Complexity and Rate-Distortion Tradeoff via Successive Refinement

Albert No, Amir Ingber and Tsachy Weissman
Department of Electrical Engineering, Stanford University
Email: {albertno, ingber, tsachy}@stanford.edu

Abstract—We demonstrate how successive refinement ideas can be used in point-to-point lossy compression problems in order to reduce complexity. We show two examples, the binary-Hamming and quadratic-Gaussian cases, in which a layered code construction results in a low complexity scheme that attains optimal performance. For example, when the number of layers grows with the block length \( n \), we show how to design an \( O(n \log(n)) \) algorithm that asymptotically achieves the rate distortion bound. We then show that with the same scheme, used with a fixed number of layers, successive refinement is achieved in the classical sense, and at the same time the second order performance (i.e. dispersion) is also tight.

Index Terms—Binary source, complexity, Gaussian source, rate-distortion, refined strong covering lemma, source dispersion, sparse regression code, successive refinement.

I. INTRODUCTION

In both source and channel coding, it is hard to find coding schemes that are both efficient and yet approach the fundamental limits. Classical random coding achievability schemes in information theory include joint typicality encoding / decoding which has exponential complexity in general. Specifically, consider the source coding problem where we wish to compress a source sequence \( x^n \in X^n \) with a loss function \( d(\cdot, \cdot) \) and a target distortion \( D \). It is known that asymptotically, we have to transmit at least \( R(D) \) bits per symbol, where \( R(D) \) denotes the rate distortion function [1]. In a typical achievability scheme, we construct a codebook \( C \) containing \( n^R \) randomly generated codewords, and for a given source sequence \( x^n \), we search over the entire codebook for a codeword \( c \in C \) that satisfies \( d(x^n, c) \leq D \). The complexity of this scheme is exponential in the block length \( n \), and even storing all codewords is challenging. Because of these two drawbacks, it is obviously impractical to implement such achievability schemes.

A lot of work has been done towards reducing the complexity of rate-distortion codes, see [2] and references within. Recently, Gupta et al. [6] proposed a scheme that achieves the rate-distortion function with low complexity by dividing source words into sub-blocks in a divide-and-conquer approach. Korada et al. [12] showed that polar codes can achieve the rate distortion function with relatively low complexity. Barron and Joseph [3] introduced sparse regression codes (SPARCs) for the AWGN channel coding problem: a scheme that achieves the Shannon capacity with low complexity. More recently, Venkataramanan et al. applied the same idea to Gaussian source coding with quadratic distortion [19], [20]. The idea behind the work in [19] is based on an encoder with several levels, where each level refines the description of the source word from the previous levels (the idea is also related to the setting of successive refinement [5][10][11], as will be discussed below). The division of labor between levels allows a tradeoff between performance and complexity. For example, a scheme is presented in [19] that attains a probability of excess distortion that vanishes exponentially in \( \frac{n}{\log n} \), with complexity of \( O \left( \left( \frac{n}{\log n} \right)^2 \right) \).

In this paper, we aim to study the general phenomenon in which a layered structure of a code, as done in SPARCs [20], results in improved complexity. We denote this approach by layered coding, a family that includes all coding schemes that can be divided into a number of stages. In particular, layered coding consists of \( L \) stages and \( L \) codebooks \( C_1, \ldots, C_L \), one for each layer. The encoding process can be described as follows:

- Find \( c_1 \in C_1 \) that minimizes a function \( \psi_1(x^n, c_1) \).
- For \( i \geq 2 \), given \( c_1, \cdots, c_{i-1} \), find \( c_i \in C_i \) that minimizes \( \psi_i(x^n, c_1, c_2, \cdots, c_i) \),

where \( \psi_1, ..., \psi_L \) are functions that depend on the specific implementation of the scheme. In general, we would like each function \( \psi_i \) to be simple, in the sense that its complexity should not depend on the size of the codebook \( C_i \). In general, we would like the size of each of the sub-codebooks \( C_i \) to be small, so that the overall complexity will be small.

One can think of the layered coding architecture as searching for an appropriate codeword over a tree structure for the appropriate codeword: the larger the tree, the faster the codeword can be found. On the other hand, the tree structure restricts the class of coding schemes, resulting in a classical tradeoff between encoding complexity (how fast can we find the codeword) and performance (how much do we end up compressing). Note that the SPARC construction is a special case of layered coding scheme, one that attains good performance.

Another setting in which a layered architecture is of interest is the successive refinement framework [5]. Here, there is a requirement that after each layer of encoding, the source will be reconstructible, where the reconstruction quality varies with each layer. Note that the additional requirement of reproduction at the intermediate layers may restrict the achievable rates, and indeed, shown in [5], not all sources are successively refinable, i.e. not always the
optimal distortion (equal to the distortion-rate function) can be attained at each stage simultaneously. Other performance measures beyond the rate-distortion function are the excess distortion exponent [14] and the source dispersion (see Ingber and Kochman [8] and Kostina and Verdú [13]). The excess distortion exponents in successive refinements were studied in [9][17]. The dispersion, however, has not been studied for the setting of successive refinement.

The contributions of this paper are as follows:

- We demonstrate two layered coding schemes, one for the quadratic-Gaussian case and one for the binary-Hamming case, for which the gap to the rate-distortion function can be bounded as a function of the number of layers. This enables a direct tradeoff with the encoder complexity (measured by the overall size of the codebooks). For example, we show how to design an $O(n^\log n)$ algorithm that asymptotically achieves the rate distortion bound. If we allow a more complex algorithm, with complexity which is approximately exponential in $\sqrt{n}$, the schemes can achieve the source dispersion.

- The analysis of the layered schemes, discussed above, reveals that if the number of layers is fixed, then for each layer, the gap to the rate-distortion function is optimal up to the $1/\sqrt{n}$ term, i.e. the source dispersion is achieved at each stage, simultaneously. This can be thought of as a stronger notion of successive refinability, and this result means that the considered sources, i.e. the binary and the Gaussian sources, are "strongly successively refinable".

The rest of paper is organized as follows. In Section II, we revisit known results about source dispersion and successive refinement. We provide our main results in Section III where proof details are given in Section IV.

II. PRELIMINARIES

A. Layered Codes

Let $n$ be the block length of the coding scheme. The codebook consists of $L$ sub-codebooks $\{C_1^n, C_2^n, \ldots, C_L^n\}$ and each sub-codebook consists of $M_i$ length-$n$ codewords for $1 \leq i \leq L$. Denote $C_i^n = \{\hat{x}_i^n(m_i) : 1 \leq m_i \leq M_i\}$. The compressed representation of the source consists of a length $L$ vector $m = (m_1, \ldots, m_L)$ which indicates the index of codewords from each sub-codebook. The encoder finds the message in a successive refinement manner as follows:

- Find $c_1 \in C_1$ that minimizes some function $\psi_1(x^n, c_1)$.
- For $i \geq 2$, given $c_1, \ldots, c_{i-1}$, find $c_i \in C_i$ that minimizes $\psi_i(x^n, c_1, \ldots, c_i)$, where $\psi_1, \ldots, \psi_i$ are functions that depend on the specific implementation of the scheme. Note that the total number of codewords is $M_1 \times \cdots \times M_L$ and the rate of scheme is $R = \sum_{i=1}^L \frac{1}{n} \log M_i$. Once the decoder receives the message, it reconstructs $\hat{x}^n(m) = \phi(\hat{x}_1^n(m_1), \ldots, \hat{x}_L^n(m_L))$ for some decoding function $\phi$.

Definition 1: An $(n, L, \{M_1, \ldots, M_L\}, D, \epsilon)$-layered code is a length $n$ coding scheme with $L$ sub-codebooks where the size of the $i$-th sub-codebook is $M_i$, and the probability of excess distortion $P(d(X^n, \hat{X}^n) > D)$ is at most $\epsilon$.

B. Source Dispersion

We shall be interested in the optimal performance of such schemes in terms of the distance of the achievable rate from the optimal rate-distortion function. This is quantified by the dispersion. Consider an i.i.d. source $X^n$ with law $P$. In classical source coding, it is well known that the optimal asymptotic rate is the rate-distortion function $R(P, D)$. In other words, given target distortion $D$, we can find a family of schemes whose excess distortion probability vanishes while the rate is converging to $R(P, D)$. However, this can be achieved only when the block length $n$ goes to infinity. Under this setting, we can consider two asymptotic behaviors: One asymptotic analysis of source coding is the excess distortion exponent [14]. In this scenario, we fix the coding rate and investigate how fast the excess distortion probability $P(d(X^n, \hat{X}^n) > D)$ is decaying. Recently, Ingber et al. studied another asymptotic behavior of source coding which is called source dispersion [8] (see also [13]). In this setting, the figure of merit is the minimum number of codewords for which there exists a coding scheme with excess distortion probability $\epsilon$. It was shown that the difference between the minimum rate for fixed $n$ and $R(P, D)$ is inversely proportional to square root of $n$. More formally, given target excess distortion probability $\epsilon$ and block length $n$, let $R_{P,D,\epsilon}(n)$ be the minimum rate of coding scheme that has excess distortion probability smaller than $\epsilon$.

Theorem 1 ([8]): Consider a discrete memoryless source with law $P$ with distortion $d(\cdot, \cdot)$ and target distortion $D$. Let $\epsilon$ be a target excess distortion probability and block length $n$. Suppose the rate-distortion function $R(P, D)$ is twice differentiable with respect to $D$ and the elements of $P$ in some neighborhood of $(P, D)$. Then for large enough $n$,

$$ R_{P,D,\epsilon}(n) = R(P, D) + \sqrt{\frac{V(P, D)}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right) $$

where $V(P, D)$ is the source dispersion, given by

$$ V(P, D) \triangleq \text{VAR}[R'(P, D)] $$

$$ = \sum_{x \in \mathcal{X}} P(x)(R'(x))^2 - \left[\sum_{x \in \mathcal{X}} P(x)R'(x)\right]^2 $$

and $R'(x)$ is the derivative of the $R(P, D)$ with respect to the probability $P(x)$:

$$ R'(x) \triangleq \left[\frac{\partial R(Q, D)}{\partial Q(x)}\right]_{Q=P}. $$

C. Successive Refinement

Consider the case where an encoder wants to send a source $x^n$ to two receivers with different target distortions (and possibly different distortion measures) due to the different
capacities of network links. For example, one receiver wants to recover \( x^n \) more finely with a link of larger capacity, while the other wants to recover the source more coarsely because his link has lower capacity. Instead of designing separate coding schemes, one might want to design a code for the second receiver and send extra information to the first receiver on top of the message of second receiver. This setting is called successive refinement, independently proposed by Koshelév [10][11] and Equitz and Cover [5]. In general, a successive refinement problem with \( L \) decoders can be formulated as follows. The encoder sends the \( L \)-tuple of messages \((m_1, \ldots, m_L)\) where \( 1 \leq m_i \leq M_i \) for all \( 1 \leq i \leq L \), and the \( i \)-th decoder takes \((m_1, \ldots, m_i)\) and based on that reconstructs \( X^n(m_1, \ldots, m_i) \). Each decoder wants to recover the source \( x^n \) with distortion \( D_i \), i.e.,

\[
d_i(x^n, \hat{X}^n_i(m_1, \ldots, m_i)) \leq D_i.
\]

The rate tuple \((R_1, \ldots, R_L)\) is simply defined as

\[
R_i = \sum_{j=1}^{i} \frac{1}{n} \log M_j.
\]

An \((n, L; (R_1, \ldots, R_L), (D_1, \ldots, D_L), \epsilon)\)-successive refinement code is a coding scheme with block length \( n \) and excess distortion probability \( \epsilon \) with rates \((R_1, \ldots, R_L)\) and target distortions \((D_1, \ldots, D_L)\). An excess distortion probability is defined as \( P(d_i(X^n, \hat{X}^n_i) > D_i) \) for some \( i \).

Definition 2: If there is a family of successive refinement codes, \((n, L; (R_1^{(n)}, \ldots, R_L^{(n)}), (D_1^{(n)}, \ldots, D_L^{(n)}), \epsilon(n))\) where

\[
\lim_{n \to \infty} R_i^{(n)} = R_i, \quad \forall i
\]

\[
\lim_{n \to \infty} \epsilon(n) = 0,
\]

then, rate-distortion \((R_1, \ldots, R_L, D_1, \ldots, D_L)\) is said to be achievable.

Theorem 2 ([5]): Consider a discrete memoryless source \( X^n \) with law \( P \) and distortion measure \( d_i \) at each stage. The rate-distortion tuple \((R_1, \ldots, R_L, D_1, \ldots, D_L)\) is achievable if and only if there is a joint law \( P_{X, \hat{X}_1, \ldots, \hat{X}_L} \) of random variables \((X, \hat{X}_1, \ldots, \hat{X}_L)\) (where \( X \) is distributed according to \( P \)) such that

\[
I(X; \hat{X}_1, \ldots, \hat{X}_L) \leq R_i
\]

\[
\mathbb{E}[d_i(X, \hat{X}_i)] \leq D_i
\]

for all \( i \).

The condition in the theorem holds for the cases of a Gaussian source with quadratic distortion and for a binary source with Hamming distortion. For the quadratic-Gaussian case, \((R(\sigma^2, D_1), \ldots, R(\sigma^2, D_L), D_1, \ldots, D_L)\) is achievable where \( R(\sigma^2, D) \) is a rate-distortion function of Gaussian source with variance \( \sigma^2 \). Note that the rates are optimal at each stage as if they were designed for each of the decoders individually. Similarly, for a binary source Bern\((p)\) with Hamming distortion, \((R(p, D_1), \ldots, R(p, D_L), D_1, \ldots, D_L)\) is achievable where \( R(p, D) \) is the rate-distortion function of the Bern\((p)\) source.

We note that this property, called ‘successive refinability’, is not shared by all sources and distortion measures. In this paper we show that the binary-Hamming and quadratic-Gaussian cases, the sources are not only refinable in the rate sense, but also in the dispersion sense.

### III. Main Results

#### A. Layered Coding Schemes

Theorem 3: For an i.i.d. Gaussian source \( N(0, \sigma^2) \), there exists a \((n, L; \{M_1, \ldots, M_L\}, D, \epsilon)\)-layered coding scheme where

\[
\frac{1}{n} \log M_1 \leq \frac{1}{2} \log \frac{\sigma^2}{D_1} + \sqrt{\frac{V(\sigma^2, D)}{n} Q^{-1}(\epsilon)} + O \left( \frac{\log n}{n} \right)
\]

\[
\frac{1}{n} \log M_i \leq \frac{1}{2} \log \frac{D_{i-1}}{D_i} + 3 \frac{\log n}{n}, \quad 2 \leq i \leq L
\]

for any \( \sigma^2 > D_1 > D_2 > \cdots > D_L = D \), where \( V(\sigma^2, D) = \frac{1}{2} D \sigma^4 \) is the dispersion of the Gaussian source.

Note that the \( O \left( \frac{\log n}{n} \right) \) term is not affected by specific choice of \( D_1 \) or \( L \). This is an important fact since we will see some cases where \( D_1 \) and \( L \) can vary with \( n \).

Corollary 4: The overall rate for the layered coding scheme is bounded by:

\[
\sum_{i=1}^{L} \frac{1}{n} \log M_i \leq \frac{1}{2} \log \frac{\sigma^2}{D} + \sqrt{\frac{V(\sigma^2, D)}{n} Q^{-1}(\epsilon)} + 3(L-1) \log \frac{n}{n} + O \left( \frac{\log n}{n} \right)
\]

Theorem 5: For an i.i.d. Bern\((p)\) source, there exists an \((n, L; \{M_1, \ldots, M_L\}, D, \epsilon)\)-layered coding scheme where

\[
\frac{1}{n} \log M_1 \leq h_2(p) - h_2(D_1) + \sqrt{\frac{V(p, D)}{n} Q^{-1}(\epsilon)} + O \left( \frac{\log n}{n} \right)
\]

\[
\frac{1}{n} \log M_i \leq h_2(D_{i-1}) - h_2(D_i) + k \frac{\log n}{n}, \quad 2 \leq i \leq L
\]

for any \( p > D_1 > D_2 > \cdots > D_L = D \), where \( V(p, D) = p(1-p) \log^2 \frac{1-p}{p} \) is the dispersion of the Bern\((p)\) source and \( h_2(p) = -p \log p - (1-p) \log (1-p) \) is the binary entropy function. Note that \( k \) is a universal constant that does not depend on any other variables. Here too, the \( O \left( \frac{\log n}{n} \right) \) term does not depend on \( L \) or \( D_1 \).

Corollary 6: The overall rate for the layered coding scheme is bounded by:

\[
\sum_{i=1}^{L} \frac{1}{n} \log M_i \leq h_2(p) - h_2(D) + \sqrt{\frac{V(p, D)}{n} Q^{-1}(\epsilon)} + k(L-1) \frac{\log n}{n} + O \left( \frac{\log n}{n} \right).
\]

1533
B. Discussion: Rate-Distortion Trade-Off

In both corollaries, it is obvious that the choice of \( L \) has an important role. If we choose large \( L \), then it will give a larger rate because of the penalty term \( L \frac{\log n}{n} \). On the other hand, the complexity of the scheme behaves in the opposite way. For simplicity, consider the case that \( M_1 = M_2 = \cdots = M_L = M \). It is easy to find \( M \) and \( D_1 > D_2 > \cdots > D_L \) which satisfy (1)-(4) with equality. For example, in the Gaussian case, we can find \( M \) and \( D_1, \cdots, D_L \) sequentially:

\[
\frac{1}{n} \log M = \frac{1}{2} \log \frac{\sigma^2}{D_1} + \sqrt{\frac{V(\sigma^2, D)}{n}} Q^{-1}(\epsilon) + O \left( \frac{\log n}{n} \right) \\
\frac{1}{n} \log M = \frac{1}{2} \log \frac{D_{i-1}}{D_i} + 3 \frac{\log n}{n}, \quad \text{for } 2 \leq i \leq L.
\]

Clearly, the number of reconstructions is \( R = \frac{1}{n} \log M \) and the rate is \( R = \frac{1}{2} L \log M \). Note that complexity is \( M \log \) since we are searching over \( M \) codewords at each stage. For fixed rate \( R \), larger \( L \) gives smaller complexity.

Since Theorems 3 and 5 hold for any \( n \) and \( L \), we can set \( L \) to be a function of \( n \). Consider the following two examples which are valid for both the Gaussian and binary cases. In the following examples, denote \( R(P, D) \) and \( V(P, D) \) be rate-distortion function and the source dispersion.

- If \( L = \frac{n}{\log n} + 1 \), then the achieved rate is
  \[
  R = R(P, D) + O \left( \frac{1}{\log n} \right),
  \]
  i.e. the scheme achieves the rate distortion function as \( n \) increases, while the coding complexity is of order \( n \log n \). Note that this complexity is near polynomial.

- If \( L = \frac{\sqrt{n}}{\log n} + 1 \), then the achieved rate is
  \[
  R = R(P, D) + \sqrt{\frac{V(\sigma^2, D)}{n}} Q^{-1}(\epsilon) + O \left( \frac{1}{\sqrt{n \log n}} \right).
  \]
  Note that \( R - R(P, D) \) is inversely proportional to \( \sqrt{n} \) with coefficient \( \sqrt{V(\sigma^2, D)} Q^{-1}(\epsilon) \), in other words, layered coding can achieve the source dispersion. On the other hand, coding complexity is order of \( \frac{1}{\sqrt{n \log n}} \) which is better than the original exponential complexity.

The following statements show an essence of the trade-off between complexity and performance.

- In the rate point of view,
  - If \( L \) is constant, then the scheme achieves the rate distortion and the dispersion as well, but the complexity is exponential (albeit with a smaller exponent).
  - If \( L \frac{\log n}{n} \to 0 \) as \( n \to \infty \), we can achieve the rate distortion function.
  - If \( L \frac{\log n}{n} \to 0 \) as \( n \to \infty \), we can achieve the source dispersion.

- For fixed \( R \), the coding complexity (or size of codebooks) scales like
  \[
  L \exp \left( \frac{nR}{L} \right)
  \]
  which is a decreasing function of \( L \) (note that we can not choose \( L \) to be linear in \( n \), since \( L \log n \) has to be vanishing).

C. Link to Successive Refinement

An important property of the layered schemes from Theorems 3 and 5 is the following: the quantities \( D_1, \cdots, D_L \) are not merely design parameters that indirectly determine the rate allocation. At each stage \( i \), the distortion attained is \( D_i \), and therefore those layered coding schemes are applicable for the successive refinement setting. Specifically, let the number of stages \( L \) be fixed (and not growing with \( n \) as before). With \( L \) being constant, the above theorems imply that we can achieve the rates

\[
R_i = \frac{1}{n} \log M_1 \cdots M_i
\]

\[
\leq R(P, D_i) + \sqrt{\frac{V(P, D_i)}{n}} Q^{-1}(\epsilon) + O \left( \frac{\log n}{n} \right), \quad \forall i
\]

at every stage \( i = 1, \ldots, L \) simultaneously, for both the Gaussian and binary cases. Note that this is the minimum rate at each stage as if it were a point to point source coding problem with target distortion \( D_i \). Therefore, we can achieve the minimum number of codewords at each stages. Moreover, if the first stage in the encoding is without an error, the remaining stages are guaranteed to be error-free.

We would like to point out that achieving dispersion in the above theorems is highly based on the fact that for the binary and Gaussian sources, the dispersion \( V(P, D) = \text{VAR}[R(P, D)] \) is independent of the choice of \( D \), i.e., \( V(P, D) = V(P, D_1) \). If \( V(P, D) \) is strictly decreasing function (with \( D \)), then \( V(P, D_1) \) will be larger than \( V(P, D) \) then layered coding might not able to achieve the source dispersion at every stage. This is a topic of further investigation.

IV. PROOF

A. Gaussian Source and quadratic distortion

In [18], maximum correlation was used for encoding and order statistics analysis was used for error analysis. Here, instead, we will set \( \sigma^2 > D_1 > \cdots > D_L = D \) to be target distortions at stage \( i \), and find a codeword \( c_{m_i} \) from sub-codebook \( C_i \) which satisfies \( ||c_m + \cdots + c_{m_i} - x^n||^2 \leq D_i \). The basic idea of the achievability scheme is the following,

**Algorithm 1 Coding Scheme.**

Set \( \sigma^2 > D_1 > D_2 > \cdots > D_L = D \), and \( A(0) = X^n \).

for \( i = 1 \) to \( L \) do

Find the codeword \( c_{m_i} \in C_i \) s.t. \( ||A(i-1) - c_{m_i}||^2 \leq D_i \).

If there is no such codeword, declare an error.

Let \( A(i) = A(i-1) - c_{m_i} \).

end for
1) Codebook Size: Let \( C_1, C_2, \ldots, C_L \) be the codebooks at each stage. We will specify the size of each codebook \( M_i = |C_i| \) later on. Note that the total number of codewords is \( M_1 M_2 \cdots M_L \). Let \( \epsilon > 0 \) be a target excess distortion probability. By [7] and [13], in the first stage, we can find a codebook such that the excess distortion probability is smaller than \( \epsilon \) with, where the codebook size satisfies
\[
\frac{1}{n} \log M_1 \leq R(\sigma^2, D_1) + \frac{\sqrt{V(\sigma^2, D_1)}}{n} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right)
\]
where \( R(\sigma^2, D) = \frac{1}{2} \log \frac{2 \pi e D}{\sigma^2} \) is the quadratic-Gaussian rate distortion function. Recall that \( V(\sigma^2, D) = \frac{1}{2} \) is a Gaussian source dispersion (and is independent of the source variance and the distortion).

If the encoding at the first stage was successful, then we know for sure that the source word resides in a \( D_1 \)-ball around the reconstruction word. Therefore, in stage 2, we can design the codebook to cover the \( D_1 \) uncertainty ball with smaller balls, of radius \( D_2 \). The rest of the stages continue similarly. The size of the codebooks at stages 2, \ldots, \( L \) is determined by the number of balls required to cover a larger ball. Specifically, Rogers [16] and Verger-Gaury [21] showed that an \( n \) dimensional ball with radius \( r \) can be covered by \( N(r) \) unit balls, where
\[
\frac{1}{n} \log N(r) \leq \log r + 3 \frac{\log n}{n}.
\]
This result implies that for \( 2 \leq i \leq L \), we can find a codebook \( C_i \) where
\[
|C_i| \leq N\left(\sqrt{\frac{D_{i-1}}{D_i}}\right)
\]
and \( C_i \) covers the larger ball of radius \( D_{i-1} \) with smaller balls of radius \( D_i \). Therefore,
\[
\frac{1}{n} \log M_i \leq N\left(\sqrt{\frac{D_{i-1}}{D_i}}\right) \leq \frac{1}{2} \log \frac{D_{i-1}}{D_i} + 3 \frac{\log n}{n}.
\]

2) Excess Distortion Probability: Note that if stage 1 is successful, the excess distortion probabilities at the rest of stages are zero. This is because at time step \( i > 1 \), \( C_i \) completely covers all points in the \( D_{i-1} \)-ball (by construction).

B. Binary source and Hamming distortion

Before going into the proof, let us introduce the useful lemma, which is a refinement of the covering lemma [4] (also refer to [4] for the definitions of a type \( P \) and of a type class \( \mathcal{T}_P \)).

**Lemma 7 (Refined Covering Lemma):** Let \( \mathcal{X} \) and \( \hat{\mathcal{X}} \) be discrete finite alphabets and let \( d_M = \max_{x, \hat{x}} d(x, \hat{x}) \). Given a type \( P \) on \( \mathcal{X} \) such that \( P(x) > \frac{1}{n} \) for all \( x \in \mathcal{X} \) and \( D \in (0, d_M) \), there exists a subset \( B_{P,D} \subseteq \hat{\mathcal{X}}^n \) such that for any \( x^n \in \mathcal{T}_P \), \( d(x^n, B_{P,D}) \leq D \) and
\[
\frac{1}{n} \log |B_{P,D}| \leq R(P, D) + k_1 \frac{\log n}{n}
\]
where
\[
k_1 = |\mathcal{X}| \cdot |\hat{\mathcal{X}}| d_M \frac{\log |\mathcal{X}|}{D} + 4 |\mathcal{X}| \cdot |\hat{\mathcal{X}}| + 8
\]
The proof of lemma is omitted and will be given in the extended version of paper.

**Remark 1:** Note that the lemma is non-asymptotic, i.e. the proof did not require large enough \( n \). This lemma holds generally for any choice of \( D \), \( n \) and type \( P \).

The coding scheme here is similar in spirit to the one used for the Gaussian case:

**Algorithm 2 Coding Scheme.**

Set \( p > D_1 > D_2 > \cdots > D_L = D \), and \( A(0) = X^n \).

for \( i = 1 \) to \( L \) do

Find a codeword \( c_{m_i} \in C_i \) s.t. \( d(A(i-1), c_{m_i}) \leq D_i \).

if there is no such codeword, declare an error.

Let \( A(i) = A(i-1) \oplus c_{m_i} \).

end for

1) Codebook Size: Let \( C_1, C_2, \ldots, C_L \) be the codebooks at each stage. We will specify the size of each codebook \( M_i = |C_i| \) later on. Let \( \epsilon > 0 \) be a target excess distortion probability. By [8, Theorem 1], in the first stage, we can find a codebook with excess distortion probability at most \( \epsilon \), with
\[
\frac{1}{n} \log M_1 \leq R(p, D_1) + \frac{\sqrt{V(p, D_1)}}{n} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right)
\]
where \( R(p, D) = h_2(p) - h_2(D) \) and \( V(p, D) = p(1-p) \log^2 \frac{1-p}{p} \).

By Lemma 7, for given type \( Q \) and \( D \), there is a codebook \( C(Q, D) \) such that \( D \)-covers all type \( Q \) sequences with
\[
\frac{1}{n} \log |C(Q, D)| \leq R(Q, D) + k_1 \frac{\log n}{n}
\]
where \( k_1 = |\mathcal{X}| \cdot |\hat{\mathcal{X}}| d_M \frac{\log |\mathcal{X}|}{D} + 4 |\mathcal{X}| \cdot |\hat{\mathcal{X}}| + 8 \).

Similarly to the Gaussian case, we will cover the hamming ball with radius \( D_{i-1} \) using hamming balls with radius \( D_i \). For \( 2 \leq i \leq L \), the codebook \( C_i \) is an union of \( C(Q, D_i) \)'s for all type \( Q \in \mathcal{T}(D_{i-1}, D_i) \) \( \triangleq \{ Q \in \mathcal{P}_n : D_i < Q(1) \leq D_{i-1} \} \) and the zero codeword \((0, 0, 0, \cdots, 0) \). Note that \( \mathcal{P}_n \) is the set of all types of length \( n \) binary sequences. Then,
\[
|C_i| \leq 1 + \sum_{Q \in \mathcal{T}(D_{i-1}, D_i)} |C(Q, D_i)|
\]
\[
\leq 1 + \sum_{Q \in \mathcal{T}(D_{i-1}, D_i)} |C(Q, D_i)|
\]
\[
\leq 1 + (nD_{i-1} - nD_i + 1) \max_{Q \in \mathcal{T}(D_{i-1}, D_i)} |C(Q, D_i)|
\]
\[
\leq (nD_{i-1} - nD_i + 2) \max_{Q \in \mathcal{T}(D_{i-1}, D_i)} |C(Q, D_i)|
\]
(5) is because \( |\mathcal{T}(D_{i-1}, D_i)| \leq nD_{i-1} - nD_i + 1 \). Also, we are only covering the types \( Q \) such that \( Q(x) > \frac{1}{n} \) for all \( x = 0, 1 \) since \( D_1 > \cdots > D_L = D > \frac{1}{n} \). Therefore,
\[
\frac{1}{n} \log M_i \leq \max_{Q \in \mathcal{T}(D_{i-1}, D_i)} R(Q, D_i)
\]
\[
\begin{align*}
&+ k_1 \log \frac{n}{n} + \frac{1}{n} \log (n D_{i-1} - n D_i + 2) \\
\leq & h_2(D_{i-1}) - h_2(D_i) \\
&+ k_1 \log \frac{n}{n} + \frac{1}{n} \log (n D_{i-1} - n D_i + 2) \\
\leq & R(D_{i-1}, D_i) + k \log \frac{n}{n} + 1
\end{align*}
\]

where \( k = k_1 + 1 \).

2) Excess Distortion Probability: Note that if stage 1 is successful, excess distortion probability in the rest of stages will be zero. This is because \( i \)-th sub-codebook \( C_i \) covers all sequences of type \( Q \in T(D_{i-1}, D_i) \) (by construction). Moreover, all other sequences of type \( Q \) with \( Q(1) \leq D_i \) are covered by zero codeword \((0, 0, \ldots, 0)\).

REFERENCES