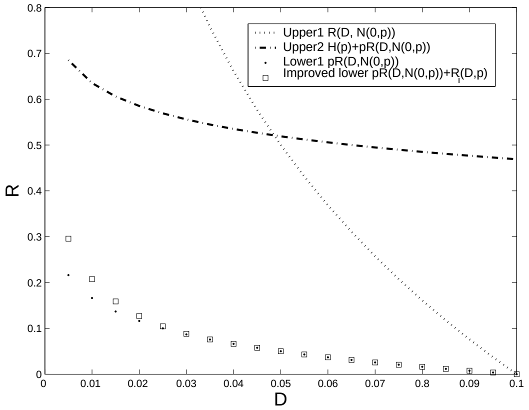

Fig. 2. Lower and upper bounds on $R(D,p)$ for $p = 0.1$ at high distortion levels, the distortion D runs from 0.005 to 0.1

Next we illustrate the asymptotic behavior of the improvement $R_I(D,p)$ as $D \to 0$. As predicted in Proposition 6, the improvement $R_I(D,p)$ converges to $p\log_2\frac{1}{p}$ as $D\to 0$. This is shown in Figure 3 for $p = 0.1$.

III. Proof of Theorem 1

Due to page limit, we only sketch our proof in this paper. Details are in Tech Report [4].

For a lossy source coding system for $x^n = \mathsf{b}^n \times \mathsf{s}^n \sim \Xi(p,1)$ as shown in Fig. 1, the output of the encoder is a^{nR}. If the distortion D is satisfied in both the average sense and the strong sense, the rate R can be lower bounded as follows:

$$\begin{aligned} R &\geq \frac{I(\mathsf{a}^{nR};\mathsf{s}^n|\mathsf{b}^n) + I(\mathsf{a}^{nR};\mathsf{b}^n)}{n} && (11) \\ &\geq pR(D,N(0,p)) + \frac{I(\mathsf{a}^{nR};\mathsf{b}^n)}{n} && (12) \\ &\geq pR(D,N(0,p)) + R_I(p,D) && (13) \end{aligned}$$

where the RHS of (13) is the final form in Theorem 1.

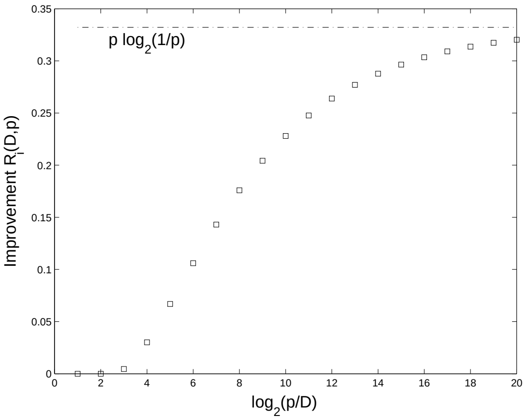

Fig. 3. The improvement $R_I(D,p)$ for $p=0.1$ at low distortion levels. As proved in Proposition 6, $R_I(D,p) \to p\log_2\frac{1}{p}$ as $D \to 0$

First in (11), we lower bound the number of bits nR by the sum of two mutual information terms. The first is a "continuous" mutual information term in (12). We lower bound $I(a^{nR}; s^n|b^n)$ by using an argument, in essence, similar to that for Proposition 4. The second mutual information term in (13) is "discrete". We lower bound $I(a^{nR}; b^n)$ by the generalized capacity of the *lossy coding channel*, which is lower bounded by $R_I(p,D)$. In proving (13), we establish the duality between a generalized-channel coding and lossy source coding which is the most interesting part of the technical proof.

A. Proof of (11)

This can be done by a fairly straightforward information theoretic argument. The output of the encoder $a^{nR} \in \{0,1\}^{nR}$, so the entropy of the random variable is upper bounded by

$$H(a^{nR}) \leq nR \tag{14}$$

Notice that a^{nR} is a a function of x^n, i.e. a function of s^n and b^n, so

$$H(a^{nR}) = H(a^{nR}) - H(a^{nR}|s^n, b^n) \tag{15}$$

Combining (14) and (15), and notice that $b^n \perp s^n$, we have:

$$\begin{aligned} nR &\geq H(a^{nR}) - H(a^{nR}|s^n, b^n) \\ &= I(a^{nR}; s^n, b^n) \\ &= I(a^{nR}; b^n) + I(a^{nR}; s^n|b^n) \end{aligned} \tag{16}$$

where (16) is true by the chain rule for mutual information [5].

B. Proof of (12)

In [4], we managed to prove a stronger inequality:

$$\frac{I(a^{nR}; s^n|b^n)}{n} \geq pR(D-(1-p)E[\hat{x}^2|b=0], N(0,p)).$$

But here we only give the proof for the weaker inequality in (12). The idea is quite similar to the proof of the lower bound in Proposition 4. Suppose that a genie reveals the *digital* part of the source, b^n, to the decoder. So for the zero part of the source, the decoder can recover them exactly. But for the non-zero part of the source, the decoder still needs to keep the *total* distortion under nD. Typically, there are roughly np non-zero entries. We denote the non zero part by $\tilde{s}^M$, the corresponding reconstructions by $\widehat{\tilde{s}}^M$, where $M \sim np$, hence $I(\tilde{s}^M; a^{nR}) \leq I(a^{nR}; s^n|b^n)$. So now we have the following Markov Chain: $\tilde{s}^M \to s^n \to a^{nR} \to \hat{x}^n \to \widehat{\tilde{s}}^M$, hence by data processing inequality:

$$I(\tilde{s}^M; \widehat{\tilde{s}}^M) \leq I(\tilde{s}^M; a^{nR}) \leq I(a^{nR}; s^n|b^n). \tag{17}$$

Note that the non-zero part $\tilde{s}^M$ is of roughly $M \sim np$ i.i.d. $N(0,1)$ random variables, and the total distortion between $\tilde{s}^M$ and $\widehat{\tilde{s}}^M$ is within nD, hence from the rate distortion theory for Gaussian random variables, we know that $I(\tilde{s}^M; \widehat{\tilde{s}}^M) \geq MR(\frac{D}{p}, N(0,1)) = MR(D, N(0,p)) \sim npR(D, N(0,p))$. Combining this with (17), we get the desired inequality.

C. Proof of (13)

Inspired by [1], we lower bound the mutual information between b^n and a^{nR}, $I(a^{nR}; b^n)$, by a generalized capacity of an imaginary with-memory channel as shown in Fig. 4. Comparing this figure with Fig. 1, we see that the Bernoulli sequence b^n is replaced by a channel input to the "lossy coding channel". So the constraint on the channel encoder F_n is that the codewords c_m obey the Berounlli-p distribution for m that is uniformly distributed on $\{1, 2, ...2^{n\tilde{R}}\}$. Now we send c_m to be multiplied by the Gaussian sequence s^n and send the resulted Gaussian Bernoulli sequence to the lossy coding system. There are two things we can do with the output a^{nR}. First, the lossy source decoder can reconstruct $\hat{x}^n$ where the distortion D constraint is satisfied in both the average and strong senses. Secondly, a channel decoder can try to recover the message m. This can only be done if the rate of the channel codebook $\tilde{R}$ is not higher than the "lossy channel capacity".

There are two steps left for lower bounding $I(a^{nR}; b^n)$. First we prove that the lossy channel capacity is indeed upper bounded by $I(a^{nR}; b^n)$ and secondly we give *a* lower bound on the channel capacity. The result is inequality (13) where $R_I(p,D)$ is defined in Theorem 1. Due to page limit, we leave the details in [4].

IV. Conclusions and Future Work

In this paper we study the rate distortion function for Bernoulli-Gaussian sequences which is a reasonable probabilistic model for sparse signals. The main result is a non trivial lower bound on the rate distortion function. The improvement over the trivial lower bound is $\sim p\log_2\frac{1}{p}$ in the low distortion regime. We also show that the gap between the trivial lower bound and an upper bound is at most $H(p)$. This is significant since $H(p)$ and $p\log_2\frac{1}{p}$ are roughly equal for small p. To derive this new lower bound, we develop a new technique to lower bound part of the rate distortion function through a randomized lossy coding channel. To further narrow the gap between the lower bound and the upper bounds, we need to develop a more sophisticated upper bounding scheme. This is left for future work.

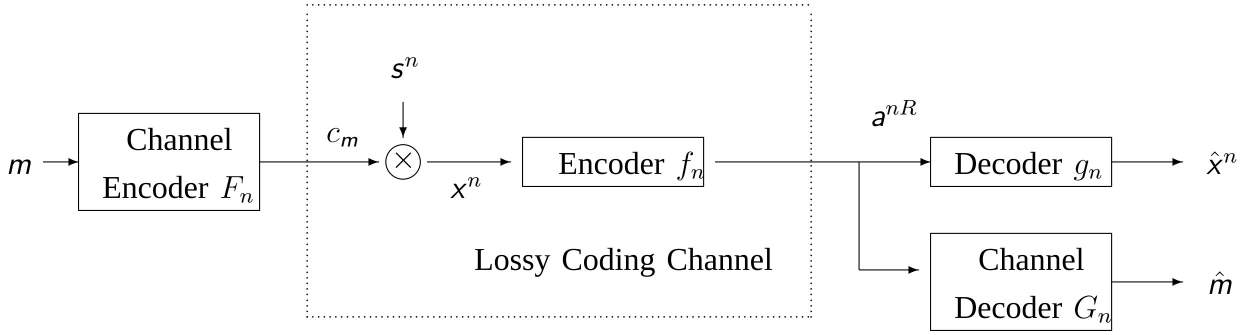

Fig. 4. A “lossy coding” channel derived from the lossy coding system for Bernoulli-Gaussian sequence $x^n = b^n \times s^n$,

Acknowledgments

The author thanks Marcelo Weinberger for introducing the problem, Erik Ordentlich for pointing out [9] and other members of the information theory group at Hewlett-Packard Labs, Palo, Alto for helpful discussions along the way.

References

[1] Mukul Agarwal, Anant Sahai, and Sanjoy Mitter. Coding into a source: a direct inverse rate-distortion theorem. *Allerton Conference*, pages 569–578, 2006.

[2] Toby Berger. *Rate Distortion Theory: A mathematical basis for data compression*. Prentice-Hall, 1971.

[3] Emmanuel Candès and Terence Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52:5406 – 5425, 2006.

[4] Cheng Chang. On the rate distortion function of bernoulli gaussian sequences. *HP Labs Tech Report*, HPL-2009-19, 2009. http://www.hpl.hp.com/techreports/2009/HPL-2009-19.pdf.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons Inc., New York, 1991.

[6] David Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289 – 1306, 2006.

[7] Alyson K. Fletcher, Sundeep Rangan, and Vivek K. Goyal. On the rate-distortion performance of compressed sensing. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, pages 885–888, 2007.

[8] Alyson K. Fletcher, Sundeep Rangan, Vivek K. Goyal, and Kannan Ramchandran. Denoising by sparse approximation: Error bounds based on rate-distortion theory. *EURASIP Journal on Applied Signal Processing*, pages 1–19, 2006.

[9] Vivek K. Goyal, Alyson K. Fletcher, and Sundeep Rangan. Compressive sampling and lossy compression. *IEEE Signal Processing Magazine*, 25:48 – 56, 2008.

[10] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948.

Appendix

A. *Proof of Proposition 5*

Notice that

$$R_I(D,p) = \max_{L\geq 0}\{\min_{U\geq L, r\in[0,1-p]:T(L,U,r)\leq D} h(L,U,r)\},$$

so for all $L \geq 0$, we define the feasible region for (U,r) as $\{(U,r)|U \geq L, r \in [0,1-p], T(L,U,r) \leq D\}$. This feasible region is bigger for larger D, hence the minimum of $h(L,U,r)$ is smaller for larger D (bigger feasible region). This is true for all $L \leq 0$, hence $R_I(D,p)$ is decreasing with D. □

B. *Proof of Proposition 6*

Same as in the proof of Proposition 5, for $L > 0$, we define a feasible region for (U,r) as: $\{(U,r)|U \geq L \geq 0, r \in [0,1-p], T(L,U,r) \leq D\}$.

We first show that as $D \to 0$, for a properly chosen L, the feasible region for (U,r) converges (shrinks) to a single point $(0,0)$. Details as follows, we pick a positive $L \ll 1$, but $L^2 \gg D$, say $L = D^{0.3}$. By the definition of $T(L,U,r)$ in (10) and the fact that $T(L,U,r) \leq D$ for any point (U,r) in the feasible region, it is obvious that $rL^2 \leq D$, hence $r \leq \frac{D}{L^2} = D^{0.4}$. So r converges to zero as D goes to zero. We now show that as D and $L = D^{0.4}$ go to zero, for any feasible point (U,r), U converges to zero. In light of the distortion constraint and that L is picked to be $D^{0.3}$, also the obvious inequality that $-2sL \geq -\frac{s^2}{4} - 4L^2$ for all s and L:

$$\begin{aligned}\frac{D}{2p} &\geq \int_L^U (s-L)^2 \frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} ds \geq \int_L^U (\frac{3s^2}{4} - 3L^2)\frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} ds \\ &= \int_{D^{0.3}}^U (\frac{3s^2}{4} - 3D^{0.6})\frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} ds\end{aligned}$$

Then by moving terms around we get:

$$\begin{aligned}\int_{D^{0.3}}^U \frac{3s^2}{4}\frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} ds &\leq \frac{D}{2p} + \int_{D^{0.3}}^U 3D^{0.6}\frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} ds \\ &\leq \frac{D}{2p} + 3D^{0.6}\end{aligned}$$

The last inequality is true because the integral of a Gaussian pdf is upper bounded by 1. Now let $D \to 0$ on both sides, the right hand side is obviously 0, the left hand side is zero if and only if $U \to 0$ as D goes to zero. Hence we just proved that if we pick $L = D^{0.3}$ together with $D \to 0$, then both U and r goes to zero if $T(L,U,r) \leq D$. This means that:

$\lim_{D\to 0} R_I(D,p)$

$$\geq \lim_{r,U\to 0}(p \times \Pr(|s| > U) + r)\mathcal{D}(\frac{p \times \Pr(|s| > U)}{p \times \Pr(|s| > U) + r}\|p)$$

$= p\mathcal{D}(1\|p) = p\log_2 \frac{1}{p}$ □