On the rate distortion function of Bernoulli Gaussian sequences

Cheng Chang

HP Laboratories
HPL- 2009-19

Keyword(s):
Rate distortion theory, duality between channel and source coding, sparse signals

Abstract:

In this paper, we study the rate distortion function of the i.i.d sequence of multiplications of a Bernoulli *p* random variable and a gaussian random variable ~ *N* (0,1). We use a new technique in the derivation of the lower bound in which we establish the duality between channel coding and lossy source coding in the strong sense. We improve the lower bound on the rate distortion function over the best known lower bound by $p\log_2\frac{1}{p}$ if distortion *D* is small. This has some interesting implications on sparse signals where *p* is small since the known gap between the lower and upper bound is *H* (*p*). This improvement in the lower bound shows that the lower and upper bounds are almost identical for sparse signals with small distortion because $\lim_{p\to 0}\frac{p\log_2\frac{1}{p}}{H(p)}=1$.

External Posting Date: February 6, 2009 [Fulltext]
Internal Posting Date: February 6, 2009 [Fulltext]

Approved for External Publication

Submitted to IEEE Transactions on Information Theory, Allerton Conference 2009

On the rate distortion function of Bernoulli Gaussian sequences

Cheng Chang

Abstract

In this paper, we study the rate distortion function of the i.i.d sequence of multiplications of a Bernoulli p random variable and a gaussian random variable $\sim N(0,1)$. We use a new technique in the derivation of the lower bound in which we establish the duality between channel coding and lossy source coding in the strong sense. We improve the lower bound on the rate distortion function over the best known lower bound by $p\log_2\frac{1}{p}$ if distortion D is small. This has some interesting implications on sparse signals where p is small since the known gap between the lower and upper bound is $H(p)$. This improvement in the lower bound shows that the lower and upper bounds are almost identical for sparse signals with small distortion because $\lim\limits_{p\to 0}\frac{p\log_2\frac{1}{p}}{H(p)}=1$.

I. Bernoulli-Gaussian model and some obvious bounds on its rate distortion functions

Notations: in this paper we use x, y, u for random variables and x, y, u for the realization of the random variables or constants. We denote by $\overset{\mathsf{x}}{\Pr}(A)$ the probability of event A under measure x. We use bit and $\log_2$ in this paper.

Consider a sequence of signals $\mathsf{x}_1, \mathsf{x}_2, ...\mathsf{x}_n$, where x_i's are zero most of the time. When x_i is non-zero, it is an arbitrary real number. In the signal processing literature, the signals x^n is called sparse if most of them are zero. In their seminal work on compressive sensing [4] and [7], Candès, Tao and Donoho show that, to exactly reconstruct the sparse signals x^n, only a fraction of n measurements are needed. 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 arbitrary real numbers without any statistical distribution assigned to them. Furthermore the compressed sensing system tries to recover the signals x^n losslessly without distortion of the reconstructed signals. These assumptions are not completely valid if the source statistics are known to the coding system, more importantly, if the goal of the sensing system is only to recover the data within a certain distortion. In the recent work by Fletcher etc. [9], [8], [10], the . What is lacking in the previous study of this problem is a systematic study of the information theoretic bounds on the rate distortion functions of the sources. In this paper, we give both lower and upper bounds on the rate distortion functions.

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 random variable and a discrete random variable. This adds to the difficulties to study the rate distortion functions of this random variable. 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 $\Xi(p,\sigma^2)$. It will be clear soon in Proposition 1 that we only need to study the rate distortion functions for $\mathsf{s}\sim N(0,1)$, i.e. the rate distortion functions for $\Xi(p,1)$. First, we review the definition of rate distortion functions in both the average distortion and strong distortion sense.

B. Review of the rate distortion theory

In the standard setup of rate distortion theory, the encoder maps n i.i.d. random variables $\mathsf{x}^n\in\mathcal{X}^n$, $\mathsf{x}\sim p_{\mathsf{x}}$, into nR bits and then the decoder reconstruct the original signal within a certain distortion. The encoder and decoder are denoted by f_n and g_n respectively:

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

Cheng Chang is with Hewlett-Packard Labs, Palo Alto. Email: cchang@eecs.berkeley.edu

and the distortion is defined as $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$.

Definition 2: Rate distortion function ([5], pg. 341): the rate distortion function $R(D)$ is the infinimum of rates R, such that (R, D) is in the rate distortion region of the source for a give distortion D. Where the rate distortion region is the closure of achievable rate distortion pairs (R, D) defined as follows. (R, D) is said to be achievable in the expected distortion sense if there exists a sequence of $(2^{nR}, n)$ rate codes (f_n, g_n), such that

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

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

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

where, in this paper, the distortion function $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$.

It turns out that the rate distortions function for both the average distortion and the strong distortion are the same for discrete random variables Chapter 13.6 [5]. We can generalize this result easily to continuous random variables whose variance is finite and the probability density function satisfies the usual regularity conditions. The proof can be carried out by quantizing the probability density function and then by using the proof for discrete random variables in [5]. A somewhat detailed sketch of how this works is in Appendix A.

A good lossy coding system in the strong sense is not *necessarily* good in the expected distortion sense. Considering the following example, a good lossy coder can miss the distortion constraint for a subset $\Upsilon_n \subseteq \mathcal{R}^n$ with asymptotically 0 measure, $\lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}(\Upsilon_n) = 0$. However the good lossy coder can *intentionally* make the distortion on Υ_n no smaller than $\frac{2D}{\overset{\mathsf{x}}{\Pr}(\Upsilon_n)}$, hence the expected distortion is at least $2D$.

However it is easy to see that given a good lossy coding system in the strong sense, we can easily make it also good in the expected sense if the mean and variance of x are finite. We sketch the proof in Appendix B. So from now on, when we say a lossy coding system is good in the strong sense, that implies that the system is also good in the expected distortion sense.

The following lemma characterizes the rate distortion function $R(D)$.

Lemma 1: Rate distortion theorem [12]:

$$R(D) = \min_{p_{\hat{\mathsf{x}}|\mathsf{x}} : \sum\limits_{x,\hat{x}} p_{\mathsf{x}}(x) p_{\hat{\mathsf{x}}|\mathsf{x}}(\hat{x}|x) d(x,\hat{x}) \le D} I(\mathsf{x}; \hat{\mathsf{x}}). \tag{3}$$

Corollary 1: Rate distortion theorem for Gaussian random variables [2]: for random variable $\mathsf{x} \sim N(0, \sigma^2)$, the rate distortion function is:

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

It is also shown that with the same variance and squared distortion measure, Gaussian random variables requires the most bits to be described. Both lower and upper bounds are given in Exercise 8 on Pg. 370 [5]. The proof can be found in [2].

Corollary 2: Rate distortion bounds for continuous random variables under square distortion measure (Exercise 8 on pg. 372 [5]): the rate distortion function $R(D)$ can be bounded as:

$$h(\mathsf{x}) - \frac{D}{2}\log(2\pi e) \le R(D) \le \max\{\frac{1}{2}\log\frac{\sigma^2}{D}, 0\} \tag{5}$$

The lower bound in Corollary 2 is known as the Shannon lower bound in the literature [5].

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

The main goal of this paper is to derive an upper and a lower bound on the rate distortion function $R(D)$ of the Bernoulli-Gaussian random variable $\Xi(p, \sigma^2)$. We denote this quantity by $R(D, \Xi(p, \sigma^2))$. We summarize some obvious properties of $R(D, \Xi(p, \sigma^2))$ in the following four propositions. The proof is in Appendix C.

First we explain why we only need to study $R(D, \Xi(p, 1))$. We 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 $R(D, \Xi(p, 1))$, simply written as $R(D, p)$. Now we give three obvious bounds on the rate distortion function $R(D, p)$.

Proposition 2: Upper bound 1 on $R(D, p)$:

$$R(D, p) \leq H(p) + pR(\frac{D}{p}, N(0, 1)) = H(p) + pR(D, N(0, p)) \tag{6}$$

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

Proposition 3: Upper bound 2 on $R(D, p)$:

$$R(D, p) \leq R(D, N(0, p)) \tag{7}$$

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

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

We give a conceptually clear explanation of these three bounds. In Proposition 2, we construct a very simple coding system that first losslessly describe the locations of the non-zero elements of $x^n \sim \Xi(p, 1)$, then lossily describe the value of these non-zero elements using a Gaussian lossy coder. In Proposition 3, we prove it by using the well known fact that for continuous random variables, with the same variance and distortion measure, Gaussian sequences require the highest rate. The difficulty is that $\Xi(p, 1)$ is not a continuous random variable. We approximate it by a sequence of continuous random variables whose rate distortion functions converge to that of $\Xi(p, 1)$. In the proof of 4, we reduce a Bernoulli-Gaussian sequence to a Gaussian sequence by letting the decoder know the non-zero locations for free and derive *a* lower bound of $R(D, p)$ from the Gaussian rate distortion function.

The more rigorous proofs of these bounds are in Appendix C. It is non trivial to bound the rate distortion function of one random variable x by the rate distortion function of another random variable y. To show that $R(D, x) \leq R(D, y)$, the technique we use in the proofs for the above four propositions is to *construct* a good lossy coding system for x from a good lossy coding system for y under the same rate-distortion constraint R and D.

Among the three bounds described in Proposition 2, 3 and 4, we find the lower bound 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 negative infinity. This paper is focused on deriving a more information-theoretically interesting lower bound on $R(D, p)$. In the next several sections, we investigate the lower bound problem. As a simple corollary of this new lower bound, we give a close form lower bound on the rate distortion function in VII that improves the previous known result by $p \log_2 \frac{1}{p}$ in the high resolution regime ($\frac{D}{p} \ll 1$).

II. An improved lower bound on $R(D, p)$

First, we reiterate the definition of a strong lossy source coding system for a Bernoulli-Gaussian sequence $x^n \sim \Xi(0, 1)$ where $x = b \times s$ and b is a Bernoulli-p random variable while $s \sim N(0, 1)$ is a Gaussian random variable. A (R, D) encoder-decoder sequence f_n, g_n does the following,

$$f_n : \mathcal{R}^n \to \{0, 1\}^{nR}, \quad f_n(x^n) = a^{nR} \quad \text{and} \quad g_n : \{0, 1\}^{nR} \to \mathcal{R}^n, \quad g_n(a^{nR}) = \hat{x}^n$$

from the definition of the rate distortion function in strong sense defined in (2), we have for all $\delta_1 > 0$:

$$\overset{x}{\Pr}\left(d(x^n, \hat{x}^n) \geq D + \delta_1\right) = \overset{x}{\Pr}\left(d(x^n, g_n(f_n(x^n))) \geq D + \delta_1\right) = e_n(\delta_1) \text{ and } \lim_{n \to \infty} e_n(\delta_1) = 0. \tag{9}$$

Recall that we can have a good lossy coder in both the strong sense and the expected distortion sense according to the discussions in Appendix B. So we assume the good coding system here f_n, g_n is good in both senses.

$$\text{So let } E_x\left(d(x^n, \hat{x}^n)\right) = E_x\left(d(x^n, g_n(f_n(x^n)))\right) = D + \varsigma_n, \text{ then } \lim_{n\to\infty} \varsigma_n = 0. \tag{10}$$

Notice that $x^n = b^n \times s^n$, where the multiplication $\times$ here is done entry by entry, so that if $b_i = 0$, the value of s_i does not have any impact on x^n. The output of the encoder f_n is a random variable that is a function of the sequence x^n, we write the output as $a^{nR} = f_n(x^n)$. our investigation of the rate distortion function relies on the properties of the encoder output a^{nR}.

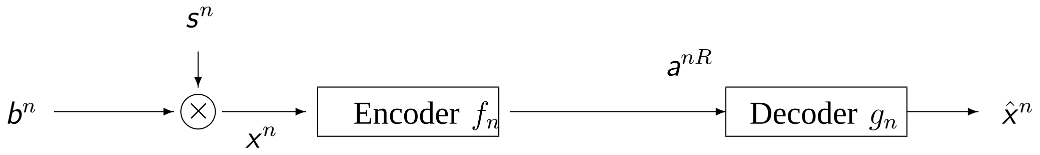

Fig. 1. A lossy source coding system for Bernoulli-Gaussian sequence $x^n = b^n \times s^n$

In Proposition 4, the lower bound is derived by letting a genie tell the decoder the non-zero positions of the Bernoulli-Gaussian sequence, i.e. the b^n part of $x^n = b^n \times s^n$, and the rate is only counted for the lossy source coding of the non-zero Gaussian subsequence $\tilde{s}^{1(b^n)}$, where $1(b^n)$ is the number of 1's in sequence b^n and $\tilde{s}_i = s_{l_i}$ if $b_{l_i} = 1$, $i = 1, 2, ..., 1(b^n)$. To tighten the lower bound in Proposition 4, we need to drop the genie who let the decoder know the entirety of b^n. In the following several sections, we attempt to tighten up the lower bound by investigating the information about b^n that *has to* be transmitted to the decoder.

First we summarize our main result in the following theorem.

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

$$\begin{aligned} & R(D,p) \geq pR(D, N(0,p)) + \tilde{R} \\ \text{where} \qquad & \tilde{R} = \max_{L\geq 0}\{ \min_{U\geq L, r\in[0,1-p]: T_1(L,U,r)\leq D} h(L,U,r)\} \end{aligned} \tag{11}$$

$$\text{where } h(L,U,r) = \{ \begin{array}{ll} (p\times \Pr(|s|>U)+r) D(\frac{p\times\Pr(|s|>U)}{p\times\Pr(|s|>U)+r}\|p) & \text{, if } \frac{p\times\Pr(|s|>U)}{p\times\Pr(|s|>U)+r} \geq p \\ 0 & \text{, if } \frac{p\times\Pr(|s|>U)}{p\times\Pr(|s|>U)+r} < p \end{array}$$

$s \sim N(0,1)$ is a Gaussian random variable.

Proof: The theorem is a corollary of the Lemma 2, 3, 4 and 5:

$$\begin{aligned} R(D,p) &\geq \frac{I(a^{nR}; s^n|b^n) + I(a^{nR}; b^n)}{n} && (12) \\ &\geq pR(D-(1-p)E[\hat{x}^2|b=0], N(0,p)) + \frac{I(a^{nR}; b^n)}{n} && (13) \\ &\geq pR(D-(1-p)E[\hat{x}^2|b=0], N(0,p)) + \tilde{R} && (14) \\ &\geq pR(D, N(0,p)) + \tilde{R} && (15) \end{aligned}$$

(12) is proved in Lemma 2. (13) is proved in Lemma 3. (14) is proved in Lemma 4 and 5, $\tilde{R}$ is defined in (11). (15) follows that rate distortion function for Gaussian random variables $R(D, N(0,p))$ is monotonically decreasing with D. ■

There are four parts in our investigation. First in Section III, we lower bound the number of bits nR by the sum of two mutual information terms. The first one is the conditional mutual information between the output of the encoder a^{nR} and the Gaussian sequence s^n given the Bernoulli sequence b^n: $I(a^{nR}; s^n|b^n)$. The second is the mutual information between the output of the encoder a^{nR} and the Bernoulli sequence b^n: $I(a^{nR};b^n)$. Then in Section IV we lower bound $I(a^{nR};s^n|b^n)$ by using a simple argument similar to that in Proposition 4. In Section V, we lower bound $I(a^{nR};b^n)$ by the capacity of the *lossy coding channel*, while the capacity of the channel is unspecified. In Section VI, we give *a* lower bound of the channel capacity by using a random coding argument. Finally in Theorem 1, we combine these bounds together to give a lower bound on the rate distortion function $R(D,p)$ for the Bernoulli-Gaussian random sequence $\Xi(p,1)$ under distortion constraint D. The investigation spans the next four sections in this paper.

III. First step: lower bounding nR by the sum of two mutual information $I(a^{nR};b^n)+I(a^{nR};s^n|b^n)$

First we have the following simple lemma that tells us that the rate is lower bounded by the sum of two mutual information terms $I(a^{nR};b^n)+I(a^{nR};s^n)$ where a^{nR} is the output of the lossy encoder and b^n and s^n are the Bernoulli sequence and the Gaussian sequence that generate the Bernoulli-Gaussian $x^n \sim \Xi(p,1)$.

Lemma 2: For a lossy coding system shown in Figure 1, the rate of the lossy source coding system can be lower bounded as follows:

$$nR \geq I(a^{nR};b^n)+I(a^{nR};s^n|b^n)$$

Proof: The output of the encoder $a^{nR}\in\{0,1\}^{nR}$, so the entropy of the random variable is upper bounded by

$$H(a^{nR}) \leq nR \tag{16}$$

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

$$H(a^{nR}) = H(a^{nR}) - H(a^{nR}|s^n,b^n) \tag{17}$$

Combining (16) and (17), and notice that $b^n \perp s^n$, we have:

$$\begin{aligned} nR &\geq H(a^{nR}) - H(a^{nR}|s^n,b^n) \\ &= I(a^{nR};s^n,b^n) \\ &= I(a^{nR};b^n)+I(a^{nR};s^n|b^n) \end{aligned} \tag{18}$$

where (18) is true by the chain rule for mutual information [5]. □

IV. Lower bounding $I(a^{nR};s^n|b^n)$, Proposition 4 revisited

In this section we lower bound the conditional mutual information term $I(a^{nR};s^n|b^n)$ in the lower bound of nR (18). From Proposition 4, we know that letting a genie tell the non-zero locations of x^n to the decoder, the coding system still needs at least $npR(D,N(0,p))$ bits to describe the values of the non-zero entries of x^n. In the proof of Proposition 4, like the proofs for other propositions in Section I-A, we use the lossy source coding system for the Bernoulli-Gaussian sequences to construct a lossy source coding system for a random sequence with known rate distortion functions.

The proof here, however is trickier in the sense that we are not bounding the rate distortion function $R(D,p)$, instead we only bound the conditional mutual information $I(a^{nR};s^n|b^n)$ which is part of the rate. Hence we cannot *construct* a lossy coder for sequence with known rate distortion using the lossy coder for the Bernoulli-Gaussian sequence. Instead, we use the classical technique in [5].

Lemma 3: Lower bound on $I(a^{nR};s^n|b^n)$

$$I(a^{nR};s^n|b^n) \geq npR(D-(1-p)E[\hat{x}^2|b=0],N(0,p)). \tag{19}$$

where

$$\begin{aligned} E[\hat{x}^2|b=0] &= \frac{1}{n}\sum_{i=1}^{n}E[\hat{x}_i^2|b_i=0] \\ &= \frac{1}{n}\left(\sum_{b^n}\Pr(b^n=b^n)\sum_{i:b_i=0}\left(E[\hat{x}_i^2|b^n=b^n]\right)\right) \end{aligned} \tag{20}$$

Proof: The proof is similar to the lower bound proof for Gaussian rate distortion function on Page 345 [5]. First, notice that the estimate $\hat{\mathsf{x}}^n = g_n(\mathsf{a}^{nR})$ is a function of a^n. And the $\mathsf{a}^{nR} = f_n(\mathsf{x}^n) = f_n(\mathsf{b}^n \times \mathsf{s}^n)$. Hence we have the following Markov Chain:

$$\mathsf{b}^n \times \mathsf{s}^n \to \mathsf{a}^{nR} \to \hat{\mathsf{x}}^n \tag{21}$$

From the data processing theorem [5], we know that $I(\mathsf{a}^{nR}; \mathsf{s}^n | \mathsf{b}^n) \geq I(\hat{\mathsf{x}}^n; \mathsf{s}^n | \mathsf{b}^n)$. For a binary sequence $b^n \in \{0,1\}^n$, let $1(b^n) = \sum\limits_{i=1}^{n} b_i$ be the number of 1's in b^n. $\mathsf{b}_i \in \{0,1\}$, so if $\mathsf{b}_i = 0$ then $\hat{\mathsf{x}}^n$ and s_i are independent because in that case $\mathsf{x}_i = \mathsf{b}_i \times \mathsf{s}_i = 0$ and s^n is i.i.d and $\hat{\mathsf{x}}^n$ is a deterministic function of x^n. Write $i_1, ..., i_{1(b^n)}$ the non-zero positions of b^n, and let $\mathcal{I}(b^n) = \{i_1, ..., i_{1(b^n)}\}$, then

$$I(\hat{\mathsf{x}}^n; \mathsf{s}^n | \mathsf{b}^n = b^n) = I(\hat{\mathsf{x}}^n; \mathsf{s}_{i_1}, ..., \mathsf{s}_{i_{1(b^n)}} | \mathsf{b}^n = b^n) = I(\hat{\mathsf{x}}^n; \mathsf{s}_{i_1, ..., i_{1(b^n)}}). \tag{22}$$

Define the ϵ_1-strong typical set $B^n_{\epsilon_1}$ for binary sequences:

$$B^n_{\epsilon_1} \triangleq \{b^n \in \{0,1\}^n : |\frac{1(b^n)}{n} - p| \leq \epsilon_1\}.$$

From the AEP [5], let $\Pr(\mathsf{b}^n \notin B^n_{\epsilon_1}) = \upsilon_n$:

$$\lim_{n\to\infty} \upsilon_n = 0 \tag{23}$$

Now we have:

$$\begin{aligned}
I(\mathsf{a}^{nR}; \mathsf{s}^n | \mathsf{b}^n) &\geq I(\hat{\mathsf{x}}^n; \mathsf{s}^n | \mathsf{b}^n) \\
&= \sum_{b^n \in \{0,1\}^n} \Pr(\mathsf{b}^n = b^n) I(\hat{\mathsf{x}}^n; \mathsf{s}^n | \mathsf{b}^n = b^n) &(24)\\
&= \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) I(\hat{\mathsf{x}}^n; \mathsf{s}^n | \mathsf{b}^n = b^n) &(25)\\
&= \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) I(\hat{\mathsf{x}}^n; \mathsf{s}_{i_1}, ..., \mathsf{s}_{i_{1(b^n)}} | \mathsf{b}^n = b^n) &(26)\\
&= \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) \left(H(\mathsf{s}_{i_1,...,i_{1(b^n)}} | \mathsf{b}^n = b^n) - H(\mathsf{s}_{i_1,...,i_{1(b^n)}} | \hat{\mathsf{x}}^n, \mathsf{b}^n = b^n)\right) \\
&\geq \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) \left(\sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j}) - \sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j} | \hat{\mathsf{x}}^n, \mathsf{b}^n = b^n)\right) &(27)\\
&= \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) \left(\sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j}) - \sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j} | \hat{\mathsf{x}}^n, \mathsf{b}^n = b^n)\right) \\
&\geq \sum_{b^n \in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n = b^n) \left(\sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j}) - \sum_{j=1}^{1(b^n)} H(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j} | \mathsf{b}^n = b^n)\right) &(28)
\end{aligned}$$

(24) follows the definition of conditional mutual information, (25) is true because mutual information is non-negative and (26) follows (22). (27) is true because s^n is i.i.d and independent of b^n. The rest are obvious. $\mathsf{s}_i \sim N(0,1)$, so $H(\mathsf{s}_i) = \frac{1}{2}\log(2\pi e)$. According Theorem 9.6.5 in [5], Gaussian random variables maximize the entropy over all distributions with thes ame covariance, so:

$$H(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j} | \mathsf{b}^n = b^n) \leq H(N(0, E[(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n]) = \frac{1}{2}\log(2\pi e E[(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n]).$$

Now (28) becomes:

$$
\begin{aligned}
I(\mathsf{a}^{nR};\mathsf{s}^n|\mathsf{b}^n) &\geq \sum_{b^n\in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n=b^n)\left(\sum_{j=1}^{1(b^n)}\frac{1}{2}\log(2\pi e)-\sum_{j=1}^{1(b^n)}\frac{1}{2}\log(2\pi e E[(\mathsf{s}_{i_j}-\hat{\mathsf{x}}_{i_j})^2|\mathsf{b}^n=b^n])\right)\\
&= \sum_{b^n\in B^n_{\epsilon_1}} \Pr(\mathsf{b}^n=b^n)\left(-\sum_{j=1}^{1(b^n)}\frac{1}{2}\log(E[(\mathsf{s}_{i_j}-\hat{\mathsf{x}}_{i_j})^2|\mathsf{b}^n=b^n])\right)\\
&= -\frac{1}{2}\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)} \Pr(\mathsf{b}^n=b^n)\left(\log(E[(\mathsf{s}_{i_j}-\hat{\mathsf{x}}_{i_j})^2|\mathsf{b}^n=b^n])\right)\\
&= -\frac{1}{2}\left(\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)}\frac{\Pr(\mathsf{b}^n=b^n)}{\sum\limits_{b^n\in B^n_{\epsilon_1}}\sum\limits_{j=1}^{1(b^n)}\Pr(\mathsf{b}^n=b^n)}\log(E[(\mathsf{s}_{i_j}-\hat{\mathsf{x}}_{i_j})^2|\mathsf{b}^n=b^n])\right)\times\\
&\qquad\left(\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)}\Pr(\mathsf{b}^n=b^n)\right)\\
&\geq -\frac{1}{2}\log\left(\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)}\frac{\Pr(\mathsf{b}^n=b^n)}{\sum\limits_{b^n\in B^n_{\epsilon_1}}\sum\limits_{j=1}^{1(b^n)}\Pr(\mathsf{b}^n=b^n)}(E[(\mathsf{s}_{i_j}-\hat{\mathsf{x}}_{i_j})^2|\mathsf{b}^n=b^n])\right)\times\\
&\qquad\left(\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)}\Pr(\mathsf{b}^n=b^n)\right) \qquad (29)
\end{aligned}
$$

(29) follows the fact that $-\log(\cdot)$ is convex $\bigcup$. We bound the two terms as follows, first:

$$
\begin{aligned}
\left(\sum_{b^n\in B^n_{\epsilon_1}}\sum_{j=1}^{1(b^n)}\Pr(\mathsf{b}^n=b^n)\right) &= \left(\sum_{b^n\in B^n_{\epsilon_1}}1(b^n)\Pr(\mathsf{b}^n=b^n)\right)\\
&\geq \left(\sum_{b^n\in B^n_{\epsilon_1}}n(p-\epsilon_1)\Pr(\mathsf{b}^n=b^n)\right)\\
&\geq n(p-\epsilon_1)(1-\upsilon_n) \qquad (30)
\end{aligned}
$$

Before bounding the other term, we have the following observation:

$$
\begin{aligned}
& \left(\sum_{b^n \in B^n_{\epsilon_1}} \sum_{j=1}^{1(b^n)} \Pr(\mathsf{b}^n = b^n)(E[(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n])\right) \\
= & \left(\sum_{b^n \in B^n_{\epsilon_1}} \sum_{j=1}^{1(b^n)} \Pr(\mathsf{b}^n = b^n)(E[(\mathsf{x}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n])\right) \\
\leq & \left(\sum_{b^n} \sum_{j=1}^{1(b^n)} \Pr(\mathsf{b}^n = b^n)(E[(\mathsf{x}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n])\right) \\
\leq & \; n(D + \varsigma_n) - \left(\sum_{b^n} \sum_{i \notin \mathcal{I}(b^n)} \Pr(\mathsf{b}^n = b^n)(E[(\mathsf{x}_i - \hat{\mathsf{x}}_i)^2 | \mathsf{b}^n = b^n])\right) \qquad (31) \\
\leq & \; n(D + \varsigma_n) - \left(\sum_{b^n} \sum_{i \notin \mathcal{I}(b^n)} \Pr(\mathsf{b}^n = b^n)(E[\hat{\mathsf{x}}_i^2 | \mathsf{b}^n = b^n])\right) \\
= & \; n(D + \varsigma_n) - nE[\hat{\mathsf{x}}^2 | \mathsf{b} = 0] \qquad (32)
\end{aligned}
$$

where $\mathcal{I}(b^n) = \{i_1, ..., i_{1(b^n)}\}$ and $\varsigma_n \to 0$ as n goes to infinity, (31) follows the fact that f_n, g_n is good in the expected distortion sense as well (10). So the first term in (29) can be lower bounded as follows, combining (30) and (32):

$$
-\frac{1}{2}\log\left(\sum_{b^n \in B^n_{\epsilon_1}} \sum_{j=1}^{1(b^n)} \frac{\Pr(\mathsf{b}^n = b^n)}{\sum\limits_{b^n \in B^n_{\epsilon_1}} \sum\limits_{j=1}^{1(b^n)} \Pr(\mathsf{b}^n = b^n)} (E[(\mathsf{s}_{i_j} - \hat{\mathsf{x}}_{i_j})^2 | \mathsf{b}^n = b^n])\right) \geq -\frac{1}{2}\log\left(\frac{(D - E[\hat{\mathsf{x}}^2 | \mathsf{b} = 0] + \varsigma_n)}{(p - \epsilon_1)(1 - \upsilon_n)}\right) \quad (33)
$$

first notice that we are lower bounding a conditional mutual information $I(\mathsf{a}^{nR}; \mathsf{s}^n | \mathsf{b}^n)$ which is non-negative, so we assume the first term being positive or else we lower bound the conditional mutual information by 0, so substituting (30) and (33) into (29), we have:

$$
I(\mathsf{a}^{nR}; \mathsf{s}^n | \mathsf{b}^n) \quad \geq \quad n(p - \epsilon_1)(1 - \upsilon_n) \max\{0, \log\left(\frac{(p - \epsilon_1)(1 - \upsilon_n)}{(D - (1-p)E[\hat{\mathsf{x}}^2 | \mathsf{b} = 0] + \varsigma_n)}\right)\} \qquad (34)
$$

Notice that ϵ_1 is an arbitrary positive real number, and both υ_n and ς_n goes to zero as n goes to infinity, so we just showed that

$$
I(\mathsf{a}^{nR}; \mathsf{s}^n | \mathsf{b}^n) \quad \geq \quad np \times \max\{0, \log\left(\frac{p}{D - (1-p)E[\hat{\mathsf{x}}^2 | \mathsf{b} = 0]}\right)\} = npR(D - (1-p)E[\hat{\mathsf{x}}^2 | \mathsf{b} = 0], N(0, p))
$$

The lemma is proved. □

As a trivial corollary of Lemma 2 and Lemma 3, we have:

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

This also proves Proposition 4.

V. Lower bounding $I(a^{nR}, b^n)$ by the randomized channel capacity of a lossy compressor

In this section we give a lower bound on the mutual information $I(a^{nR}; b^n)$ from a channel capacity perspective. This is partly inspired by the seminal work in [1]. First we have another look at the whole lossy coding system in Figure 1, we single out the binary randomness b^n and make the rest of the system a "lossy coding channel" as shown in Figure 2. The channel input is a binary sequence $b^n \in \{0,1\}^n$, and the channel output is $a^{nR} \in \{0,1\}^{nR}$. What the channel does is to first multiply b^n by a Gaussian random sequence s^n and then send it to a good lossy encoder f_n. The output is the output of the lossy coding encoder f_n.

Notice that this is not a standard communication channel. It is in some sense a arbitrarily varying channel. The constraint on the channel is such that the lossy coder pair f_n, g_n is good in both the strong and expected distortion sense. **The goal in this section is to lower bound the mutual information $I(a^{nR}, b^n)$ by the number bits (channel capacity) that can be reliably communicated across the channel in average over a randomized codebook.**

More interestingly, the input sequence b^n obeys the statistics of a Bernoulli process with non-zero probability p. So it will be soon obvious that we need to investigate the channel capacity for the randomized codebooks where each code word is chosen according to its probability under i.i.d Bernoulli-p.

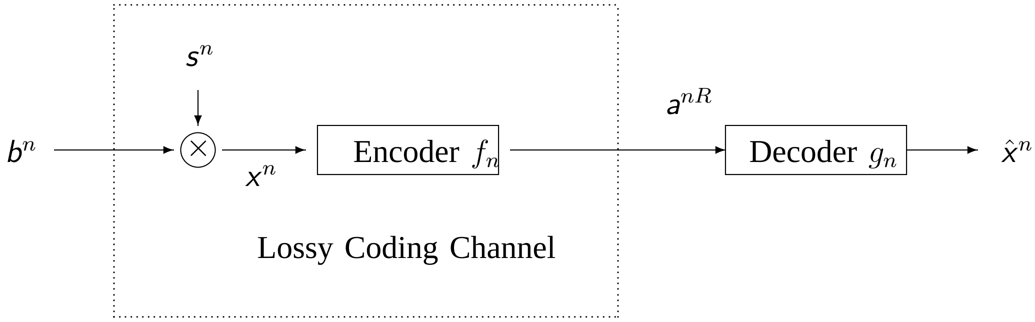

Fig. 2. A "lossy coding" channel derived from the lossy coding system for Bernoulli-Gaussian sequence $x^n = b^n \times s^n$,

As shown in Figure 3, we have a channel coding problem. A message m is a random variable uniformly distributed on $\{1, 2, ..., 2^{nR}\}$. The constraint on the channel encoder F_n is that the code word b^n is chosen for message m with probability

$$p^{1(b^n)}(1-p)^{n-1(b^n)},$$

where $1(b^n)$ is the number of 1's in sequence b^n, this will be explained in details in Definition 3. The constraint on the *lossy coding channel* is such that the estimate of the Bernoulli Gaussian random sequence $x^n = b^n \times s^n$, through the lossy coding system f_n, g_n: $\hat{x}^n$ is within a distortion $D + \delta_1$ of the true sequence x^n with probability 1 for all $\delta_1 > 0$ asymptotically. Before giving the lemma on the lower bound of the mutual information $I(a^{nR}; b^n)$,

we give the following definition of randomized channel capacity for the lossy source channel.

Definition 3: Randomized channel capacity for the lossy source channel is written as $\tilde{R}_p$[1]: let $\mathcal{B}_n = \{0,1\}^n$, let $\mathcal{C}(n)$ be the codebook set of rate $\tilde{R}$: $\mathcal{C}(n) = \mathcal{B}_n^{2^{n\tilde{R}}}$ is the set product of $2^{n\tilde{R}}$ many $\mathcal{B}_n$'s: $\mathcal{B}_n \times \mathcal{B}_n \times ... \times \mathcal{B}_n$, a codebook $C \in \mathcal{C}(n)$, $C = (c_1, c_2, ...c_{2^{n\tilde{R}}})$ is such that the codeword for message m, $m = 1, 2, ...2^{n\tilde{R}}$, is the i'th entry of C: c_m. From the definition $c_m \in \mathcal{B}_n$ for all n. We let C_p be a random variable distributed on $\mathcal{C}(n)$, such that a codebook $C = (c_1, c_2, ...c_{2^{n\tilde{R}}}) \in \mathcal{C}(n)$ is chosen as the codebook, i.e. $C_p = C$ with the following probability:

$$\Pr(C_p = C) = \prod_{m=1}^{2^{n\tilde{R}}} p^{1(c_m)}(1-p)^{n-1(c_m)} \tag{35}$$

[1] Note: in this section we use $\tilde{R}$ to denote the channel capacity of the lossy coding channel. This is not the rate of the lossy coding system R.

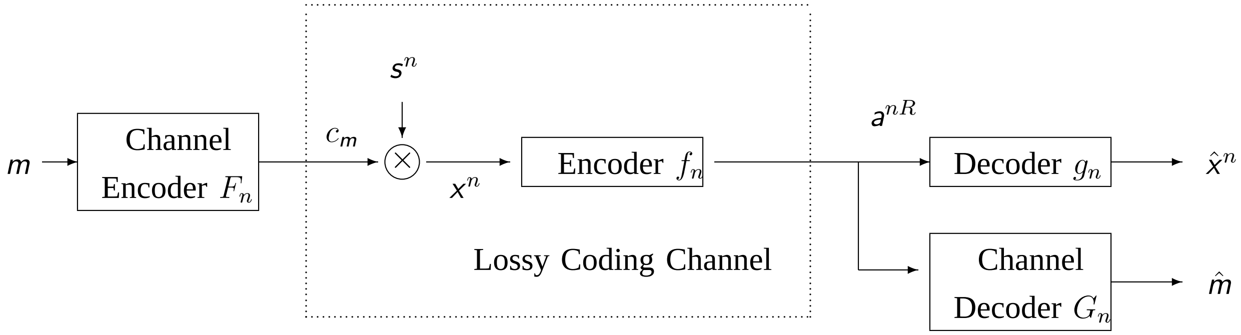

Fig. 3. A channel coding system for the "lossy coding" channel

the average error probability of the randomized coding with uniform distributed C_p is defined as:

$$\begin{aligned} e_{p,n}(\tilde{R}) &= \sum_{C\in\mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(C_p = C)\left(\frac{1}{2^{n\tilde{R}}}\sum_{m=1}^{2^{n\tilde{R}}} \overset{s}{\Pr}(m \neq \hat{m}(a^{nR}(c_m \times s^n)))\right) \\ &= \sum_{C\in\mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(C_p = C)\left(\Pr(m \neq \hat{m}(a^{nR})|C_p = C)\right) \end{aligned} \quad (36)$$

where the error probability is over all codebooks $\mathcal{C}(n) = \mathcal{B}_q^{2^{n\tilde{R}}}$ with distribution defined in (35) and all messages $m \in \{1, 2, ..., 2^{n\tilde{R}}\}$, i.e. the random variable m is uniformly distributed in (36). Notice that in Figure 3, a codebook C is chosen and known to both the encoder and the decoder. The output from the channel encoder is $F_n(m) = c_m$, the output from the lossy encoder is a random sequence $f_n(c_m \times s^n) = a^{nR}(c_m \times s^n)$, and the estimate of m is $\hat{m}(a^{nR}(c_m \times s^n)) = G_n(a^{nR}(c_m \times s^n)))$.

The randomized channel capacity for the lossy coding system f_n, g_n is $\tilde{R}_p$, if for all $\tilde{R} < \tilde{R}_p$, there exists a channel decoder G_n, such that the average error goes to zero as n goes to infinity:

$$\lim_{n\to\infty} e_{p,n}(\tilde{R}) = 0, \text{ equivalently: } \tilde{R}_p = \sup_{\lim\limits_{n\to\infty} e_{p,n}(\tilde{R})=0} \{\tilde{R}\}.$$

The following lemma summarizes the main result in this section.

Lemma 4: Lower bounding the mutual information $I(a^{nR}, b^n)$ by the randomized capacity: for any $\epsilon > 0$ the mutual information is lower bounded by the minimum randomized lossy coding channel capacity:

$$\liminf_{n\to\infty} \frac{1}{n} I(a^{nR}; b^n) \geq \tilde{R}_p = \sup_{\lim\limits_{n\to\infty} e_{p,n}(\tilde{R})=0} \{\tilde{R}\} \quad (37)$$

Proof: : to show 37, from the definition of $\tilde{R}_p$, we know that it is enough to show that for all $\tilde{R}$, such that $\lim\limits_{n\to\infty} e_{p,n}(\tilde{R}) = 0$:

$$\liminf_{n\to\infty} \frac{1}{n} I(a^{nR}; b^n) \geq \tilde{R}.$$

First we take a new perspective of the Bernoulli sequence b^n. Instead of letting b^n be i.i.d generated from the Bernoulli p random process, we first generate two auxiliary random variables C_p and m and then the b^n is a function of the two auxiliary random variables in a way such that b^n is an i.i.d Bernoulli p sequence.

We first generate a codebook random variable C_p according to the distribution described in (35), where the code book $C_p = C = (c_1, ..., c_{2^{n\tilde{R}}})$ with the following probability:

$$\Pr(C_p = C) = \prod_{m=1}^{2^{n\tilde{R}}} p^{1(c_m)}(1-p)^{n-1(c_m)}.$$

Then we pick the message random variable m according that is uniform on $\{1, 2, ..., 2^{n\tilde{R}}\}$. Finally we let the binary sequence b^n be a function of C_p and m, such that for $\mathsf{C}_p = C = (c_1, ..., c_{2^{n\tilde{R}}})$ and $\mathsf{m} = m$, $b^n = c_m$. It is easy to see that b^n chosen this way have the following distribution:

$$\Pr(\mathsf{b}^n = b^n) = p^{1(b^n)}(1-p)^{n-1(b^n)}.$$

So we have the following Markov Chain:

$$(\mathsf{C}_p, \mathsf{m}) \to \mathsf{b}^n \to \mathsf{a}^{nR} \tag{38}$$

So from the data processing lemma and the chain rule for mutual information, we know that:

$$\begin{aligned} I(\mathsf{a}^{nR}; \mathsf{b}^n) &\geq I(\mathsf{a}^{nR}; \mathsf{C}_p, \mathsf{m}) \\ &= I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p) + I(\mathsf{a}^{nR}; \mathsf{C}_p) \\ &\geq I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p) \end{aligned} \tag{39}$$

where the last inequality follows that mutual information is always non-negative. Now the overall error probability is, as defined in (36):

$$e_{p,n}(\tilde{R}) = \sum_{C \in \mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(\mathsf{C}_p = C) \left(\Pr(\mathsf{m} \neq \hat{\mathsf{m}}(\mathsf{a}^{nR})|\mathsf{C}_p = C) \right) \tag{40}$$

where $\Pr(\mathsf{C}_p = C)(\Pr(\mathsf{m} \neq \hat{\mathsf{m}}(\mathsf{a}^{nR})|\mathsf{C}_p = C))$ is the decoding error when the code book C is chosen. Hence this is a standard communication problem that we can use the technique detailed in Chapter 8.9 [5] to lower bound the mutual information $I(\mathsf{a}^{nR}; \mathsf{b}^n)$ by the rate $\tilde{R}$ that a reliable communication is possible. Notice that if the codebook C is chosen, we have the following Markov Chain:

$$\mathsf{m} \to \mathsf{b}^n \to \mathsf{a}^{nR} \to \hat{\mathsf{m}}, \tag{41}$$

more specifically b^n is a deterministic function of m, $\hat{\mathsf{m}}$ is a deterministic function of a^{nR}. So we can apply Fano's inequality (Theorem 2.11.1 [5] for any fixed codebook C:

$$H(\mathsf{m}|\mathsf{a}^{nR}, \mathsf{C}_p = C) \leq 1 + \Pr(\mathsf{m} \neq \hat{\mathsf{m}}(\mathsf{a}^{nR})|\mathsf{C}_p = C)n\tilde{R} \tag{42}$$

Now, from the standard information theoretic equalities:

$$\begin{aligned} n\tilde{R} &= H(\mathsf{m}) \\ &= H(\mathsf{m}|\mathsf{C}_p = C) \\ &= H(\mathsf{m}|\mathsf{a}^{nR}, \mathsf{C}_p = C) + I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p = C) \\ &\leq 1 + \Pr(\mathsf{m} \neq \hat{\mathsf{m}}(\mathsf{a}^{nR})|\mathsf{C}_p = C)n\tilde{R} + I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p = C) \end{aligned}$$

Multiply both sides by $\Pr(\mathsf{C}_p = C)$ and sum over all $C \in \mathcal{B}_n^{2^{n\tilde{R}}}$, we have:

$$\begin{aligned} n\tilde{R} &\leq 1 + \sum_{C \in \mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(\mathsf{C}_p = C) \Big(\Pr(\mathsf{m} \neq \hat{\mathsf{m}}(\mathsf{a}^{nR})|\mathsf{C}_p = C)n\tilde{R} + I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p = C) \Big) \\ &= 1 + n\tilde{R} \times e_{p,n}(\tilde{R}) + I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p) \end{aligned} \tag{43}$$

Finally, substitute (39) into (43), we have:

$$I(\mathsf{a}^{nR}; \mathsf{b}^n) \geq I(\mathsf{a}^{nR}; \mathsf{m}|\mathsf{C}_p) \geq n\tilde{R} - 1 - n\tilde{R} \times e_{p,n}(\tilde{R})$$

So, if the randomized lossy coding capacity is above $\tilde{R}$, i.e. $\lim_{n\to\infty} e_{p,n}(\tilde{R}) = 0$, then

$$\liminf_{n\to\infty} \frac{1}{n} I(\mathsf{a}^{nR}; \mathsf{b}^n) \geq \tilde{R}$$

□

VI. RANDOMIZED CHANNEL CAPACITY OF A LOSSY COMPRESSOR, A LOWER BOUND

In the previous section, we showed the relation between the mutual information $I(\mathsf{a}^{nR}, \mathsf{b}^n)$ is lower bounded by the randomized lossy coding capacity if the input codewords look like an i.i.d Bernoulli p sequence. What was missing in the previous section is a lower bound on the randomized capacity. In this section we study the capacity, in particular the lower bound on the capacity. Notice that the encoder is using a randomized code book according to the distribution in (35). We only need to design the decoder G_n in Figure 3. If we could show that for some $\tilde{R}$, the average error probability $e_{p,n}(\tilde{R})$ goes to zero as n goes to infinity, then whatever the $\tilde{R}$ is, it is a lower bound on the randomized lossy coding capacity $\tilde{R}_p$. We give a lower bound on $\tilde{R}_p$. As will be clear soon from our derivation of the lower bound, this bound is not tight. However, this is our first effort to derive a non-trivial lower bound to the rate distortion function $R(D,p)$.

Lemma 5: A lower bound on the randomized lossy coding capacity:

$$\tilde{R}_p \geq \underline{\tilde{R}} = \max_{L\geq 0}\{\min_{U\geq L, r\in[0,1-p]: T_1(L,U,r)\leq D} h(L,U,r)\}$$

$$\text{where } h(L,U,r) = \{ \begin{array}{cc} (p\times \Pr(|\mathsf{s}|>U)+r)D(\frac{p\times \Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r}\|p) & \text{, if } \frac{p\times \Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r}\geq p \\ 0 & \text{, if } \frac{p\times \Pr(|\mathsf{s}|>U)}{p\times\Pr(|\mathsf{s}|>U)+r}< p \end{array} \tag{44}$$

$$\mathsf{s} \text{ in (44) is Gaussian } N(0,1) \text{ and } T_1(L,U,r) = rL^2 + 2p\int_L^U (s-L)^2\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds$$

Or equivalently, for all $\tilde{R}\leq \underline{\tilde{R}}$, the decoding error defined in (36) for the randomized coding scheme converges to zero as n goes to infinity:

$$\lim_{n\rightarrow\infty} e_{p,n}(\tilde{R}) = 0$$

Proof: we first describe the decoder G_n. The codebook C is chosen, i.e. $\mathsf{C}_p = C$. As shown in Figure 3, if a message m is to be sent, where $\mathsf{m} \in \{1,2,...,2^{n\tilde{R}}\}$ with equal probability, the binary output to the channel encoder F_n is c_{m}. After the modulation of the Gaussian sequence s^n and the lossy source coding encoder f_n, the channel decoder G_n receives a^{nR}. The first step of G_n is to run the lossy source decoder g_n and get the lossy estimate of $\mathsf{x}^n = c_{\mathsf{m}}\times \mathsf{s}^n$, $\hat{\mathsf{x}}^n = g_n(\mathsf{a}^{nR})$. The second step of G_n is to estimate m from $\hat{\mathsf{x}}^n$. We pick the code word with the most entries' absolute value above the positive real number L:

$$\hat{\mathsf{m}}(a^{nR}(c_1\times\mathsf{s}^n)) = \hat{\mathsf{m}}(\hat{x}^n) = \arg\max_i \sum_{k=1}^n 1(|c_i(k)\hat{\mathsf{x}}_k)|\geq L) \tag{45}$$

where $c_i\in\{0,1\}^n$ is the codeword for message i in the chosen codebook C and $c_i(k)\in\{0,1\}$ is the k-th entry of the codeword c_i. Now we analyze the average error probability of the above coding system over all codebooks according the the codebook distribution in (35) and the over all Gaussian sequence s^n. The average error probability is hence as shown in (36):

$$\begin{aligned} e_{p,n}(\tilde{R}) &= \sum_{C\in\mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(\mathsf{C}_p = C)\left(\frac{1}{2^{n\tilde{R}}}\sum_{m=1}^{2^{n\tilde{R}}}\overset{\mathsf{s}}{\Pr}(m\neq\hat{m}(a^{nR}(c_m\times\mathsf{s}^n)))\right) \\ &= \sum_{C\in\mathcal{B}_n^{2^{n\tilde{R}}}} \Pr(\mathsf{C}_p = C)\left(\overset{\mathsf{s}}{\Pr}(1\neq\hat{m}(a^{nR}(c_1\times\mathsf{s}^n)))\right) \qquad (46) \\ &= \overset{\mathsf{C}_p,\mathsf{s}}{\Pr}(1\neq\hat{m}(a^{nR}(c_1\times\mathsf{s}^n))) \qquad (47) \end{aligned}$$

where (46) follows the symmetry of the system.

We decompose (47) into four parts. We sketch the partitions then give a detailed analysis.

1) The atypical behavior of codeword c_1. The typicality is defined in the usual way [5] for finite discrete random sequences. The concentration theorem is well established in the literature.
2) The atypical behavior of $\widetilde{s}^{1(c_1)}$ while c_1 is typical, where $\widetilde{s}^{1(c_1)}$ is the non-zero subsequence of $x^n = c_1 \times s^n$ where $\widetilde{s}_1 = s_{i_1}, ..., \widetilde{s}_{1(c_1)} = s_{i_{1(c_1)}}$, where $i_1, ..., i_{1(c_1)}$ are the non-zero locations of c_1. The typicality for a Gaussian $N(0,1)$ sequence is defined in Appendix D. We prove the concentration result in Lemma 6.
3) The atypical behavior of the lossy source coding while both c_1 and $s_{i_1}, ..., s_{i_{1(c_1)}}$ are typical. i.e. the distortion of the Bernoulli-Gaussian sequence $d(c_1 \times s^n, \hat{x}^n) = d(x^n, \hat{x}^n) > D$, the concentration of the typical behavior of the lossy source coding is established in (9) for good lossy coders.
4) The probability that there exists a message $\underline{m}$ that has a higher score than message 1 according to the decoding rule in (45) while everything else (the codeword for message 1, c_1, the subsequence $s_{i_1}, ..., s_{i_{1(c_1)}}$, and the distortion $d(c_1 \times s^n, \hat{x}^n)$ are typical. We bound this error by a union bound argument.

The first part is the atypicality of the codeword for message 1, c_1, the second part is the error probability for $c_1 \in \mathcal{B}_\epsilon^n$, where

$$B_\epsilon^n \triangleq \{b^n \in \{0,1\}^n : |\frac{1(b^n)}{n} - p| \leq \epsilon\}.$$

Under the codebook probability C_p, all c_i's are binary sequences of length-n with distribution such that for all $b^n \in \{0,1\}^n$:

$$\overset{C_p}{\Pr}(c_i = b^n) = p^{1(b^n)}(1-p)^{n-1(b^n)}, \quad i = 1, 2, ...2^{n\tilde{R}}. \tag{48}$$

so we obviously have [5]:

$$\lim_{n\to\infty} \overset{C_p}{\Pr}(c_1 \notin B_\epsilon^n) = 0 \tag{49}$$

The second part is the atypicality of the Gaussian subsequence $s_{i_1}, ..., s_{i_{1(c_1)}}$, where $i_1, ..., i_{1(c_1)}$ are the non-zero locations of c_1, while c_1 is typical, $c_1 \in B_\epsilon^n$). The typical Gaussian $N(0,1)$ set is defined as follows, first we have two definitions: for a real sequence s^n and s.t. $-\infty \leq S \leq T \leq \infty$, the l-th moment of entries in s^n within interval $[S,T]$ is denoted by

$$n_{s^n}^l(S,T) = \frac{\sum_{i=1}^n 1(S < s_i < T) s_i^l}{n}.$$

Then the ϵ-typical set for Gaussian $N(0,1)$ is defined as:

$$S_\epsilon(n) = \left\{ s^n : \max_{l=0,1,2} \left\{ \sup_{S,T} \left| n_{s^n}^l(S,T) - \int_S^T s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| < \epsilon \right\} \right\}$$

We prove the concentration result in Lemma 6 in Appendix D: $\lim_{n\to\infty} \overset{s}{\Pr}(s^n \notin S_\epsilon(n)) = 0$. c_1 and s^n are independent, and if $c_1 \in B_\epsilon^n$, then $1(c_1) \geq p(n - \epsilon)$, so if n goes to infinity, $1(c_1)$ goes to infinity too, so

$$\begin{aligned} \lim_{n\to\infty} \overset{C_p,s}{\Pr}\left(c_1 \in B_\epsilon^n, \widetilde{s}^{1(c_1)} \notin S_\epsilon(1(c_1))\right) &\leq \lim_{n\to\infty} \overset{C_p,s}{\Pr}\left(\widetilde{s}^{1(c_1)} \notin S_\epsilon(1(c_1)) | c_1 \in B_\epsilon^n\right) \\ &= 0 \end{aligned} \tag{50}$$

where the first inequality follows that conditional probability is bigger than joint probability.

The third part is the atypical behavior of the lossy coding system. Following the definition of a good lossy source coder in the strong sense in (9) and that $x^n = c_1 \times s^n$, we have, for all $\delta_1 > 0$

$$\lim_{n\to\infty} \overset{x}{\Pr}\left(d(x^n, \hat{x}^n) \geq D + \delta_1\right) = \lim_{n\to\infty} \overset{C_p,s}{\Pr}\left(d(c_1 \times s^n, \hat{x}^n) \geq D + \delta_1\right) = 0$$

This implies that:

$$\lim_{n\to\infty} \overset{\mathsf{C}_p,\mathsf{s}}{\Pr}\left(\mathsf{c}_1 \in B_\epsilon^n, \widetilde{\mathsf{s}}^{1(\mathsf{c}_1)} \in S_\epsilon(1(\mathsf{c}_1)), d(\mathsf{c}_1 \times \mathsf{s}^n, \hat{\mathsf{x}}^n) \geq D + \delta_1\right) = 0 \tag{51}$$

The fourth part is when the code word c_1, the Gaussian subsequence $\widetilde{\mathsf{s}}^{1(\mathsf{c}_1)}$, and the distortion $d(\mathsf{c}_1 \times \mathsf{s}^n, \hat{\mathsf{x}}^n)$ are all typical, the decoding error for the channel decoder following the decoding rule in (45).

The output of the lossy source coding decoder is $\hat{\mathsf{x}}^n = g_n(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n))$, from the decoding rule in (45), the estimate of the message $\hat{m}(a^{nR}(c_1 \times \mathsf{s}^n)))$ is not equal to the true message 1, if and only if there exists a message $\underline{m} \neq 1$, such that

$$\sum_{k=1}^{n} 1(|c_{\underline{m}}(k)\hat{x}_k)| \geq L) \geq \sum_{k=1}^{n} 1(|c_1(k)\hat{x}_k)| \geq L) \tag{52}$$

Notice that the codebooks are symmetric to the messages, i.e. over all the codebooks, the probability that the estimation of the message $\hat{\mathsf{m}} = i$ is equal to the probability that $\hat{\mathsf{m}} = j$ for all $i, j \in \{1, 2, ..., 2^{n\tilde{R}}\}$ and $i \neq 1$, $j \neq 1$. So we can union bound the decoding error probability of the event shown in (52) as follows:

$$\overset{\mathsf{s},\mathsf{C}_p}{\Pr}(1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n))) \quad \leq \quad 2^{n\tilde{R}} \overset{\mathsf{s},\mathsf{C}_p}{\Pr}\left(\sum_{k=1}^{n} 1(|\mathsf{c}_2(k)\hat{x}_k| \geq L) \geq \sum_{k=1}^{n} 1(|\mathsf{c}_1(k)\hat{x}_k| \geq L)\right) \tag{53}$$

where the probability is calculated over all possible codebooks over the measure C_p and the Gaussian sequences s^n. First, for a codeword c_1, and the lossy coding estimate of $c_1 \times s^n$, $\hat{x}^n$, denote by u and v the number of entries of the estimate $\hat{x}_k$ with absolute value above L where $c_1(k)$ is 1 and 0 respectively:

$$\begin{aligned} u &= \sum_{k=1}^{n} 1(|c_1(k)\hat{x}_k)| \geq L) \\ v &= \sum_{k=1}^{n} 1(|\hat{x}_k)| \geq L \text{ and } c_1(k) = 0). \end{aligned} \tag{54}$$

With u and v fixed(here we fix the codeword c_1, the sequence s^n and the estimate $\hat{x}^n$), we union bound the probability of the following event that there exists a message $\underline{m} \neq 1$, such that (52) is true:

$$\begin{aligned} \overset{\mathsf{s},\mathsf{C}_p}{\Pr}(1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n))|\mathsf{c}_1 = c_1, \mathsf{s}^n = s^n) \quad &\leq \quad 2^{n\tilde{R}} \overset{\mathsf{C}_p}{\Pr}\left(\sum_{k=1}^{n} 1(|\mathsf{c}_2(k)\hat{x}_k| \geq L) \geq \sum_{k=1}^{n} 1(|c_1(k)\hat{x}_k| \geq L)\right) \\ &= \quad 2^{n\tilde{R}} \overset{\mathsf{C}_p}{\Pr}\left(\sum_{k=1}^{n} 1(|\mathsf{c}_2(k)\hat{x}_k| \geq L) \geq u\right) && (55) \\ &= \quad 2^{n\tilde{R}} \sum_{l=u}^{u+v} \begin{pmatrix} u+v \\ l \end{pmatrix} p^l (1-p)^{u+v-l} && (56) \\ &\leq \quad 2^{n\tilde{R}} \times n \max_{l: u \leq l \leq u+v} \{\begin{pmatrix} u+v \\ l \end{pmatrix} p^l (1-p)^{u+v-l}\} && (57) \\ &\leq \quad 2^{n\tilde{R}} \times n 2^{-(u+v) \min\limits_{l: u \leq l \leq u+v} D(\frac{l}{u+v} \| p)} && (58) \\ &= \quad 2^{n\tilde{R}} \times \{ \begin{array}{cl} n & \text{, if } \frac{u}{u+v} \leq p \\ n 2^{-(u+v) D(\frac{u}{u+v} \| p)} & \text{, if } \frac{u}{u+v} > p \end{array} && (59) \end{aligned}$$

(55) follows the definition of u. (56) follows that $\mathsf{c}_2 \in \{0, 1\}^n$ is an i.i.d. Bernoulli p sequence. (57) is because $u + v \leq n$. (58) and (59) follows basic information theoretic inequalities [6]. From Lemma 7 in Appendix E, we know that the $(u + v)D(\frac{u}{u+v}\|p)$ is monotonically increasing with u and monotonically decreasing with v. $\frac{u}{u+v}$ is also monotonically increasing with u and monotonically decreasing with v, so the expression in (59) is monotonically **decreasing** with u and monotonically **increasing** with v.

(59) is true for all codeword c_1 and sequence $\tilde{s}^{1(c_1)}$, typical or not. So it is also true for all those $c_1 \in B_\epsilon^n$, $\tilde{s}^{1(c_1)} \in S_\epsilon(1(c_1))$ and $d(c_1 \times s^n, \hat{x}^n) \leq D + \delta_1$ in this case, we can give a feasible region for u and v, i.e. then give a bound on (59). We further investigate the distortion for the said typical sequences:

$$\begin{aligned} n(D+\delta_1) &\geq nd(c_1 \times s^n, \hat{x}^n) \\ &= \sum_{k=1}^{n}(c_1(k)s_k - \hat{x}_k)^2 \\ &= \sum_{k:c_1(k)=1}(c_1(k)s_k - \hat{x}_k)^2 + \sum_{k:c_1(k)=0}\hat{x}_k^2 \\ &= \sum_{k:c_1(k)=1}(c_1(k)s_k - \hat{x}_k)^2 + \sum_{k:c_1(k)=0,x_k\geq L}\hat{x}_k^2 + \sum_{k:c_1(k)=0,x_k<L}\hat{x}_k^2 \\ &\geq \sum_{k:c_1(k)=1}(c_1(k)s_k - \hat{x}_k)^2 + vL^2 \qquad (60) \end{aligned}$$

where (60) follows the definition of v. Notice that by definition $x_k = c_1(k)s_k$, so $x_k > 0$ implies that $c_1(k) = 1$, the first term of (60) is:

$$\begin{aligned} \sum_{k:c_1(k)=1}(x_k - \hat{x}_k)^2 &\geq \sum_{k:|x_k|\geq L\geq|\hat{x}_k|}(x_k - \hat{x}_k)^2 \\ &\geq \sum_{k:|x_k|\geq L\geq|\hat{x}_k|}(|x_k| - L)^2 \end{aligned}$$

We rewrite (60) as:

$$n(D+\delta_1) \geq \sum_{k:|x_k|\geq L\geq|\hat{x}_k|}(|x_k| - L)^2 + vL^2 \qquad (61)$$

From the definition of u: we know that $u = \sum_{k=1}^{n} 1(|c_1(k)\hat{x}_k)| \geq L)$ hence

$$\begin{aligned} \sum_{k=1}^{n} 1(|x_k| \geq L \geq |\hat{x}_k|) &\geq \sum_{k=1}^{n} 1(|x_k| > 0) - \sum_{k=1}^{n} 1(0 < |x_k| \leq L) - \sum_{k=1}^{n} 1(|c_1(k)\hat{x}_k)| \geq L) \\ &= \sum_{k=1}^{n} 1(|x_k| > L) - u \\ &\triangleq n(|x_k| > L) - u \qquad (62) \end{aligned}$$

Recall that $\tilde{s}_1, ...\tilde{s}_{1(c_1)}$ are the none-zero entries of x^n, without out loss of generality, let $|\tilde{s}_1|, ...|\tilde{s}_{n(|x_k|>L)-u}|$ be the smallest $n(|x_k| > L) - u$ many $|x_k|$'s that are larger than L, without loss of generality let $|\tilde{s}_1| \geq .... \geq$

$|\tilde{s}_{n(|x_k|>L)-u}| \geq L$. Then substituting (62) into (61) and denote by $\tilde{U} = |\tilde{s}_1|$, we have:

$$\begin{align*}
n(D+\delta_1) &\geq \sum_{j=1}^{n(|x_k|>L)-u} (|\tilde{s}_j| - L)^2 + vL^2 \\
&= \sum_{j: L < |\tilde{s}_j| \leq \tilde{U}} (|\tilde{s}_j| - L)^2 + vL^2 \tag{63} \\
&= \sum_{j: L < |\tilde{s}_j| \leq \tilde{U}} (|\tilde{s}_j|^2 - 2L|\tilde{s}_j| + L^2) + vL^2 \\
&\geq 2 \times 1(c_1) \left(\int_L^{\tilde{U}} (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds - \epsilon(1+L)^2 \right) + vL^2 \tag{64} \\
&\geq 2 \times n(p-\epsilon) \left(\int_L^{\tilde{U}} (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds - \epsilon(1+L)^2 \right) + vL^2 \tag{65} \\
&\geq n \left(2p \int_L^{\tilde{U}} (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds + \frac{v}{n} L^2 \right) - n\epsilon K_1(p, L) \tag{66}
\end{align*}$$

(63) follows the definition of $\tilde{s}^{1(c_1)}$, (64) is true because $\tilde{s}^{1(c_1)} \in S_\epsilon(1(c_1))$ is ϵ-typical Gaussian $N(0,1)$. (65) is true because $c_1 \in B_\epsilon^n$. Finally in (66), $K_1(p, L)$ is a finite function of p and L, we do not need $\tilde{U}$ in the picture because we can replace $\tilde{U}$ with ∞ when bounding the the residue. We rewrite (66) as:

$$2p \int_L^{\tilde{U}} (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds + \frac{v}{n} L^2 \leq D + \delta_1 + \epsilon K_1(p, L) \tag{67}$$

Meanwhile, because $\tilde{U} = |\tilde{s}_1| \geq ... \geq |\tilde{s}_{n(|x_k|>L)-u}| \geq L$ are the smallest $n(|x_k| > L) - u$ many $|x_k|$'s that are larger than L, $\tilde{s}^{1(c_1)}$ is a ϵ-typical Gaussian sequence, so $n(|x_k| > L) - u \leq 1(c_1)(\Pr(L < |s| < \tilde{U}) + \epsilon)$, hence:

$$\begin{align*}
u &> n(|x_k| > L) - 1(c_1)(\Pr(L < |\mathsf{s}| < \tilde{U}) + \epsilon) \\
&\geq n(p-\epsilon)(\Pr(|\mathsf{s}| > L) - \epsilon) - n(p+\epsilon)(\Pr(L < |\mathsf{s}| < \tilde{U}) + \epsilon) \\
&= np \Pr(|\mathsf{s}| > \tilde{U}) - n\epsilon K_2(p, L) \tag{68}
\end{align*}$$

The above analysis are true for all δ_1 and ϵ, we let both be small, we have

$$\begin{align*}
u \geq n\big(p \Pr(|\mathsf{s}| > U) - \epsilon_2\big) \tag{69} \\
\text{s.t.: } 2p \int_L^U (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds + \frac{v}{n} L^2 \leq D \tag{70}
\end{align*}$$

where $\lim\limits_{\delta, \epsilon \to 0} \epsilon_2 = 0$, this is true because for any $\tilde{U}$ that satisfies (67), it either also satisfies the more stringent constraint in (70) or the gap between $\tilde{U}$ and the biggest U that satisfies (70) is small when δ_1 and ϵ are small. Then (70) follows the continuity of $\Pr(|\mathsf{s}| > U)$ in U.

Notice that (59) holds for all codeword c_1 and s^n, in particular it is true for the typical ones, $c_1 \in B_\epsilon^n$ and $\tilde{s}^{1(c_1)} \in S_\epsilon(1(c_1))$ and $d(c_1 \times s^n, \hat{x}^n) \leq D + \delta_1$, also (59) is monotonically decreasing with u, with (69) and let $r = \frac{v}{n}$, recall the definition of v in (54), for $c_1 \in B_n^\epsilon$, $v \leq n - n(p-\epsilon) = n(1-p+\epsilon)$ or equivalently $r \in [0, 1-p]$, we rewrite (59):

$$\begin{align*}
&\Pr^{\mathsf{s}, C_p}\left(1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n)) | \mathsf{c}_1 = c_1 \in B_\epsilon^n, \mathsf{s}^n = s^n \in S_\epsilon(1(c_1)), d(c_1 \times s^n, \hat{x}^n) \leq D + \delta_1\right) \\
\leq\ & 2^{n\tilde{R}} \times \left\{ \begin{array}{cl} n & \text{, if } \frac{u}{u+v} \leq p \\ n 2^{-(u+v) D(\frac{u}{u+v} \| p)} & \text{, if } \frac{u}{u+v} > p \end{array} \right. \\
\leq\ & 2^{n\tilde{R}} \times \left\{ \begin{array}{cl} n & \text{, if } \frac{\Pr(|\mathsf{s}| > U)}{\Pr(|\mathsf{s}| > U) + r} \leq p \\ n 2^{-n\left((\Pr(|\mathsf{s}| > U) + r) D(\frac{\Pr(|\mathsf{s}| > U)}{\Pr(|\mathsf{s}| > U) + r} \| p) - \epsilon_3\right)} & \text{, if } \frac{\Pr(|\mathsf{s}| > U)}{\Pr(|\mathsf{s}| > U) + r} > p \end{array} \right. \tag{71}
\end{align*}$$

with (70) being satisfied, where $\lim_{\epsilon_2 \to 0} \epsilon_3 = 0$ because the exponent in (71) is continuous in u, we know that $\lim_{\delta,\epsilon \to 0} \epsilon_2 = 0$, so $\lim_{\delta,\epsilon \to 0} \epsilon_3 = 0$ as well.

Notice that the coding system can pick arbitrary L, it picks the best possible L, we have, if

$$\tilde{R} < \underline{\tilde{R}} = \max_{L \geq 0} \{ \min_{U \geq L, r \in [0, 1-p] : T_1(L,U,r) \leq D} h(L,U,r) \}$$

$$\text{where } h(L,U,r) = \{ \begin{array}{cc} (p \times \Pr(|\mathsf{s}| > U) + r) D(\frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U) + r} \| p) & \text{, if } \frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U)+r} \geq p \\ 0 & \text{, if } \frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U)+r} < p \end{array}$$

then

$$\lim_{n \to \infty} \overset{\mathsf{s},\mathsf{C}_p}{\Pr} (1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n)) | \mathsf{c}_1 = c_1 \in B^n_\epsilon, \mathsf{s}^n = s^n \in S_\epsilon(1(c_1)), d(c_1 \times s^n, \hat{x}^n) \leq D + \delta_1) = 0 \tag{72}$$

The above inequality is true for all those $c_1 \in B^n_\epsilon$, $\tilde{s}^{1(c_1)} \in S_\epsilon(1(\mathsf{c}_1))$ and $d(c_1 \times s^n, \hat{x}^n) \leq D + \delta_1$, so

$$\lim_{n \to \infty} \overset{\mathsf{s},\mathsf{C}_p}{\Pr} (1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n)), \mathsf{c}_1 \in B^n_\epsilon, \mathsf{s}^n \in S_\epsilon(1(c_1)), d(\mathsf{c}_1 \times \mathsf{s}^n, \hat{\mathsf{x}}^n) \leq D + \delta_1) = 0 \tag{73}$$

Finally we can upper bound the overall error probability of the randomized coding scheme. The decoding error $e_{p,n}(\tilde{R})$ is defined in (36) which is equivalent to (46) because of the symmetry. We decompose the error event into 4 atypical events as illustrated at the beginning of the proof. For any $\tilde{R} < \underline{\tilde{R}}$,

$$\begin{aligned} e_{p,n}(\tilde{R}) &= \overset{\mathsf{C}_p,\mathsf{s}}{\Pr} (1 \neq \hat{m}(a^{nR}(c_1 \times \mathsf{s}^n))) & (74) \\ &\leq \overset{\mathsf{C}_p}{\Pr}(\mathsf{c}_1 \notin B^n_\epsilon) \\ &\quad + \overset{\mathsf{C}_p,\mathsf{s}}{\Pr} (\mathsf{c}_1 \in B^n_\epsilon, \widetilde{\mathsf{s}}^{1(\mathsf{c}_1)} \notin S_\epsilon(1(\mathsf{c}_1))) \\ &\quad + \overset{\mathsf{C}_p,\mathsf{s}}{\Pr} \left(\mathsf{c}_1 \in B^n_\epsilon, \widetilde{\mathsf{s}}^{1(\mathsf{c}_1)} \in S_\epsilon(1(\mathsf{c}_1)), d(\mathsf{c}_1 \times \mathsf{s}^n, \hat{\mathsf{x}}^n) > D + \delta_1 \right) \\ &\quad + \overset{\mathsf{s},\mathsf{C}_p}{\Pr} (1 \neq \hat{m}(a^{nR}(\mathsf{c}_1 \times \mathsf{s}^n)), \mathsf{c}_1 \in B^n_\epsilon, \mathsf{s}^n \in S_\epsilon(1(c_1)), d(\mathsf{c}_1 \times \mathsf{s}^n, \hat{\mathsf{x}}^n) \leq D + \delta_1) & (75) \end{aligned}$$

where (74) follows (46). The asymptotic behaviors of the four terms in (74) are shown in (49), (50), (51) and (73) respectively. δ_1 can be arbitrarily small, so we can finally claim that: for a good lossy source coding system in the strong sense with distortion constraint D, the randomized channel coding error converges to zero as n goes to infinity:

$$\lim_{n \to \infty} e_{p,n}(\tilde{R}) = 0$$

This concludes the proof of Lemma 5. □

VII. Discussions and Numerical Result

Now we have two upper bounds and two lower bounds on the rate distortion function $R(D,p)$. We reiterate the bounds,

$$R(D,p) \leq H(p) + pR(D, N(0,p)) \tag{76}$$

$$R(D,p) \leq R(D, N(0,p)) \tag{77}$$

$$R(D,p) \geq pR(D, N(0,p)) \tag{78}$$

$$R(D,p) \geq pR(D, N(0,p)) + \max_{L \geq 0} \{ \min_{U \geq L, r \in [0,1-p] : T_1(L,U,r) \leq D} h(L,U,r) \} \triangleq pR(D, N(0,p)) + R_i(D,p) \tag{79}$$

$$\text{where } h(L,U,r) = \{ \begin{array}{cc} (p \times \Pr(|\mathsf{s}| > U) + r) D(\frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U) + r} \| p) & \text{, if } \frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U)+r} \geq p \\ 0 & \text{, if } \frac{p \times \Pr(|\mathsf{s}|>U)}{p \times \Pr(|\mathsf{s}|>U)+r} < p \end{array} \tag{80}$$

s is Gaussian $N(0,1)$ and $T_1(L,U,r) = rL^2 + 2p \int_L^U (s-L)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$

where $R(D,N(0,p))$ is the rate distortion function for zero mean variance p Gaussian random sequence with distortion constraint D, $R(D,N(0,p))=\max\{0,\frac{1}{2}\log_2\frac{p}{D}\}$. (76), (77) and (78) are derived in Propositions 2, 3 and 4 respectively, (79) is the main result in Theorem 1.

A. Properties of the improvement $R_i(D,p)$

The improvement of our new lower bound, the second term $R_i(D,p)$ in (79), has a game theoretic interpretation. In a two player zero sum game, the first player (the coding system) chooses L, the second player (adversary) chooses U and r with string attached in (80), the payoff to player one is $h(U,L,r)$. First we argue that the improvement of our lower bound, the second term $R_i(p,D)$ in (79), is monotonically decreasing with D and if for some D, the improvement is zero.

Corollary 3: $R_i(D,p)$ is monotonically decreasing with D, i.e. for $D_1>D_2$, $R_i(D_1,p)\leq R_i(D_2,p)$

Proof: $R_i(D,p)$ is of the form of

$$\max_{L\geq 0}\{\min_{U\geq L, r\in[0,1-p]:T_1(L,U,r)\leq D} h(L,U,r)\},$$

so for all $L\geq 0$, if the pair (U,r) is feasible for D_2, it is also feasible for D_1, hence the minimum of $h(L,U,r)$ for D_1 is no bigger than that for D_2. □

More importantly the improvement is within $[0,H(p)]$ in light of the upper bound in (76). In the low distortion regime, i.e. $\frac{D}{p}\ll 1$. We argue that the improvement $R_i(D,p)$ is close to $p\log_2\frac{1}{p}$.

Corollary 4: Asymptotic behavior of $R_i(D,p)$ in the low distortion regime , for any $p>0$

$$\lim_{D\to 0} R_i(D,p)=p\log_2\frac{1}{p}$$

Proof: We only give a sketch of proof here. The coding system pick a positive $L\ll 1$, but $L^2\gg D$, say $L=D^{0.3}$ The distortion constraint on $T_1(L,U,r)$ implies that $D\geq rL^2$, hence

$$r\leq\frac{D}{L^2}=D^{0.4}.$$

So r goes to zero as D goes to zero. Similarly we argue that U goes to zero as D goes 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:

$$\frac{D}{2p}\geq\int_L^U(s-L)^2\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds\geq\int_L^U(\frac{3s^2}{4}-3L^2)\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds=\int_{D^{0.3}}^U(\frac{3s^2}{4}-3D^{0.6})\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds$$

hence:

$$\int_{D^{0.3}}^U\frac{3s^2}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds\leq\frac{D}{2p}+\int_{D^{0.3}}^U 3D^{0.6}\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}ds\leq\frac{D}{2p}+3D^{0.6}$$

take limit on both side when $D\to 0$, the right hand side is 0, the left hand side is zero if and only if $U\to 0$ as D goes to zero. We just showed that if we pick $L=D^{0.3}$ and D goes to zero, then both U and r goes to zero if the distortion constraint be satisfied. This means that the in this case:

$$\lim_{D\to 0}R_i(D,p)=\lim_{r,U\to 0}(p\times\Pr(|s|>U)+r)D(\frac{p\times\Pr(|s|>U)}{p\times\Pr(|s|>U)+r}\|p)=pD(1\|p)=p\log_2\frac{1}{p}$$

□

A simple corollary of Corollary 4 is as follows. For small p, the sparse signal studied in the compressive sensing literature:

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

So the gap between the improved lower bound in (79) and the upper bound in (76) is at most $\log_2(e)p$ which is dominated by the improvement $p\log_2\frac{1}{p}$ for small p.

B. Numerical Results

We plot the bounds in (76)- (79) for $p = 0.1$. As shown in Figure 4, the rate distortion function $R(D, p)$ is bounded by the lower and upper bounds in (76)- (79)

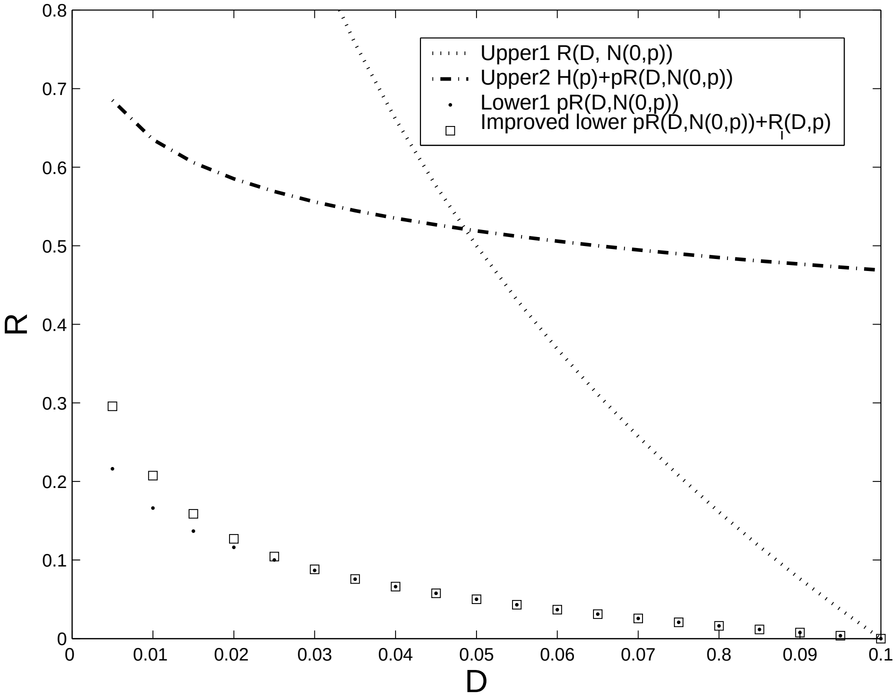

Fig. 4. Lower and upper bounds on $R(D, p)$ for $p = 0.1$ at high distortion levels, the distortion D runs from 0.005 to 0.1

VIII. Conclusions and Future Work

In this paper we study the rate distortion function for Bernoulli-Gaussian sequences. The main result is an improved lower bound on the rate distortion function. The improvement over the known best lower bound is $p \log_2 \frac{1}{p}$ if D is small. This is significant since the currently known gap between the lower bound and upper bound is $H(p)$, hence the improved lower bound is almost tight for sparse signals where $p \ll 1$. To derive this lower bound, we develop a new technique to lower bound part of the rate distortion function through a randomized lossy coding channel. This is, to our knowledge, the first work on this topic. This new lower bound and the obvious upper bounds do not match. The lower bounding technique we use in this paper can be improved if we can relax the near-zero error probability constraint on the randomized channel coding. A potentially useful direction is to replace the channel coding part with a lossy source coder. This is left for future work. There is another interesting result we developed on the way to prove the main result. We showed the equivalence of the rate distortion functions in strong sense and expected distortion sense for continuous random variables with finite variances.

Acknowledgments

The author thanks Marcelo Weinberger for introducing the problem, Erik Ordentlich for pointing out [10] and other members of the information theory group at Hewlett-Packard Labs, Palo, Alto for helpful discussions along the way.

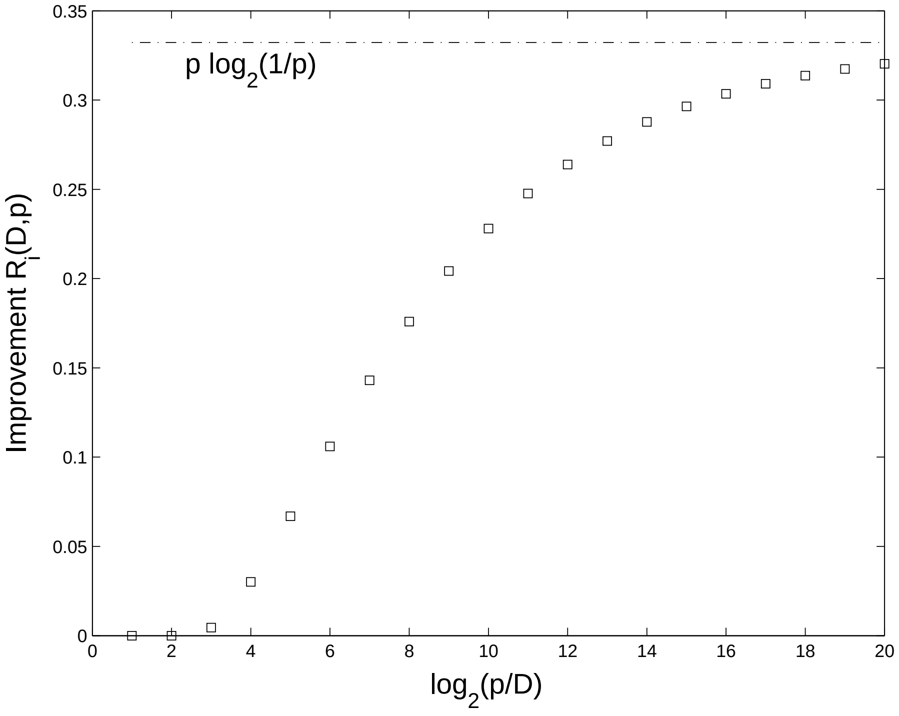

Fig. 5. The improvement $R_i(D,p)$ for $p = 0.1$ at low distortion levels. As proved in Corollary 4, $R_i(D,p) \to p \log_2 \frac{1}{p}$ as $D \to 0$

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] Toby Berger and Jerry D. Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44:2693 – 2723, 1998.

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

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

[6] Imre Csiszár and János Körner. *Information Theory*. Akadémiai Kiadó, Budapest, 1986.

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

[8] 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.

[9] 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.

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

[11] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325 – 2383, 1998.

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

Appendix

A. Rate distortion function in the strong sense for continuous random variables

It is shown that the rate distortions function for both the average distortion and the strong distortion are the same for discrete random variables Chapter 13.6 [5]. However it is not obvious if it is also true for continuous random variables. In this section, we give a sketch on why it is also true for continuous(mixed) random variables. Since we have not seen similar results in the classic literature on rate distortion function [3] [2] and [11], we feel it is necessary to give a sketch of proof here.

As shown in Figure 6, to make it more general, we let x be a mixture of a continuous probability function $p(x)$ and finite many discrete values with positive probabilities ($\Pr(x = a_i) = p_i > 0$ shown as impulses in the figure). We need the mean and the variance of x to be finite: $E(x) < \infty$ and $E(x^2) < \infty$.

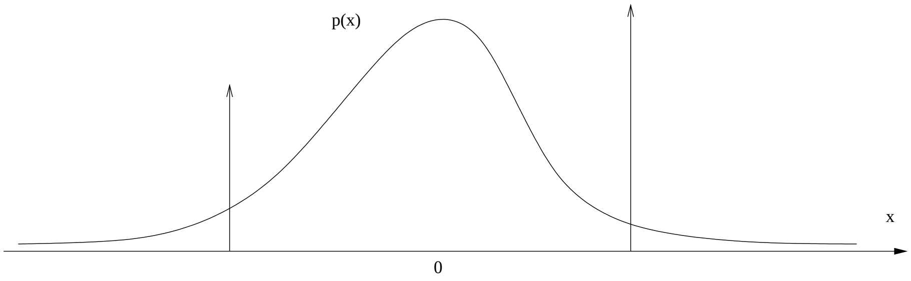

Fig. 6. Probability density function $p(x)$ of a continuous random variable x

First, we argue that the rate distortion function in the expected distortion sense exists for the mixed random variables by approximating the impulses in the pdf by a sharp step function[2] so we have a continuous pdf and the rate distortion theorem can be applied. It remains to be shown that the continuous rate distortion function converges to the one for x as $m \to \infty$. This can be easily proved by noticing that the approximation error is at most $\frac{1}{2m}$ for this approximation, hence the rate distortion function of the continuous random variable converges to the mixed one.

Now we show that the rate distortion function in the strong sense for continuous(mixed) random variable x, denoted by $R_S(D, \mathsf{x})$ is equal to the rate distortion function in the expected sense, denoted by $R_E(D, \mathsf{x})$.

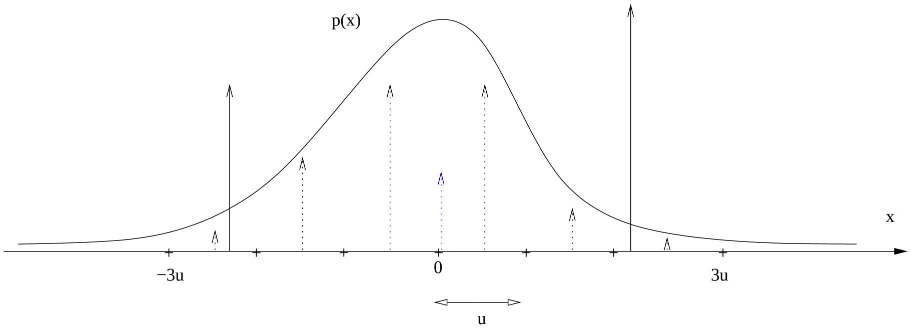

Fig. 7. Quantization of a probability density function $p(x)$ of a mixed random variable x, 7 level quantization for the continuous part and exact representation of the discrete part.

As shown in Figure 7, for the continuous part of the probability density function, we quantize the real line into $(2K+1)$ quantization levels with the interval size d. The intervals are: $[-Ku, -(K-1)u], ..., [-u, 0], [0, u], ..., [(K-1)u, Ku]$ and the "tail" interval $(-\infty, -Ku] \bigcup [Ku, \infty)$. For each interval, the representation value is the middle point of the interval, specifically for the "tail" interval, the representation value is 0. We use the following function $q_{K,u}$ to map a mixed random variable to a discrete random variable:

$$q_{K,u}(x) = \begin{array}{cc} x, & p_{\mathsf{x}}(x) > 0, \\ (k+\frac{1}{2})u, & p_{\mathsf{x}}(x) = 0 \text{ and } x \in [ku, (k+1)u), \quad k = -K, ..., K-1 \\ 0, & p_{\mathsf{x}}(x) = 0 \text{ and } x \in (-\infty, -Ku] \bigcup [Ku, \infty) \end{array}$$

For a random variable x, the output of the map $\mathsf{y}_{K,u} = q_{K,u}(\mathsf{x})$ is a discrete random variable. Hence we know that the rate distortion functions in the strong sense, denoted by $R_S(D, \mathsf{y}_{K,u})$ and the expected distortion sense, denoted by $R_E(D, \mathsf{y}_{K,u})$, are the same.

Now we have four rate distortion functions, the rate distortion function for the mixed (continuous) random variable x, $R_S(D, \mathsf{x})$ and $R_E(D, \mathsf{x})$, and the rate distortion functions for the quantized discrete random variables

[2] For an impulse $\Pr(\mathsf{x} = a_i) = p_i > 0$, we add the continuous pdf $p(x)$ by the following step function $p_i(x)$: $p_i(x) = m$ if $x \in [a_i - \frac{1}{2m}, a_i + \frac{1}{2m}]$, $p_i(x) = 0$ otherwise.

$R_S(D, y_{K,u})$ and $R_E(D, y_{K,u})$. **The goal is to show that** $R_S(D, x) = R_E(D, x)$. First, from the discussion in Appendix B, we know that $R_S(D, x) \geq R_E(D, x)$. It remains to be shown that $R_S(D, x) \leq R_E(D, x)$. We will use the discrete random variable $y_{K,u}$'s rate distortion functions as bridges to show that. We will show that when $u \to 0$ and $Ku \to \infty$: $R_S(D, x) \leq R_S(D, y_{K,u})$ and $R_E(D, y_{K,u}) \leq R_E(D, x)$. And knowing that for discrete random variables $y_{K,u}$, $R_S(D, y_{K,u}) = R_E(D, y_{K,u})$. We will have:

$$R_S(D, x) \leq R_S(D, y_{K,u}) = R_E(D, y_{K,u}) \leq R_E(D, x).$$

This will conclude our proof that $R_S(D, x) = R_E(D, x)$. Now we only need to show that $R_S(D, x) \leq R_S(D, y_{K,u})$ and $R_E(D, y_{K,u}) \leq R_E(D, x)$.

1) $R_S(D, x) \leq R_S(D, y_{K,u})$: We only need to show that if at a rate-distortion pair (R, D), there is a good lossy coder $\tilde{f}_{n_{K,u}}, \tilde{g}_{n_{K,u}}$ in the strong sense for $y_{K,u}$, then there is a good lossy coder f_n, g_n in the strong sense for x.

From the definition of the good lossy coder in the strong sense, we know that for any $\epsilon > 0$,:

$$\lim_{n \to \infty} \overset{y_{K,u}}{\Pr} \left(d(y^n_{K,u}, \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(y^n_{K,u}))) \geq D + \delta_0 \right) = 0$$

Notice that $y^n_{K,u} = q_{K,u}(x)$, so the above equation becomes:

$$\lim_{n \to \infty} \overset{x}{\Pr} \left(d(q_{K,u}(x^n), \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(x^n)))) \geq D + \delta_0 \right) = 0 \tag{81}$$

where the quantizer $q_{K,u}(\cdot)$ is illustrated in Figure 7. Now we show the following encoder decoder pair $f_{n_{K,u}}, g_{n_{K,u}}$ is good in the strong sense for x when u goes to zero and Ku goes to infinity. Where

$$f_{n_{K,u}}(\cdot) = \tilde{f}_{n_{K,u}}(q_{K,u}(\cdot)), \text{ and } g_{n_{K,u}}(\cdot) = \tilde{g}_{n_{K,u}}(\cdot).$$

Notice that the distortion $d(\cdot, \cdot)$ is the mean square of the difference, so almost surely:

$$\begin{aligned} d(x^n, g_{n_{K,u}}(f_{n_{K,u}}(x^n))) &= d(x^n, \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(x^n)))) \\ &\leq d(x^n, q_{K,u}(x^n)) + d(q_{K,u}(x^n), \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(x^n)))) \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - q_{K,u}(x_i))^2 + d(q_{K,u}(x^n), \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(x^n)))) \end{aligned} \tag{82}$$

We analyze the first term in (82). We decompose the sum square depending on how x_i is quantized, remember for $x > Ku$, the quantization is 0 and we assume that Ku is big enough that no discrete part of x is larger than Ku:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (x_i - q_{K,u}(x_i))^2 &= \frac{1}{n} \sum_{i:|x_i| \leq Ku} (x_i - q_{K,u}(x_i))^2 + \frac{1}{n} \sum_{i:|x_i| > Ku} x_i^2 \\ &\leq u + \frac{1}{n} \sum_{i:|x_i| > Ku} x_i^2 \end{aligned}$$

Pick $u < \delta_0$ and Ku big enough such that $E_x(1(|x| > Ku)x^2) < \delta_0 - u$, this is clearly doable because $E_x(x^2) < \infty$. Now we use the weak law of large numbers,

$$\begin{aligned} \overset{x}{\Pr}\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - q_{K,u}(x_i))^2 > \delta_0\right) &\leq \overset{x}{\Pr}\left(\frac{1}{n} \sum_{i:|x_i| > Ku} x_i^2 > \delta_0 - u\right) \\ &= \overset{x}{\Pr}\left(\frac{1}{n} \sum_{i=1}^{n} (1(|x| > Ku)x^2 > \delta_0 - u\right) \\ &\to 0 \qquad \text{as } n \to \infty \end{aligned} \tag{83}$$

Now we can bound the following probability:

$$
\begin{aligned}
\overset{\mathsf{x}}{\Pr}(d(\mathsf{x}^n, g_{n_{K,u}}(f_{n_{K,u}}(\mathsf{x}^n))) > 2\delta_0) &\leq \overset{\mathsf{x}}{\Pr}(\frac{1}{n}\sum_{i=1}^{n}(\mathsf{x}_i - q_{K,u}(\mathsf{x}_i))^2 + d(q_{K,u}(\mathsf{x}^n), \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(\mathsf{x}^n)))) > 2\delta_0 && (84)\\
&\leq \overset{\mathsf{x}}{\Pr}(\frac{1}{n}\sum_{i=1}^{n}(\mathsf{x}_i - q_{K,u}(\mathsf{x}_i))^2 > \delta_0)\\
&\qquad + \overset{\mathsf{x}}{\Pr}(d(q_{K,u}(\mathsf{x}^n), \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(q_{K,u}(\mathsf{x}^n)))) > \delta_0) && (85)\\
&\to 0 \quad \text{as } n \to \infty && (86)
\end{aligned}
$$

where (84) follows (82). (85) is true because $\Pr(\mathsf{x}+\mathsf{y} > 2\epsilon_0) \leq \Pr(\mathsf{x} > \epsilon_0 \text{ or } \mathsf{y} > \epsilon_0) \leq \Pr(\mathsf{x} > \epsilon_0) + \Pr(\mathsf{y} > \epsilon_0)$, while (86) follows (81) and (83).

2) $R_E(D, \mathsf{y}_{K,u}) \leq R_E(D, \mathsf{x})$: We only need to show that if at a rate-distortion pair (R, D), there is a good lossy coder f_n, g_n in the expected distortion sense for x, then there is a good lossy coder $\tilde{f}_{n_{K,u}}, \tilde{g}_{n_{K,u}}$ in the strong sense for $\mathsf{y}_{K,u}$.

From the definition of the good lossy coder in the expected distortion sense, we know that

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

Now we construct a good lossy coder in the expected distortion sense, we implement the following “inverse” map of $q_{K,u}$, denoted by $\mathsf{w}_{K,u}$. Where $\mathsf{w}_{K,u}$ is a random map, for any real sequence y^n generated by the random variable $\mathsf{y}_{K,u}$, y_i can only take values on $A = \{ku : k = -K, ..., 0, ..., K$ and $a \in \mathcal{R}$ where $p_{\mathsf{x}}(a) > 0$, the inverse map $\mathsf{w}_{K,u} : A \to \mathcal{R}$, such that: $\mathsf{w}_{K,u}(\mathsf{y}_{K,u}) \sim \mathsf{x}$ and for all $y \in A$: $\mathsf{w}_{K,u}(y) \in \{x \in \mathcal{R} : q_{K,u}(x) = y\}$. Pictorically the inverse map maps the impulses in Figure 7 back to the mixed random variable with probability density function in Figure 6. The good lossy coder in the expected distortion sense for $\mathsf{y}_{K,u}$ is for all $y^n \in A^n$:

$$
\begin{aligned}
&\tilde{f}_{n_{K,u}}(y^n) = f_n(\mathsf{w}_{K,u}(y^n))\\
&\tilde{g}_{n_{K,u}} = g_n
\end{aligned}
$$

Now we analyze the expected distortion of such coder.

$$
\begin{aligned}
E\left(d(\mathsf{y}^n_{K,u}, \tilde{g}_{n_{K,u}}(\tilde{f}_{n_{K,u}}(\mathsf{y}^n_{K,u})))\right) &= E\left(d(\mathsf{y}^n_{K,u}, g_n(f_n(\mathsf{w}_{K,u}(\mathsf{y}^n_{K,u}))))\right)\\
&\leq E\left(d(\mathsf{y}^n_{K,u}, \mathsf{w}_{K,u}(\mathsf{y}^n_{K,u}))\right) + E\left(d(\mathsf{w}_{K,u}(\mathsf{y}^n_{K,u}), g_n(f_n(\mathsf{w}_{K,u}(\mathsf{y}^n_{K,u}))))\right) && (87)
\end{aligned}
$$

The second term in (87) converges to D as n goes to infinity because $\mathsf{w}_{K,u}(\mathsf{y}^n_{K,u})) \sim \mathsf{x}^n$ and f_n, g_n is good for x^n in the expected distortion sense. As for the first term in (87), we show it converges to zero for small u and big Ku as n goes to infinity.

$$
\begin{aligned}
E_{\mathsf{y}_{K,u}}\left(d(\mathsf{y}^n_{K,u}, \mathsf{w}_{K,u}(\mathsf{y}^n_{K,u}))\right) &= E_{\mathsf{y}_{K,u}}(\frac{1}{n}\sum_{i=1}^{n}(\mathsf{y}_{K,u}(i) - \mathsf{w}_{K,u}(\mathsf{y}_{K,u}(i)))^2)\\
&= E_{\mathsf{y}_{K,u}}((\mathsf{y}_{K,u} - \mathsf{w}_{K,u}(\mathsf{y}_{K,u}))^2)\\
&= E_{\mathsf{y}_{K,u}}(1(\mathsf{w}_{K,u}(\mathsf{y}_{K,u}) \leq Ku)(\mathsf{y}_{K,u} - \mathsf{w}_{K,u}(\mathsf{y}_{K,u}))^2)\\
&\qquad + E_{\mathsf{y}_{K,u}}(1(\mathsf{w}_{K,u}(\mathsf{y}_{K,u}) > Ku)(\mathsf{y}_{K,u} - \mathsf{w}_{K,u}(\mathsf{y}_{K,u}))^2)\\
&\leq \frac{u^2}{4} + E_{\mathsf{y}_{K,u}}(1(\mathsf{w}_{K,u}(\mathsf{y}_{K,u}) > Ku)(\mathsf{y}_{K,u} - \mathsf{w}_{K,u}(\mathsf{y}_{K,u}))^2) && (88)\\
&= \frac{u^2}{4} + E_{\mathsf{x}}(1(\mathsf{x} > Ku)\mathsf{x}^2) && (89)\\
&\to 0 \text{ as } u \to 0 \text{ and } Ku \to \infty && (90)
\end{aligned}
$$

(88) is true because if $|\mathsf{w}_{K,u}(\mathsf{y}_{K,u})| \leq Ku$, then the quantization error is no bigger than $\frac{u}{2}$. (89) follows that $\mathsf{w}_{K,u}(\mathsf{y}_{K,u}) \sim \mathsf{x}$. (90) is true because the variance of x is finite. (90) and (87) gives us the desired result that the expected distortion of $\tilde{g}_{n_{K,u}}, \tilde{f}_{n_{K,u}}$ converges to D if u goes to zero, Ku goes to infinity.

B. Constructing a good lossy source coding in the expected distortion sense from a good one in the strong sense

The construction here is a general proof. It works for both continuous, discrete and mixed random variables. By constructing a good lossy source coder in the expected distortion sense from a good lossy coder in the strong sense at the same rate-distortion point (R, D), we can easily see that the rate distortion function in the strong sense is not smaller than the rate distortion function in the expected distortion sense. This fact is used in the proof in Appendix A.

Assume both the first and second order moment of x are finite, i.e. $E(\mathsf{x}) = \mu_{\mathsf{x}} < \infty$ and $E(\mathsf{x}^2) = \sigma_{\mathsf{x}} < \infty$. If f_n, g_n is good in the strong sense for $R(D)$, then we denote by $\Upsilon_n \subseteq \mathcal{R}^n$, the subset the distortion constraint is not satisfied, i.e. $\Upsilon_n = \{x^n \in \mathcal{R}^n : d(x^n, g_n(f_n(x^n)) \geq D + \delta\}$. Denote by $e_n = \overset{\mathsf{x}}{\Pr}(\Upsilon_n)$, then $e_n \to 0$. A good lossy coder might have $g_n(f_n(x^n))$ arbitrarily faraway from x^n for $x^n \in \Upsilon_n$ as pointed out in Section I-B and cause the expected distortion arbitrarily large. We build a new lossy coding system $\tilde{f}_n, \tilde{g}_n$, such that $\tilde{g}_n(\tilde{f}_n(x^n)) = g_n(f_n(x^n))$ for $x^n \notin \Upsilon_n$ and $\tilde{g}_n(\tilde{f}_n(x^n)) = 0$ for $x^n \in \Upsilon_n$. Obviously is good in the strong sense, we only need to show that $\tilde{f}_n, \tilde{g}_n$ is also good in the expected distortion sense. The expected distortion of $\tilde{f}_n, \tilde{g}_n$ is:

$$\begin{aligned} E(d(x^n, \tilde{g}_n(\tilde{f}_n(x^n)))) &= \Pr(x^n \in \Upsilon_n^C) E(d(x^n, \tilde{g}_n(\tilde{f}_n(x^n)))|x^n \in \Upsilon_n^C) + \Pr(x^n \in \Upsilon_n) E(d(x^n, \tilde{g}_n(\tilde{f}_n(x^n)))|x^n \in \Upsilon_n) \\ &\leq (1 - e_n)(D + \delta) + \Pr(x^n \in \Upsilon_n) E(d(x^n, \tilde{g}_n(\tilde{f}_n(x^n)))|x^n \in \Upsilon_n) \\ &= (1 - e_n)(D + \delta) + \Pr(x^n \in \Upsilon_n) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n) \end{aligned} \tag{91}$$

Now we upper bound the second term, first according to the weak law of large numbers and the variance and the mean of x are finite, we know that for any $\epsilon > 0$, there exists $n_\epsilon < \infty$, s.t for all $n > n_\epsilon$:

$$\overset{\mathsf{x}}{\Pr}(|\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 - \sigma_{\mathsf{x}}| > \epsilon) < \epsilon. \tag{92}$$

This implies that for any subset $\Gamma \in \mathcal{R}^n$ with measure $\overset{\mathsf{x}}{\Pr}(\Gamma) \geq 1 - \epsilon$, then there is a subset $\Gamma_1 \subseteq \Gamma$, such that $\overset{\mathsf{x}}{\Pr}(\Gamma_1) \geq 1 - 2\epsilon$ and for all $x^n \in \Gamma_1$: $|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \sigma_{\mathsf{x}}| \leq \epsilon$.

From the definition of e_n, we know that for large enough n, $e_n = \overset{\mathsf{x}}{\Pr}(\Upsilon_n) < \epsilon$ or equivalently $\overset{\mathsf{x}}{\Pr}(\Upsilon_n^C) \geq 1 - \epsilon$. From the above discussion, there exists subset $\Gamma_1 \in \Upsilon_n^C$, such that $\overset{\mathsf{x}}{\Pr}(\Gamma_1) \geq 1 - 2\epsilon$ and for all $x^n \in \Gamma_1$: $|\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \sigma_{\mathsf{x}}| \leq \epsilon$. So the expectation of the mean variance of x^n can be decomposed:

$$\begin{aligned} \sigma_{\mathsf{x}} &= E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2) \\ &= \Pr(x^n \in \Upsilon_n^C) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n^C) + \Pr(x^n \in \Upsilon_n) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n) \\ &\geq \Pr(x^n \in \Gamma_1) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Gamma_1) + \Pr(x^n \in \Upsilon_n) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n) \\ &\geq (1 - \epsilon)(\sigma_{\mathsf{x}} - \epsilon) + \Pr(x^n \in \Upsilon_n) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n) \end{aligned}$$

Hence:

$$\Pr(x^n \in \Upsilon_n) E(\frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i^2 | x^n \in \Upsilon_n) \leq \epsilon(1 + \sigma_{\mathsf{x}}) \tag{93}$$

Substituting (91) into (93), we have:

$$E(d(x^n, \tilde{g}_n(\tilde{f}_n(x^n)))) \leq (1 - e_n)(D + \delta) + \epsilon(1 + \sigma_{\mathsf{x}}) \leq D + \delta + \epsilon(1 + \sigma_{\mathsf{x}})$$

Note that the above is true for all ϵ and δ, so we can let both be arbitrarily small and the expected distortion of $\tilde{f}_n, \tilde{g}_n$ is arbitrarily close to D. Hence we just constructed a good lossy coding system in the expected distortion sense from a good lossy coding system in the strong sense. □

C. Proof of the simple bounds: proof of Propositions 1, 2, 3 and 4

Proof of Proposition 1:

To show $R(D,\Xi(p,\sigma^2)) \geq R(\frac{D}{\sigma^2},\Xi(p,1))$, we only need to construct a sequence of good, in the strong sense of rate distortion in (2), encoder/decoder pairs $(f'_n,g'_n),\ n=1,2,...$, for $\Xi(p,1)$ from that for $\Xi(p,\sigma^2)$, $(f_n,g_n),\ n=1,2,....$ Let f'_n and g'_n be as follows, for all $x^n\in\mathcal{X}^n$ and $a^{nR}\in\{0,1\}^{nR}$:

$$f'_n(x^n)=f_n(\sigma x^n),\quad g'_n(a^{nR})=\frac{1}{\sigma}g_n(a^{nR})$$

So for $\mathsf{x}\sim\Xi(p,0,1)$

$$\begin{aligned}\Pr\left(d(\mathsf{x}^n,g'_n(f'_n(\mathsf{x}^n)))\geq\frac{D+\delta}{\sigma^2}\right) &= \Pr\left(d(\mathsf{x}^n,\frac{1}{\sigma}g_n(f_n(\sigma\mathsf{x}^n)))\geq\frac{D+\delta}{\sigma^2}\right)\\ &= \Pr\left(d(\sigma\mathsf{x}^n,g_n(f_n(\sigma\mathsf{x}^n)))\geq D+\delta\right) \qquad (94)\end{aligned}$$

where (94) is because the distortion measure $d(x,y)=(x-y)^2$ in this paper.

Obviously for $\mathsf{x}\sim\Xi(p,1)$, $\sigma\mathsf{x}\sim\Xi(p,\sigma^2)$, and if f_n and g_n are good in the strong sense, defined in (2), for $\Xi(p,\sigma^2)$, then for all $\delta>0$:

$$\lim_{n\to\infty}\Pr\left(d(\sigma\mathsf{x}^n,g_n(f_n(\sigma\mathsf{x}^n))\geq D+\delta\right)=0. \qquad (95)$$

Combining (94) and (95), we have:

$$\lim_{n\to\infty}\Pr\left(d(\mathsf{x}^n,g'_n(f'_n(\mathsf{x}^n)))\geq\frac{D+\delta}{\sigma^2}\right)=0.$$

Notice that δ is an arbitrary positive number and σ is constant, we just show that $R(D,\Xi(p,\sigma^2))\geq R(\frac{D}{\sigma^2},\Xi(p,1))$. Similarly we can show that $R(D,\Xi(p,\sigma^2))\leq R(\frac{D}{\sigma^2},\Xi(p,1))$. This complete the proof that $R(D,\Xi(p,\sigma^2))=R(\frac{D}{\sigma^2},\Xi(p,1))$. □

Proof of Proposition 2: for a Bernoulli-Gaussian random sequence x^n, by Definition 1, we know that $\mathsf{x}_i=\mathsf{b}_i\times\mathsf{s}_i$, $\mathsf{b}_i\sim Bernoulli-p$ and $\mathsf{s}_i\sim N(0,1)$ are i.i.d random variables. The encoder f_n works as follows. It is consisted of two parts. First the encoder encode b^n *losslessly* using a fixed length code-book. Then the encoder encode *lossily* the subsequence of s^n where $\mathsf{b}_i\neq 0$ by applying standard Gaussian lossy source coding.

We now describe the coding scheme f_n,g_n, in details. If b^n is ϵ_1-strong typical, and write $1(b^n)$ as the number of 1's in sequence b^n. i.e.:

$$b^n\in B^n_{\epsilon_1}\triangleq\{b^n\in\{0,1\}^n:|\frac{1(b^n)}{n}-p|\leq\epsilon_1\}.$$

then f_n one-to-one maps b^n to a binary sequence of length $n(H(p)+\tau(\epsilon_1))$ excluding the all zero signal, otherwise $b^n\notin B^n_{\epsilon_1}$, f_n sends the all zero signal, where $\tau(\epsilon_1)\to 0$ if $\epsilon_1\to 0$, this is guaranteed by the standard lossless source coding theorem. Obviously for all $\epsilon_1>0$:

$$\lim_{n\to\infty}\overset{\mathsf{b}}{\Pr}(\mathsf{b}^n\notin B^n_{\epsilon_1})=0 \qquad (96)$$

Now for each $x^n=b^n\times s^n$, if $b^n\in B^n_{\epsilon_1}$, we know that $n(p-\epsilon_1)\leq 1(b^n)\leq n(p+\epsilon_1)$. Denote by a new sequence $\tilde{s}_1,...\tilde{s}_{1(b^n)}$ the non zero entries of x^n. Then the encoder f_n passes $\tilde{s}^{1(b^n)}$ to a good Gaussian lossy encoder-decoder pair $\tilde{f}_{1(b^n)},\tilde{g}_{1(b^n)}$ with rate $R(\frac{D}{p},N(0,1))$ for a sequence of length $1(b^n)$. If output of $\tilde{f}_{1(b^n)}$, when $1(b^n)<n(p+\epsilon_1)$, is shorter than $n(p+\epsilon_1)R(\frac{D}{p},N(0,1))$, f_n just pad zeros at the end. The total block length for x^n is

$$n(H(p)+\tau(\epsilon_1))+n(p+\epsilon_1)R(\frac{D}{p},N(0,1)). \qquad (97)$$

If the output form the encoder is not a all zero sequence, the decoder g_n first looks at the first $n(H(p)+\tau(\epsilon_1))$ bits and recover b^n exactly and hence $1(b^n)$. Then g_n discards the padded zeros at the end and pass the rest to the Gaussian lossy decoder $\tilde{g}_{1(b^n)}$ with rate $R(\frac{D}{p},N(0,1))$ for a sequence of length $1(b^n)$. Then g_n put the outputs of $\tilde{g}_{1(b^n)}$ to the non-zero locations of b^n one by one. By using the coding system described above, we have for $b^n \in B^n_{\epsilon_1}$,

$$nd(x^n, g_n(f_n(x^n))) = 1(b^n)d(\tilde{s}^{1(b^n)}, \tilde{g}_{1(b^n)}(\tilde{f}_{1(b^n)}(\tilde{s}^{1(b^n)}))) \tag{98}$$

and because s^n and b^n are independent and the coding system $\tilde{f}_{1(b^n)}, \tilde{g}_{1(b^n)}$ is good, for all fixed $b^n \in B^n_{\epsilon_1}$, for all $\delta_0 > 0$:

$$\lim_{n\to\infty} \overset{\tilde{\mathsf{s}}}{\Pr}\left(d(\tilde{\mathsf{s}}^{1(b^n)}, \tilde{g}_{1(b^n)}(\tilde{f}_{1(b^n)}(\tilde{\mathsf{s}}^{1(b^n)}))) \geq \frac{D}{p} + \delta_0\right) = 0. \tag{99}$$

Now we evaluate the performance of f_n, g_n, for all $\delta_1 > 0$:

$$\begin{aligned}
\lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1\right) &\leq \lim_{n\to\infty} \{\overset{\mathsf{b}}{\Pr}\left(\mathsf{b}^n \notin B^n_{\epsilon_1}\right) + \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 | \mathsf{b}^n \in B^n_{\epsilon_1}\right)\} && (100)\\
&= \lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 | \mathsf{b}^n \in B^n_{\epsilon_1}\right) && (101)\\
&= \lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}\left(d(\tilde{\mathsf{s}}^{1(b^n)}, \tilde{g}_{1(b^n)}(\tilde{f}_{1(b^n)}(\tilde{\mathsf{s}}^{1(b^n)}))) \geq \frac{n(D+\delta_1)}{1(\mathsf{b}^n)} | \mathsf{b}^n \in B^n_{\epsilon_1}\right) && (102)\\
&\leq \lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}\left(d(\tilde{\mathsf{s}}^{1(b^n)}, \tilde{g}_{1(b^n)}(\tilde{f}_{1(b^n)}(\tilde{\mathsf{s}}^{1(b^n)}))) \geq \frac{n(D+\delta_1)}{n(p+\epsilon_1)} | \mathsf{b}^n \in B^n_{\epsilon_1}\right) && (103)\\
&= \lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}\left(d(\tilde{\mathsf{s}}^{1(b^n)}, \tilde{g}_{1(b^n)}(\tilde{f}_{1(b^n)}(\tilde{\mathsf{s}}^{1(b^n)}))) \geq \frac{D}{p} + \frac{\delta_1 p - D\epsilon_1}{p(p+\epsilon_1)} | \mathsf{b}^n \in B^n_{\epsilon_1}\right)\\
&= 0 && (104)
\end{aligned}$$

(100) is because for events A and B, $\Pr(A) = \Pr(A, B) + \Pr(A, B^c) \leq \Pr(B) + \Pr(A|B^c)$. (101) is true because of (96). (98) implies (102). (103) is true because $1(b^n) \leq n(p + \epsilon_1)$ if $b^n \in B^n_{\epsilon_1}$. Finally, for any δ_1, by letting ϵ_1 small enough, hence $\frac{\delta_1 p - D\epsilon_1}{p(p+\epsilon_1)} > 0$ and by (99) and the fact that $\tilde{\mathsf{s}}^{1(\mathsf{b}^n)}$ is induced by x^n, we have (104).

(97) together with (104) implies that:

$$R(D, p) \leq H(p) + pR(\frac{D}{p}, N(0, 1)) + \tau(\epsilon_1) + \epsilon_1 R(\frac{D}{p}, N(0, 1)).$$

Notice that we can pick ϵ_1 arbitrarily small and hence $\tau(\epsilon_1)$ arbitrarily small, we have $R(D, p) \leq H(p) + pR(\frac{D}{p}, N(0, 1)) = H(p) + pR(D, N(0, p))$. □

Proof of Proposition 3: this is a direct corollary of the upper bound in Corollary 2. Notice that the variance of a Bernoulli-Gaussian random variable $\Xi(p, 1)$ is p, so according to Corollary 2, if $\Xi(p, 1)$ is a continuous random variable, we would have:

$$R(D, p) \leq \max\{\frac{1}{2}\log\frac{p}{D}, 0\} = R(D, N(0, p)) \tag{105}$$

The technicality here is that $\Xi(p, 1)$ is not a continuous random variable, but the fix is quite easy. Let random variable y_m be the p-mixture of a Gaussian $N(0, 1)$ and a uniformly distributed random variable on $[-\frac{1}{m}, \frac{1}{m}]$. i.e. with probability $1 - p$, $\mathsf{y}_m \sim N(0, 1)$ and with probability p, $\mathsf{y}_m \sim U[-\frac{1}{m}, \frac{1}{m}]$. The pdf of y_m is

$$p_{\mathsf{y}_m}(y) = \{ \begin{array}{cc} \frac{1-p}{2}e^{\frac{y^2}{2}}, & |y| > \frac{1}{m}, \\ \frac{1-p}{2}e^{\frac{y^2}{2}} + \frac{pm}{2}, & |y| \leq \frac{1}{m}. \end{array} \tag{106}$$

Obviously, y_m is a continuous random variable with variance $p + \frac{1}{3m^2}$. Now according to Corollary 2, we know that the rate distortion function for y_m, $R_{\mathsf{y}_m}(D)$ is upper bounded by

$$\max\{\frac{1}{2}\log\frac{p}{D}, 0\} = R(D, N(0, p + \frac{1}{3m^2})). \tag{107}$$

Now we upper bound $R(D,p)$ by constructing a good randomized lossy coding system for $x \sim \Xi(p,1)$ in the average sense, f_n^u, g_n^u, from a good lossy coding system for y_m, f_n^y, g_n^y. Given $x^n \sim \Xi(p,1)$, f_n^u applies the following operation on it, for $i = 1,2,...,n$, let

$$z_i = \{ \begin{array}{cc} x_i, & x_i \neq 0, \\ x_i + u_i, & x_i = 0 \end{array} \tag{108}$$

where u_i's are independent and $u_i \sim U[\frac{-1}{m}, \frac{1}{m}]$ is a uniform distributed random variable. It is clear that z_i has the same distribution as y_m. Now f_n^u passes z^n to encoder f_n^y. The decoder $g_n^u = g_n^y$. Now we analyze the performance of the coding system f_n^u, g_n^u.

First, because f_n^y, g_n^y is a good in the strong sense for y_m, we have, for any $\delta_1 > 0$:

$$\lim_{n\to\infty} \overset{z}{\Pr}\left(d(z^n, g_n^y(f_n^y(z^n)) \geq D + \delta_1\right) = 0.$$

From the construction of f_n^u, g_n^u, we know that $g_n^y(f_n^y(z^n)) = g_n^u(f_n^u(x^n))$ a.s., where z^n is induced from x^n and u^n, denote $g_n^u(f_n^u(x^n))_i$ or equivalently $g_n^y(f_n^y(z^n))_i$ by w_i, so :

$$\lim_{n\to\infty} \overset{x,u}{\Pr}\left(d(z^n, g_n^u(f_n^u(x^n)) \geq D + \delta_1)\right) = \lim_{n\to\infty} \overset{x,u}{\Pr}\left(d(z^n, w^n) \geq D + \delta_1)\right) = 0. \tag{109}$$

Secondly, from the construction of z^n, we know that for all i, $|x_i - z_i| \leq \frac{1}{m}$ a.s.. So we have a.s.:

$$\begin{aligned} d(x^n, g_n^u(f_n^u(x^n))) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - w_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - z_i + z_i - w_i)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}(|x_i - z_i| + |z_i - w_i|)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{m} + |z_i - w_i|)^2 \\ &\leq \frac{1}{m} + \frac{1}{n}\sum_{i=1}^{n}(z_i - w_i)^2 + \frac{2}{nm}\sum_{i=1}^{n}|z_i - w_i| \end{aligned} \tag{110}$$

By the Cauchy-Schwartz inequality, a.s.:

$$(\sum_{i=1}^{n}|z_i - w_i|)^2 \leq (\sum_{i=1}^{n}|z_i - w_i|^2)(\sum_{i=1}^{n}1)$$

hence a.s.

$$\frac{1}{n}\sum_{i=1}^{n}|z_i - w_i| \leq \sqrt{\frac{1}{n}(\sum_{i=1}^{n}|z_i - w_i|^2)}. \tag{111}$$

Now for a realization of x^n and u^n: x^n and u^n, the induced realization of z^n and $g_n^u(f_n^u(x^n))$ are z^n and w^n respectively. If $d(z^n, w^n) < D + \delta_1$, then combining (110) and (111), we have:

$$\begin{aligned} d(x^n, g_n^u(f_n^u(x^n))) &\leq \frac{1}{m} + (D + \delta_1) + \frac{2\sqrt{D+\delta_1}}{m} \\ &\leq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m} \end{aligned}$$

This means that

$$\overset{x,u}{\Pr}\left(d(z^n, w^n) \geq D + \delta_1\right) \geq \overset{x,u}{\Pr}\left(d(x^n, w^n) \geq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m}\right) \tag{112}$$

The above coding system is a randomized coding system where the performance is measured under the distribution of the "dithering" random variable u. Now if we take the above average "dithering", i.e.: for each $x^n \in R^n$,

$$\text{if} \qquad \overset{\mathsf{u}}{\Pr}\left(d(x^n, \mathsf{w}^n) \geq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m} \right) < 1,$$

there exists $u^n(x^n) \in [\frac{-1}{m}, \frac{1}{m}]$ and of course $u_i(x^n) = 0$ if $x_i \neq 0$, such that the distortion between x^n and the output of the the lossy coding system $f_n^{\mathsf{y}}, g_n^{\mathsf{y}}$ with input $x^n + u^n(x^n)$ is no bigger than $D + \delta_1 + \frac{1+2\sqrt{D+\delta_1}}{m}$:

$$d(x^n, g_n^{\mathsf{y}}(f_n^{\mathsf{y}}(x^n + u^n(x^n)))) \leq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m}$$

$$\text{otherwise, simply let} \qquad u^n(x^n) = 0.$$

Finally, let f_n and g_n be such that, for all x^n, $f_n(x^n) = f_n^{\mathsf{y}}(x^n + u^n(x^n))$ and $g_n = g_n^{\mathsf{y}}$. The construction of g_n, f_n implies that

$$\overset{\mathsf{x}}{\Pr}(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m}) \leq \overset{\mathsf{x},\mathsf{u}}{\Pr}\left(d(\mathsf{x}^n, \mathsf{w}^n) \geq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m} \right) \tag{113}$$

Now combining (109), (112) and (113), we have:

$$\lim_{n\to\infty} \overset{\mathsf{x}}{\Pr}(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m}) = 0$$

Note that the rate of the coding system is $R_{\mathsf{y}_m}(D)$ which is upper bounded by $R(D, N(0, p + \frac{1}{3m^2}))$ in (107). So

$$R(D + \delta_1 + \frac{1 + 2\sqrt{D+\delta_1}}{m}, p) \leq R(D, N(0, p + \frac{1}{3m^2})) \tag{114}$$

while (114) is true for all $\delta_1 > 0$ and $m \in \mathcal{N}$. Note that the Gaussian rate distortion function $R(D, N(0, \sigma^2))$ is continuous in σ^2 and the Bernoulli-Gaussian rate distortion function $R(D, p)$ is monotonically decreasing and bounded in D, hence continuous with measure 1. By letting $m \to \infty$ and $\delta_1 \to 0$, we have: $R(D, p) \leq R(D, N(0, p))$. □

Proof of Proposition 4: For a good lossy coding system f_n, g_n for Bernoulli-Gaussian sequence $\mathsf{x}^n = \mathsf{b}^n \times \mathsf{s}^n \sim \Xi(p, 1)$ defined in Definition 1 and distortion constraint D, the rate is $R(D, p)$, i.e.

$$f_n : R^n \to \{0,1\}^{nR(D,p)} \qquad g_n : \{0,1\}^{nR(D,p)} \to \mathcal{R}^n, \qquad \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1\right) = e_n$$

$$\text{and for all } \delta_1 > 0\text{: } \lim_{n\to\infty} e_n = 0. \tag{115}$$

We use the same notations as those in the proof of Proposition 2. We construct a good length $m_n \in [n(p - \epsilon_1), n(p + \epsilon_1)]$ lossy source coding system $\tilde{f}_{m_n}, \tilde{g}_{m_n}$ for $\tilde{\mathsf{s}}^{m_n} \sim N(0, p)$ under the same distortion constraint D, where m_n will be determined later. First we decompose e_n, by (96), we know that there exists $n_{\epsilon_1} < \infty$, such that for all $n > n_{\epsilon_1}$, $\overset{\mathsf{x}}{\Pr}(\mathsf{b}^n \in B^n_{\epsilon_1}) \geq \frac{1}{2}$, so for all $n > n_{\epsilon_1}$:

$$\begin{aligned}
e_n &= \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1\right) \\
&\geq \overset{\mathsf{x}}{\Pr}\left(\mathsf{b}^n \in B^n_{\epsilon_1}\ ,\ d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1\right) && (116)\\
&= \sum_{b^n \in B^n_{\epsilon_1}} \overset{\mathsf{x}}{\Pr}\left(\mathsf{b}^n = b^n\ ,\ d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1\right) && (117)\\
&= \overset{\mathsf{x}}{\Pr}(\mathsf{b}^n \in B^n_{\epsilon_1}) \sum_{b^n \in B^n_{\epsilon_1}} \frac{\overset{\mathsf{x}}{\Pr}(\mathsf{b}^n = b^n)}{\overset{\mathsf{x}}{\Pr}(\mathsf{b}^n \in B^n_{\epsilon_1})} \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 | \mathsf{b}^n = b^n\right) && (118)\\
&\geq \frac{1}{2} \sum_{b^n \in B^n_{\epsilon_1}} \phi(b^n) \overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 | \mathsf{b}^n = b^n\right) && (119)
\end{aligned}$$

(116), (117) and (118) are obvious, in (119), we denote $\phi(b^n)$ by $\frac{\overset{\mathsf{x}}{\Pr}(\mathsf{b}^n=b^n)}{\overset{\mathsf{x}}{\Pr}(\mathsf{b}^n\in B^n_{\epsilon_1})}$. Notice that $\phi()$ is a probability measure on $B^n_{\epsilon_n}$. Hence there exists $\bar{b}^n \in \mathcal{B}^n$, write $1(\bar{b}^n) = m_n \in [n(p-\epsilon_1), n(p+\epsilon_1)]$, such that:

$$\overset{\mathsf{x}}{\Pr}\left(d(\mathsf{x}^n, g_n(f_n(\mathsf{x}^n))) \geq D + \delta_1 | \mathsf{b}^n = \bar{b}^n\right) \leq 2e_n. \tag{120}$$

We bound the distortion of x^n as follows, let $l_1 < l_2 < ... < l_{m_n}$, $L = \{l_1, ...l_{m_n}\}$, be the positions of the non-zero elements of $\bar{b}^n$,

$$nd(x^n, g_n(f_n(x^n))) \geq \sum_{i=1}^{m_n} (x_{l_i} - g_n(f_n(x^n))_{l_i})^2. \tag{121}$$

Substituting (121) into (120), we have:

$$\overset{\mathsf{x}}{\Pr}\left(\sum_{i=1}^{m_n} (\mathsf{x}_{l_i} - g_n(f_n(x^n))_{l_i})^2 \geq n(D+\delta_1) | \mathsf{b}^n = \bar{b}^n\right) \leq 2e_n. \tag{122}$$

Now we are ready to construct a good lossy source coding system $\tilde{f}_{m_n}, \tilde{g}_{m_n}$ for $\mathsf{s}^{m_n} \sim N(0,p)$. The encoder $\tilde{f}_{m_n}$ works as follows, for any sequence $\tilde{s}^{m_n} \in \mathcal{R}^{m_n}$, $\tilde{f}_{m_n}(\tilde{s}^{m_n}) = f_n(T(s^{m_n}))$, for a binary sequence $a^{R(D,p)n} \in \{0,1\}^{R(D,p)n}$: $\tilde{g}_{m_n}(a^{R(D,p)n}) = T^{-1}g_n(a^{R(D,p)n}))$, where T is a one-to-one map from $\mathcal{R}^{m_n}$ to $\mathcal{R}^n$:

$$\begin{aligned}
T(\tilde{s}^{m_n}) &= s^n, \text{ where } s_{l_i} = \tilde{s}_i, \quad i = 1, 2, ..., m_n \text{ and } s_i = 0, \ i \notin L \\
T^{-1}(s^n) &= \tilde{s}^{m_n}, \text{ where } \tilde{s}_i = s_{l_i}, \quad i = 1, 2, ..., m_n
\end{aligned}$$

$\mathsf{x}^n = \mathsf{b}^n \times \mathsf{s}^n$, so if $\mathsf{b}^n = \bar{b}^n$ then $\mathsf{x}_i = 0$ for all $i \notin L$, and by the memorylessness of x^n. We have:

$$\begin{aligned}
\overset{\tilde{\mathsf{s}}}{\Pr}\left(m_n d(\mathsf{s}^{m_n}, \tilde{g}_{m_n}(\tilde{f}_{m_n}(\mathsf{s}^{m_n}))) \geq n(D+\delta_1)\right) &= \overset{\tilde{\mathsf{s}}}{\Pr}\left(\sum_{i=1}^{m_n} (\tilde{\mathsf{s}}_i - \tilde{g}_{m_n}(\tilde{f}_{m_n}(\tilde{\mathsf{s}}^n))_i)^2 \geq n(D+\delta_1)\right) \\
&= \overset{\mathsf{x}}{\Pr}\left(\sum_{i=1}^{m_n} (\mathsf{x}_{l_i} - g_n(f_n(x^n))_{l_i})^2 \geq n(D+\delta_1) | \mathsf{b}^n = \bar{b}^n\right) \\
&\leq 2e_n.
\end{aligned} \tag{123}$$

where the inequality is by (122). Notice that $m_n = 1(\tilde{b}^n) \in [n(p-\epsilon_1), n(p+\epsilon_1)]$, so $\frac{n}{m_n} \in [\frac{1}{p+\epsilon_1}, \frac{1}{p-\epsilon_1}]$. So (123) and (115) tells us:

$$\begin{aligned}
0 &= \lim_{n\to\infty} \overset{\tilde{\mathsf{s}}}{\Pr}\left(m_n d(\mathsf{s}^{m_n}, \tilde{g}_{m_n}(\tilde{f}_{m_n}(\mathsf{s}^{m_n}))) \geq n(D+\delta_1)\right) \\
&= \lim_{n\to\infty} \overset{\tilde{\mathsf{s}}}{\Pr}\left(d(\mathsf{s}^{m_n}, \tilde{g}_{m_n}(\tilde{f}_{m_n}(\mathsf{s}^{m_n}))) \geq \frac{n}{m_n}(D+\delta_1)\right) \\
&\geq \lim_{n\to\infty} \overset{\tilde{\mathsf{s}}}{\Pr}\left(d(\mathsf{s}^{m_n}, \tilde{g}_{m_n}(\tilde{f}_{m_n}(\mathsf{s}^{m_n}))) \geq \frac{1}{p-\epsilon_1}(D+\delta_1)\right).
\end{aligned} \tag{124}$$

The encoder decoder pair $\tilde{f}_{m_n}, \tilde{g}_{m_n}$ use $nR(D,p)$ bits, so the rate of this coding system is $\frac{nR(D,p)}{m_n} \leq \frac{R(D,p)}{p-\epsilon_1}$. (124) is true for all δ_1 and ϵ_1, by letting $\epsilon_1 \to 0$, we just construct a rate $\frac{R(D,p)}{p}$, distortion $\frac{D}{p}$ coding system for i.i.d Gaussian random variables $\tilde{\mathsf{s}}^{m_n} \sim N(0,1)$. From Corollary 1 we know that $\frac{R(D,p)}{p} \geq R(\frac{D}{p}, N(0,1))$, i.e.

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

□

D. *Strong Typical Gaussian Sequences*

In this appendix we define and investigate the properties of the so called strong typical Gaussian sequences. For a sequence $s^n \in \mathcal{R}^n$, for a real number $T \in \mathcal{R}$, the empirical l-th moment of entries in s^n within interval $[T, \infty]$ is denoted by

$$n^l_{s^n}(T) = \frac{\sum_{i=1}^n 1(s_i > T) s_i^l}{n}.$$

Definition 4: ϵ-typical Gaussian sequences: A sequence s^n is said to be ϵ typical for $N(0,1)$, if the followings are true: for any real number $T \geq -\infty$,

$$\max_{l=0,1,2} \left\{ \sup_T \left| n^l_{s^n}(T) - \int_T^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| < \epsilon \right\} \tag{125}$$

The ϵ-typical set of $N(0,1)$ is denoted by $S_\epsilon(n)$, similar to the strong typical set for random sequences with finite alphabet, we have the following concentration lemma. Note that the convergence is uniform convergence, in the sense that we ask the sequence to be typical for all real numbers T simultaneously.

An almost equivalent "double-sided" definition of ϵ-typical Gaussian sequence is as follows. First, for any $-\infty \leq S \leq T \leq \infty$, we denote by

$$n^{l*}_{s^n}(S,T) = \frac{\sum_{i=1}^n 1(s_i \in [S,T]) s_i^l}{n}.$$

Similar to that in Definition 4, we define the typical set $S^*_\epsilon(n)$ as the set of all sequence s^n, s.t.

$$\max_{l=0,1,2} \left\{ \sup_{S \leq T} \left| n^*_{s^n}(S,T) - \int_S^T s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| < \epsilon \right\} \tag{126}$$

We now illustrate the equivalence of the two typical sets $S_\epsilon(n)$ and $S^*_\epsilon(n)$. First, obviously $S^*_\epsilon(n) \subseteq S_\epsilon(n)$. Secondly,

$$\begin{aligned}
\sup_{S \leq T} \left| n^{l*}_{s^n}(S,T) - \int_S^T s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| &= \sup_{S \leq T} \left| n^l_{s^n}(S) - n^l_{s^n}(T) - \int_S^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds + \int_T^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| \\
&\leq \sup_{S \leq T} \left| n^l_{s^n}(S) - \int_S^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| + \left| n^l_{s^n}(T) - \int_T^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| \\
&\leq 2 \sup_T \left| n^l_{s^n}(T) - \int_T^\infty s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right|
\end{aligned}$$

This means $S_\epsilon(n) \subseteq S^*_{2\epsilon}(n)$, so the concentration of the "double-sided" and the "one-sided" typical sets are equivalent. We use the latter definition of ϵ-typical set in the main body of the paper. However, for the sake of simplicity of notations, we prove the concentration of the ϵ-typical set of the "one-sided" definition.

Lemma 6: Concentration of Gaussian sequences: for i.i.d $N(0,1)$ random sequence s^n, for all $\epsilon > 0$

$$\lim_{n \to \infty} \Pr(s^n \in S_\epsilon(n)) = 1 \tag{127}$$

Proof: we give a sketch of the proof here. The idea is to first quantize the real line for the Gaussian $N(0,1)$ random variable then apply the concentration result for i.i.d discrete finite random sequences. The quantization goes as follows, we study the following intervals: $\{(-\infty, -K\omega], [-K\omega, -(K-1)\omega], ..., [(K-1)\omega, K\omega], [K\omega, \infty]\}$, i.e. the end points of the intervals are defined as follows: for an integer j within range $[-K-1, K+1]$ we denote $\omega(j) = j\omega$ if $j = -K, ..., K$ and $\omega(-K-1) = -\infty$ and $\omega(K+1) = \infty$. We can obviously let ω be small enough and K be big enough such that the following two integrals are true for all $j = -K-1, ...K$

$$\left| \int_{\omega(j)}^{\omega(j+1)} s^l \frac{1}{\sqrt{2\pi}} e^{\frac{-s^2}{2}} ds \right| < \frac{\epsilon}{2} \text{ for } l = 0, 1, 2. \tag{128}$$

We let $S_\epsilon^{\omega,K}(n)$ be the set that the typicality condition in (126) is true for $T=\omega(j)$ for all $j\in\{-K-1,...,K+1\}$ simultaneously, i.e.

$$S_\epsilon^{\omega,K}(n)=\{s^n: \max_{l=0,1,2}\left\{\sup_{j=-K-1,\dots,K+1}\left|n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|<\epsilon\right\}\} \tag{129}$$

We show that

$$\lim_{n\to\infty}\Pr(\boldsymbol{s}^n\in S_\epsilon^{\omega,K}(n))=1 \tag{130}$$

This is true because from the weak law of large numbers we know that for $l=0,1,2$:

$$\lim_{n\to\infty}\Pr\left(\left|n_{\boldsymbol{s}^n}^l(T)-\int_T^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|<\epsilon\right)=1 \tag{131}$$

for all $T\in\mathcal{R}\bigcup\{-\infty,\infty\}$, in particular for all $T=\omega(j)$, $j=-K-1,...,K+1$. This is a finite set, so

$$\begin{aligned}&\lim_{n\to\infty}\Pr(\boldsymbol{s}^n\in S_\epsilon^{\omega,K}(n))\\&=\lim_{n\to\infty}\Pr(\max_{l=0,1,2}\left\{\sup_{j=-K-1,\dots,K+1}\left|n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|\right\}<\epsilon)\\&=1\end{aligned} \tag{132}$$

(130) is proved. In particular:

$$\lim_{n\to\infty}\Pr(\boldsymbol{s}^n\in S_{\frac{\epsilon}{2}}^{\omega,K}(n))=1 \tag{133}$$

Now we are ready to use (133) to prove the lemma.

For any s^n and a real number $T\in[\omega(j),\omega(j+1)]$, $j\in\{-K-1,...,K+1\}$, then obviously

$$n_{s^n}^l(T)\in[n_{s^n}^l(\omega(j+1)),n_{s^n}^l(\omega(j))],\text{ so for }l=0,1,2:$$

$$\begin{aligned}n_{s^n}^l(T)-\int_T^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds &\le n_{s^n}^l(\omega(j))-\int_T^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\\&= n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds+\int_{\omega(j)}^{T}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\\&\le n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds+\int_{\omega(j)}^{\omega(j+1)}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\\&\le \left|n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|+\frac{\epsilon}{2},\end{aligned} \tag{134}$$

where (134) follows (128), similarly we have for $l=0,1,2$:

$$n_{s^n}^l(T)-\int_T^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds \ \ge\ -\left|n_{s^n}^l(\omega(j+1))-\int_{\omega(j+1)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|-\frac{\epsilon}{2}. \tag{135}$$

(134) and (135) tells us that for $l=0,1,2$:

$$\left|n_{s^n}^l(T)-\int_T^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right| \ \le\ \sup_{j=-K-1,\dots,K+1}\left|n_{s^n}^l(\omega(j))-\int_{\omega(j)}^{\infty}s^l\frac{1}{\sqrt{2\pi}}e^{\frac{-s^2}{2}}ds\right|+\frac{\epsilon}{2}. \tag{136}$$

Notice the definitions of $S_\epsilon(n)$ and $S_{\frac{\epsilon}{2}}^{\omega,K}(n)$, (136) implies that $S_\epsilon(n)\supseteq S_{\frac{\epsilon}{2}}^{\omega,K}(n)$, hence:

$$\lim_{n\to\infty}\Pr(\boldsymbol{s}^n\in S_\epsilon(n)) \ \ge\ \lim_{n\to\infty}\Pr(\boldsymbol{s}^n\in S_{\frac{\epsilon}{2}}^{\omega,K}(n))=1$$

The lemma is proved. □

E. Properties of $(u+v)D(\frac{u}{u+v}\|p)$

In this section we show some properties of $(u+v)D(\frac{u}{u+v}\|p)$, summarized in the following lemma.

Lemma 7: If $u, v \geq 0$, $\frac{u}{u+v} > p$, then $(u+v)D(\frac{u}{u+v}\|p)$ is monotonically increasing with u and monotonically decreasing with v.

Proof: First, both $\frac{u}{u+v}$ and $D(\frac{u}{u+v}\|p)$ are positive and monotonically increasing with u if $\frac{u}{u+v} > p$. Hence $(u+v)D(\frac{u}{u+v}\|p)$ is monotonically increasing with u.

Secondly, using basic calculus, we have:

$$\begin{aligned}
\frac{d(u+v)D(\frac{u}{u+v}\|p)}{dv} &= \frac{d\left(u\log(\frac{u}{(u+v)p}) + v\log(\frac{v}{(u+v)(1-p)})\right)}{dv} \\
&= -\frac{u}{u+v} - \frac{v}{u+v} + 1 + \log\left(\frac{v}{(u+v)(1-p)}\right) \\
&= \log\left(\frac{1-\frac{u}{u+v}}{1-p}\right) \\
&< 0 \qquad (137)
\end{aligned}$$

The last inequality is true because $\frac{u}{u+v} > p$ hence $1 - \frac{u}{u+v} < 1 - p$. □