A Study of Source Coding Algorithms with Fidelity Criterion

（忠実度規範つき情報源符号化アルゴリズムに関する研究）

古　賀　弘　樹

A Study of Source Coding Algorithms with Fidelity Criterion

（忠実度規範つき
情報源符号化アルゴリズムに関する研究）

古賀 弘樹

Abstract

In this thesis data compression algorithms with fidelity criterion are proposed for a certain class of information sources. Properties of the data compression schemes are analyzed from a viewpoint of the rate-distortion theory. In the framework of the rate-distortion theory, the rate-distortion function represents an achievable lower bound of compression efficiency for a class of stationary and ergodic sources. Though the source coding theorem guarantees the existence of an asymptotically optimal block code that achieves the rate-distortion bound with increasing the blocklength, little is known about construction of the asymptotically optimal encoding scheme. The algorithms proposed in this thesis are intended to realize the asymptotic optimal encoding scheme in a universal way.

Chapter 1 is devoted to a general introduction for data compression with fidelity criterion. Three problems considered in the following chapters are briefly formulated so that readers can grasp outline of this thesis.

In Chapter 2 a block coding algorithm for memoryless gaussian sources is proposed under the squared-error criterion. Source blocks are treated as elements in the n-dimensional Euclidean space. First, the algorithm is defined for the sources of known mean and variance. It consists of a scalar quantizer for encoding of block gain and a set of points on the n-dimensional unit hypersphere for encoding block shape. It is proved that there exists a set on the n-dimensional unit hypersphere that yields the asymptotic optimality of the algorithm. The most dominant term in the asymptotic performance only depends on parameters of the scalar quantizer. As a byproduct, the rate-distortion function is naturally deduced as an achievable bound of data compression efficiency. The algorithm is easily extended to a universal encoding scheme for memoryless gaussian sources of unknown mean or variance. In the algorithm, before encoding a source block the encoder estimates the unknown parameters by using the maximum-likelihood method and transmits the estimated ones in a quantized form to the decoder. Though the method results in a slight loss of rate, it leads to the same asymptotic rate-distortion property even if the parameters are unknown. This property shows a robustness includeed in the encoding scheme.

A data compression algorithm based on string matching is treated in Chapter 3.

In the algorithm both an encoder and a decoder have a buffer whose contents are independently generated according to a probability distribution on a source alphabet. Given a source block of blocklength n, the encoder searches for the minimum index of the buffer with property that distortion between the source block and n consecutive symbols beginning from the index is within an acceptable level. It transmits the index in the fixed-length binary form to the decoder. The decoder reproduces the source block as n consecutive symbols from the transmitted index. Rate required for transmission of indices is analyzed in detail for discrete memoryless sources with finite alphabet and memoryless gaussian sources. Under the assumption that probability distribution of such sources are known, a sufficient condition that the rate converges *in probability* to the rate-distortion function are deduced. In the case of discrete memoryless sources, for this convergence contents of the buffer must be generated according to another probability distribution that is originated from the test channel matrix. In the case of gaussian sources, however, any memoryless gaussian sequence of known mean and variance guarantees the convergence by modifying the way of matching.

In Chapter 4 two universal estimation algorithms of the probability distribution that make the data compression scheme proposed in Chapter 3 asymptotically optimal in the case of discrete memoryless sources without any knowledge on the source. The two algorithms use output sequences of the source and an auxiliary memoryless source putting out all symbols in the same alphabet. When contents of the buffer are drawn according to a probability distribution different from the optimal one, the divergence from the distribution to the optimal one means a cost in rate. Criterion of estimation imposed on one of the algorithms is similar to the PAC learning models in the computational learning theory. For arbitrarily fixed $\varepsilon > 0$ and $\delta \in (0,1)$ the algorithm outputs an estimate of the optimal probability distribution with probability at least $1-\delta$ that satisfies the divergence being not greater than ε. Lower bounds on length of the sequences required by the algorithm are obtained as a function of ε and δ. The other algorithm outputs a probability distribution arbitrarily close to the optimal one with high probability if appropriate parameters are chosen in the algorithm. Though an infinite auxiliary sequence is required by the algorithm, the algorithm turns out to stop *with probability one.*

Acknowledgements

The author would like to express his sincere gratitude of constant encouragements and suggestions to Professor Suguru Arimoto, who has been his supervisor for five years in the master course and the doctor course in University of Tokyo.

He wishes to thank to Professor Hirosuke Yamamoto for giving him many helpful suggestions and comments on this study. He also thanks to Professor Shun-ichi Amari, Professor Yasunori Okabe and Professor Koichiro Deguchi for reading the entire text in its original form and giving him valuable comments.

He acknowledges introduction of data compression with fidelity criterion with Dr. Hisashi Suzuki in Chuo University. Dr. Takashi Tsubouchi in Tsukuba University and Dr. Tomohide Naniwa gave him helpful comments on this study from the viewpoint of computer science and robotics. He would like to express his appreciation to other members of Arimoto Laboratory for their help in preparation of this thesis.

Finally, the author expresses his sincere thank to his parents and sister who encouraged and supported him on his study.

Hiroki Koga
Febrary, 1995
Tokyo, Japan

Contents

1 **Introduction** 1

1.1 What Is Data Compression with Fidelity Criterion? 1

1.2 Outline of the Thesis 6

2 **Asymptotic Optimality of Modified Spherical Code with Scalar Quantizer of Gain for Memoryless Gaussian Sources** 9

2.1 Introduction 9

2.2 Encoding Scheme 11

2.3 Analysis of Asymptotic Behavior 15

2.3.1 Conditions for the Scalar Quantizer 16

2.3.2 Basic Lemmata 17

2.3.3 Asymptotic Optimality of the Encoding Scheme 21

2.4 A Universal Data Compression Algorithm 26

2.4.1 Modified Encoding Algorithm for the Sources with Known Parameters 27

2.4.2 Universal Encoding Algorithm for the Sources with Unknown Parameters 32

2.5 Encoding of Other Continuous Sources 42

2.5.1 Encoding of Memoryless Laplacian Sources 43

2.5.2 Geometrical Interpretation of the Shannon Lower Bound 44

2.6 Conclusion 46

3 **Asymptotic Properties of Data Compression Algorithms with Fidelity Criterion Based on String Matching** 49

3.1 Introduction 49
3.2 Data Compression for Discrete Memoryless Sources 52
3.2.1 Definitions 52
3.2.2 Analysis 54
3.2.3 Encoding Scheme 62
3.3 Data Compression for Memoryless Gaussian Sources 64
3.3.1 Definitions and Analysis 64
3.3.2 Simulation Results 67
3.4 Conclusion 68

4 Universal Estimation of the Optimal probability distribution for compression of Discrete Memoryless Source with Fidelity Criterion 71
4.1 Introduction 71
4.2 A Universal Estimation Algorithm 74
4.2.1 Definition of the Algorithm 75
4.2.2 Comparison with the PAC Learning Models 78
4.2.3 Analysis of the Algorithm 79
4.3 Another Universal Estimation Algorithm 90
4.3.1 Definition and Properties of the Algorithm 90
4.3.2 Analysis of the Algorithm 93
4.4 Conclusion 97

5 Conclusion 99

A For Section 2.3 102

B For Section 2.4 104

C For Section 2.5 107

Bibliography 115

Chapter 1

Introduction

1.1 What Is Data Compression with Fidelity Criterion?

Further development of data compression schemes for efficient transmission or storage of information is required as communication networks spread out and become mature. There are a lot of data compression tools available on computers. For example, large and rarely used text files may be compressed for saving hard disks and can be quickly reproduced in the same form if necessary. Such kind of data compression is called *lossless data compression* since no information is lost through compression and reproduction.

On the other hand, in compression of image or speech signals distortion between an original signal and its reproduced form can be permitted as far as the distortion is acceptable. Data compression that permits an acceptable distortion is called *data compression with fidelity criterion* or *lossy data compression*. It usually leads to high compression efficiency compared with the lossless case, though it becomes impossible to reproduce the original signal itself. This thesis is concerned with data compression schemes with fidelity criterion.

The block diagram of a data transmission system assumed throughout the thesis is given in Fig. 1.1. The source generates a sequence of symbols according to a stochastic rule. If all of the symbols belong to a finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ and are

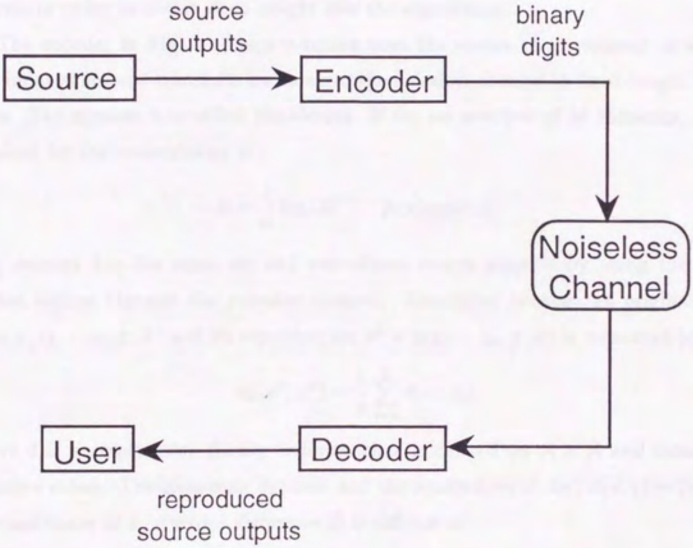


Figure 1.1 Block diagram of a transmission system

independently drawn according to an identical distribution, the source is called *discrete memoryless source.* If each symbol in the sequence is independently drawn according to a normal distribution, the source is called *memoryless gaussian source.* In case of the memoryless gaussian sources, alphabet $\mathcal{A}$ is equal to the set of all real numbers. These two classes of source are subclasses of *stationary ergodic sources* that generate a stationary and ergodic sequences. Performance of source coding algorithms proposed in this thesis is mainly analyzed for discrete memoryless sources and memoryless gaussian sources in order to obtain deep insight into the algorithms.

The encoder in Fig. 1.1 maps n-tuples from the source to an element of a finite and indexed set and transmits its index to the noiseless channel in fixed-length binary form. The number n is called *blocklength.* If the set consists of M elements, *rate* R required for the transmission is

$$R = \frac{1}{n}\log_2 M \qquad \text{[bit/symbol]}. \tag{1.1}$$

The decoder has the same set and reproduces source outputs by using the transmitted indices through the noiseless channel. Distortion between an original block $x^n = x_1x_2\cdots x_n \in \mathcal{A}^n$ and its reproduction $y^n = y_1y_2\cdots y_n \in \mathcal{A}^n$ is measured by

$$d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i), \tag{1.2}$$

where d is a *single-letter fidelity criterion* that is defined on $\mathcal{A} \times \mathcal{A}$ and takes non-negative values. The Hamming distance and the squared-error, i.e., $d(x, y) = (x - y)^2$ are candidates of d. *Average distortion* $\bar{D}$ is defined as

$$\bar{D} = E[d_n(X^n, Y^n)], \tag{1.3}$$

where E denotes expectation with respect to a random vector X^n from the source and Y^n means that a reproduction of X^n. Performance of data compression schemes with fidelity criterion is characterized by a pair $(R, \bar{D})$ defined by (1.1) and (1.3). The rate R and the average distortion $\bar{D}$ has a trade-off relationship, that is, a small average distortion is realized by a high rate while a low rate results in a large average distortion.

For a class of the stationary ergodic sources the rate-distortion function is a basic bound of compression efficiency for data compression schemes with fidelity criterion.

It describes an achievable bound of the schemes as a function of acceptable distortion levels and plays the same role as the entropy rate plays in lossless data compression. In the case of discrete memoryless source of alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ and probability distribution $\boldsymbol{p} = (p(a_1), p(a_2), \ldots, p(a_J))$, for any $D \geq 0$ it is defined as follows:

$$R(\boldsymbol{p}, D) = \min_{W \in \mathcal{W}(\boldsymbol{p}, D)} I(\boldsymbol{p}; W) \tag{1.4}$$

where $I(\boldsymbol{p}; W)$ is the mutual information defined by

$$I(\boldsymbol{p}; W) = \sum_{k=1}^{J} \sum_{j=1}^{J} p(a_j) W(a_k|a_j) \log_2 \frac{W(a_k|a_j)}{\sum_{l=1}^{J} p(a_l) W(a_j|a_l)} \tag{1.5}$$

for any $J \times J$ stochastic matrix W,

$$\mathcal{W}(\boldsymbol{p}, D) = \Big\{ W \text{ is a } J \times J \text{ stochastic matrix} \,|\, \sum_{j=1}^{J} \sum_{k=1}^{J} p(a_j) W(a_k|a_j) d(a_j, a_k) \leq D \Big\}, \tag{1.6}$$

and d is a single-letter fidelity criterion. Definition of the rate-distortion function for memoryless gaussian sources is parallel to (1.4) though summation included are replaced by integral. The rate-distortion function is convex and monotone decreasing function. For an arbitrary distortion level Δ the source coding theorem claims the following:

Theorem 1.1 (The Source Coding Theorem: converse part) *For any distortion level $\Delta > 0$ and $n > 0$ there is no encoding scheme satisfying both*

$$R < R(p, \Delta) \tag{1.7}$$

and

$$\bar{D} < \Delta. \tag{1.8}$$

Theorem 1.2 (The Source Coding Theorem: direct part) *For any distortion level $\Delta > 0$, $\varepsilon \geq 0$ and $\delta \geq 0$ satisfying $\delta + \varepsilon > 0$ there exists an encoding scheme that satisfies*

$$R \leq R(\boldsymbol{p}, \Delta) + \delta \tag{1.9}$$

and

$$\bar{D} < \Delta + \varepsilon \tag{1.10}$$

if n is sufficiently large.

Though Theorem 1.1 and Theorem 1.2 are stated in a terminology for discrete memoryless sources, these claims make sense for a class of stationary ergodic sources by replacing $R(\boldsymbol{p}, \Delta)$ with the rate distortion function for the sources. The two theorems show that the rate-distortion function is an essential bound for data compression with fidelity criterion. Existence of an *asymptotically optimal code* in a sense that the rate-distortion function is achievable is guaranteed by the direct part. Since proofs of the source coding theorem can be found in textbooks on the information theory, it is not given here. (The proof for stationary ergodic sources can be found in [1, 2]. Only memoryless sources are treated in [1, 2, 3, 4, 5].) Proving the converse part is easier than establishing the direct part. In proofs of the direct part a randomly generated code is always used for establishing achievablity of the rate-distortion function.

However, constructing an encoding scheme with the asymptotic optimality that requires reasonable computational costs for execution is one of open problems in the Shannon theory originating from [6]. For instance, time and space complexity of the random code used in the proofs is of exponential order of n. The construction remains unsolved even for simple memoryless sources while many lossless data compression methods that asymptotically achieve the entropy rate have been developed for the class of stationary ergodic sources. Difficulty of finding the asymptotically optimal schemes consists in both a probabilistic structure of the sources and a geometric aspect included in distortion measures. Main objective of the thesis is throwing light on such difficulty by considering encoding algorithms for simple sources and obtaining deep insight into not only the asymptotic optimal encoding scheme but also improvement of practical data compression schemes such as image and speech coding.

1.2 Outline of the Thesis

This section is devoted to a brief formulation of problems treated in this thesis so that readers can acquire a bird's-eye view on the thesis. The following three topics on data compression with fidelity criterion are discussed:

- an asymptotically optimal encoding scheme for memoryless gaussian sources,
- conditions for the asymptotic optimality of encoding schemes based on string matching,
- universal estimation algorithms of the optimal probability distribution for compression of discrete memoryless sources.

The first topic is treated in Chapter 2. It is motivated by a simple question that "what kind of data compression algorithm that asymptotically achieves the rate-distortion function for memoryless gaussian sources under the squared-error criterion?" It is well-known that the rate-distortion function for memoryless gaussian sources is an upper-bound of the rate-distortion functions for other continuous-alphabet memoryless sources of the same variance. Therefore, finding an asymptotically optimal data compression algorithm for the memoryless gaussian sources means designing the best scheme for the least compressible sources subject to a power constraint. As is proved in Chapter 2, for memoryless gaussian sources of known mean and variance a data compression algorithm essentially equal to the shape-gain vector quantization turns out to be asymptotically optimal under a certain condition. Another interesting aspect of the algorithm is unveiled when it is extended to a universal data compression method for the sources of unknown mean or variance. A robustness included in the algorithm and application of the algorithm to compression of other continuous memoryless sources are also discussed.

Chapter 3 is devoted to the second topic. The Lempel-Ziv algorithm proposed in 1977 [7] is a famous universal lossless data compression scheme that asymptotically achieves the entropy rate for a class of stationary ergodic sources without any knowledge on the sources. In the chapter extension of the Lempel-Ziv algorithm to a data compression scheme with fidelity criterion is discussed. A sufficient condition for the scheme being asymptotically optimal is deduced for discrete memoryless sources and memoryless gaussian sources, which clarifies an essential difference between the asymptotically optimal encoding schemes in the two cases. That is, if another output sequence from the sources is available to both an encoder and a decoder, the asymptotically optimal encoding is realized in the case of gaussian sources, while it is quite difficult (probably impossible) in the case of discrete sources.

Chapter 4 deals with the last topic. For a class of discrete memoryless sources conventional proofs of the source coding theorem include a random coding technique generating a random code according to the probability distribution

$$p^*(a_k) = \sum_{j=1}^{J} p(a_j) W^*(a_k|a_j) \quad \text{for all} \quad k = 1, 2, \ldots, J \tag{1.11}$$

in order to establish the existence of the asymptotically optimal code, where $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ denotes alphabet of the sources, $\boldsymbol{p} = (p(a_1), p(a_2), \ldots, p(a_J))$ is probability distribution on $\mathcal{A}$, W^* denotes the $J \times J$ stochastic matrix achieving the minimum in (1.4) at $D = \Delta$ and Δ is an acceptable distortion level arbitrarily fixed. Hence, knowledge on $\boldsymbol{p}^*$ leads to obtaining an intuition on encoding schemes with the asymptotic optimality. In the chapter two algorithm that universally estimates $\boldsymbol{p}^*$ for an arbitrary Δ are proposed. One of the algorithms is related to the encoding scheme proposed in Chapter 3. For any $\varepsilon > 0$ and $\delta \in (0, 1)$ it outputs an estimate of $\boldsymbol{p}^*$, denoted by $\hat{\boldsymbol{p}^*}$, satisfying $D(\boldsymbol{p}^*||\hat{\boldsymbol{p}^*}) \leq \varepsilon$ with probability at least $1 - \delta$, where $D(\cdot||\cdot)$ denotes the divergence. Since the divergence means a cost of the data compression scheme, the estimate asymptotically enables to transmit source outputs in rate $R(\boldsymbol{p}, \Delta) + \varepsilon$ and average distortion close to Δ per source symbol. The algorithm provides another view as a first attempt to apply the PAC learning models to data compression with fidelity criterion in an feasible way. The other universal estimation algorithm can estimate $\boldsymbol{p}^*$ of arbitrary precision with high probability if appropriate parameters are chosen in the algorithm. Both of the two algorithms are analyzed in detail in Chapter 4.

Readers can start to read these three chapters in any order since they are written as if they were independent topics. Notations are introduced in each chapter. Detailed description on background of each topics is also given at the beginning of each chapter.

Relations between the topics are stated in corresponding parts in case that cross-reference between chapters is desirable.

Chapter 2

Asymptotic Optimality of Modified Spherical Code with Scalar Quantizer of Gain for Memoryless Gaussian Sources

2.1 Introduction

Data compression for discrete-time analogue signals is important for the sake of practical applications. In many applications such as image coding and speech coding, it is desirable to encode signals efficiently within an acceptable distortion. The framework of rate-distortion theory, which originated from Shannon [8] and described in detail by Berger [2], provides a basis to develop efficient encoding methods. According to the rate-distortion theory, lower bound of compression efficiency becomes a function of distortion level called a rate-distortion function if sources are stationary and ergodic. The rate-distortion function is usually defined as the minimum of the mutual information subject to a constraint on expected single-letter distortion. The rate-distortion theory also guarantees the existence of an asymptotically optimal block code that achieves the rate-distortion bound with increasing the blocklength for an arbitrarily distortion level. This claim is known as the source coding theorem.

Though it has been an open problem to construct the asymptotically optimal code guaranteed by the source coding theorem, little is known about construction of the asymptotically optimal code even for simple memoryless sources. There are two main reasons for the source coding theorem not being constructively proved. One is that the rate-distortion function used to be only calculated by minimizing the mutual information without constructing any codes. The other is that the source coding theorem is usually proved for a large class of sources such as stationary ergodic sources and memoryless sources. In 1968 Sakrison [9] proved the source coding theorem for memoryless gaussian sources as a special case by using a geometrical approach. However, there still remain many basic problems to be unveiled about the construction of the asymptotically optimal code since his result is not so sharp.

Recent studies of vector quantization [10, 11, 12], which has deep connection with the rate-distortion theory, usually treat vector quantizers for practical analogue sources with a finite number of parameters in a vector space of fixed dimension. Those concern optimization schemes of parameters and structures for the vector quantizers and hardly analyze asymptotic behaviors of the block quantizers theoretically with increasing the blocklength infinitely.

In another aspect of vector quantization, rate-distortion properties of block quantizers in a fixed dimension are analyzed approximately in the case that the number of quantization levels is sufficiently large [13, 14, 15, 16]. However, from the viewpoint of the rate-distortion theory, the analyses are valid only for sufficiently small distortion level and gives little intuition about the construction of the asymptotically optimal code in a sense of the source coding theorem.

This chapter is concerned with characterizing a class of optimal fixed-to-fixed length codes for memoryless gaussian sources under the squared-error criterion. This is the first attempt to prove the source coding theorem in a constructive way. In an encoding scheme defined in Section 2.2 source blocks of blocklength n are treated as elements of n-dimensional Euclidean space. They are encoded by two steps: a set of point on the n-dimensional unit hypersphere encodes block shapes and a scalar quantizer encodes block shapes. In Section 2.3 asymptotic behavior of rate and average distortion of the encoding scheme is evaluated in detail. It is proved that the method has the asymptotic optimality in a sense of source coding theory; it achieves the rate-distortion function with increasing the blocklength. The asymptotic behavior of this encoding algorithm describes a trade-off relationship between the rate and the average distortion more clearly and tightly than Sakrison's results. It is also surprising to note that the rate-distortion function is naturally deduced as a lower bound of compression efficiency without considering the minimum of mutual information.

The encoding scheme proposed in Section 2.2 is easily extended to a universal data compression scheme for memoryless gaussian sources with unknown mean or variance. The extension of the encoding scheme is treated in Section 2.4. There are two stages in the universal scheme for encoding of an n-tuples from the source. First, an encoder estimates the unknown parameters and transmits them to the decoder in their quantized form. After transmitting the unknown parameter, it encodes the source block by using the estimated and quantized parameters. A rate-distortion property of the universal scheme asymptotically becomes the same as the one obtained in non-universal case. This result clarifies that the encoding scheme has a robustness against perturbation of the transmitted parameters.

Encoding of other memoryless continuous sources is discussed in Section 2.5. Though the source coding theorem holds for a class of stationary ergodic sources with abstract alphabet, difficulty arises even for simple memoryless sources such as memoryless Laplacian sources with known parameters. Though the asymptotic optimal code is not characterized for such sources, a geometrical interpretation of the Shannon lower bound is deduced by using the approach developed for compression of memoryless gaussian sources. A class of memoryless continuous sources that this approach is meaningful is also given in Section 2.5.

2.2 Encoding Scheme

This section is devoted to proposal of a block coding scheme for memoryless gaussian source with known mean and variance. Without loss of generality, the memoryless gaussian sources with zero mean and unit variance is assumed. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$ be an n-tuple of source output and assume $n \geq 3$. From the assumption of the source, each element of $\boldsymbol{x}$ is drawn according to $N(0,1)$ and take real values. The probability density function of $\boldsymbol{x}$ is written as

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)\right]. \quad (2.1)$$

For $\nu \geq 1$ the norm of $\boldsymbol{x}$ is defined as follows:

$$\|\boldsymbol{x}\|_\nu = \sqrt[\nu]{\sum_{i=1}^{n} |x_i|^\nu}. \quad (2.2)$$

When $\nu = 2$, $\|\boldsymbol{x}\|_\nu$ becomes a usual Euclidean norm. For any $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \boldsymbol{R}^n$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \in \boldsymbol{R}^n$ define inner-product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} x_i y_i. \quad (2.3)$$

Let S^{n-1} be the n-dimensional unit hypersphere in the sense of $\nu = 2$, i.e.,

$$S^{n-1} = \{\boldsymbol{x} \in \boldsymbol{R}^n \mid \|\boldsymbol{x}\|_2 = 1\}. \quad (2.4)$$

For any $\boldsymbol{x} \in \boldsymbol{R}^n - \{\mathbf{o}\}$, the orthogonal projection of $\boldsymbol{x}$ to S^{n-1} is defined as

$$\tilde{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}. \quad (2.5)$$

Note that $\tilde{\boldsymbol{x}}$ is uniformly distributed on S^{n-1} if $\boldsymbol{x}$ is generated from the source.

Next, three sets used in the block coding scheme are defined. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_L\}$ and $\mathcal{C} = \{c_1, c_2, \ldots, c_L\}$ be two arbitrary sets of L real numbers satisfying

$$0 = a_1 \leq c_1 < a_2 \leq c_2 < \cdots < a_{L-1} \leq c_{L-1} < a_L \leq c_L < \infty. \quad (2.6)$$

It is convenient to interpret a_{L+1} as infinity. The two sets $\mathcal{A}$ and $\mathcal{B}$ describe a scalar quantizer for block gain $\|\boldsymbol{x}\|_2$. Elements of $\mathcal{A}$ and $\mathcal{B}$ define quantization intervals and quantization levels, respectively. Let $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ be an arbitrary set of cardinality M satisfying $\boldsymbol{y}_m \in S^{n-1}$ for all $m = 1, 2, \ldots, M$. Elements of $\mathcal{Y}$ are used for quantizing block shape $\tilde{\boldsymbol{x}}$. Encoding is defined as a mapping $\varphi : \boldsymbol{R}^n \to \mathcal{Y} \times \mathcal{C}$ specified by two mappings $\varphi_1 : \boldsymbol{R}^n \to \mathcal{Y}$ and $\varphi_2 : \boldsymbol{R}^n \to \mathcal{C}$. The mappings φ_1, φ_2 and φ are defined as follows:

$$\begin{aligned}
\varphi_1(\boldsymbol{x}) &= \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle, & (2.7)\\
\varphi_2(\boldsymbol{x}) &= c_l \quad \text{if} \quad a_l \leq \langle \boldsymbol{x}, \varphi_1(\boldsymbol{x}) \rangle < a_{l+1}, & (2.8)\\
\varphi(\boldsymbol{x}) &= \varphi_2(\boldsymbol{x})\, \varphi_1(\boldsymbol{x}), & (2.9)
\end{aligned}$$

where arg $\max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ in (2.7) means the element of $\mathcal{Y}$ maximizing the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and the right hand side of (2.9) denotes the multiplication of $\varphi_2(\boldsymbol{x})$ by $\varphi_1(\boldsymbol{x})$. Firstly, for a source block $\boldsymbol{x}$ the encoder searches for the elements of $\boldsymbol{y} \in \mathcal{Y}$ maximizing the inner-product first. Ties in (2.7) are arbitrarily broken. Secondly, the encoder searches for the quantization interval of the scalar quantizer that $\langle \boldsymbol{x}, \varphi_1(\boldsymbol{x}) \rangle$ belongs. Indices of $\varphi_1(\boldsymbol{x})$ and $\varphi_2(\boldsymbol{x})$ are transmitted to a decoder in a fixed-length binary form. The decoder reproduces $\boldsymbol{x}$ as $\varphi(\boldsymbol{x}) = \varphi_2(\boldsymbol{x})\varphi_1(\boldsymbol{x})$. Rate R required for transmitting the indices is written as

$$R = \frac{1}{n}\log_2 LM = \frac{1}{n}\log_2 L + \frac{1}{n}\log_2 M. \tag{2.10}$$

In (2.8) assumption of non-negativity of $\langle \boldsymbol{x}, \varphi_1(\boldsymbol{x}) \rangle$ for all $\boldsymbol{x} \in \boldsymbol{R}^n$ seems strange. However, in the case of $M \geq 2$ it is possible to make the inner-product non-negative by choosing an arbitrary point on S^{n-1} as $\boldsymbol{y}_1$ and its antipodal point as $\boldsymbol{y}_2$. Since sufficiently large n is supposed in the following sections, assumption on the non-negativity does not cause any problems. Figure 2.1 gives an illustration of quantization regions of this encoding scheme. It indicates that blocks are encoded to the nearest point in the sense of Euclidean distance.

Let $d(x, y)$ be a single-letter fidelity criterion. Distortion between $\boldsymbol{x}$ and $\varphi(\boldsymbol{x})$ is defined by

$$d_n(\boldsymbol{x}, \varphi(\boldsymbol{x})) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, (\varphi(\boldsymbol{x}))_i), \tag{2.11}$$

where x_i and $(\varphi(\boldsymbol{x}))_i$ denote the i-th component of $\boldsymbol{x}$ and $\varphi(\boldsymbol{x})$, respectively. The squared-error criterion, i.e., $d(x, y) = (x - y)^2$ is assumed for encoding of gaussian sources. Then, d_n can be expressed as

$$d_n(\boldsymbol{x}, \varphi(\boldsymbol{x})) = \frac{1}{n}\|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2, \tag{2.12}$$

which means the squared-error per source symbol. Average distortion $\bar{D}$ caused by the mapping φ is defined by

$$\begin{aligned} \bar{D} &= \int_{\boldsymbol{R}^n} d_n(\boldsymbol{x}, \varphi(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{n}\int_{\boldsymbol{R}^n} \|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2 p(\boldsymbol{x})d\boldsymbol{x}. \end{aligned} \tag{2.13}$$

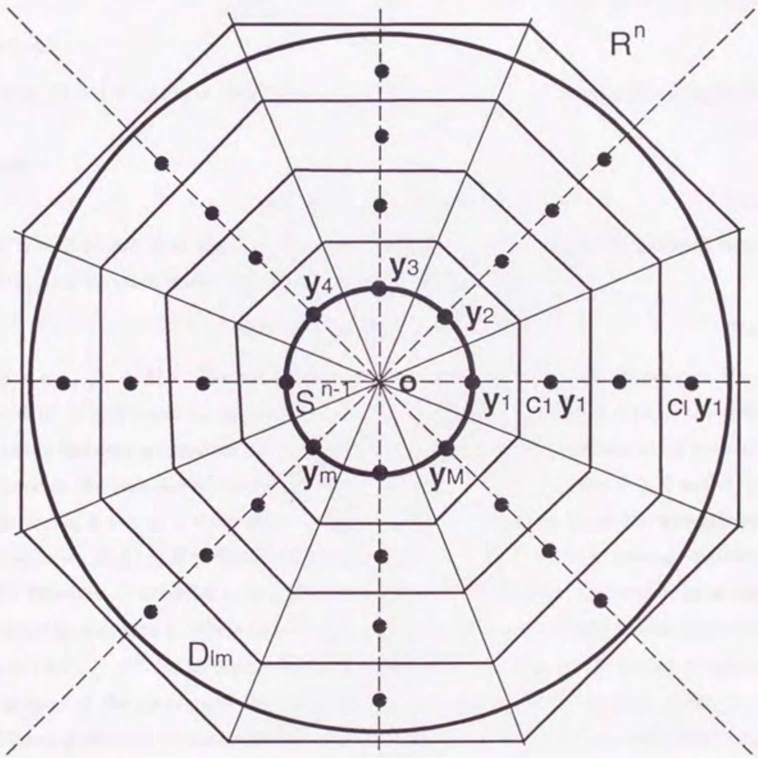


Figure 2.1 Quantization Regions

The rate-distortion function of the source with probability density function $p(x)$ with respect to the distortion measure d is defined by

$$\begin{aligned} R(D) &= \inf_{W \in \mathcal{W}(p,D)} I(p;W) \\ &= \inf_{W \in \mathcal{W}(p,D)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)W(y|x)\log_2 \frac{W(y|x)}{q(y)} dydx, \end{aligned} \tag{2.14}$$

where

$$\mathcal{W}(p,D) = \left\{ W(y|x) \text{is a conditional probability} \,\Big|\, \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)W(y|x)d(x,y)dydx \le D \right\} \tag{2.15}$$

and

$$q(y) = \int_{-\infty}^{\infty} p(x)W(y|x)dx. \tag{2.16}$$

It is well-known that the rate-distortion function of the memoryless gaussian source with unit variance under the squared-error criterion is written as

$$R(D) = \frac{1}{2}\log_2 \frac{1}{D}, \qquad D \in (0,1]. \tag{2.17}$$

(See e.g., [2, 4, 5].) The rate-distortion function $R(D)$ in (2.17) is positive for all $D \in (0,1)$ and equal to zero for $D = 1$. For any distortion level $\Delta \in (0,1]$ the source coding theorem guarantees the existence of an asymptotically optimal block code that achieves the rate-distortion bound $R(\Delta)$. In other words, for any $\delta \ge 0$ and $\varepsilon \ge 0$ satisfying $\delta + \varepsilon > 0$ there exists a block code of sufficiently large blocklength with properties $R(\Delta) \le R \le R(\Delta) + \delta$ in rate R and $\Delta \le \bar{D} \le \Delta + \varepsilon$ in average distortion $\bar{D}$. However, it is trivial to describe an encoder and a decoder that make the average distortion equal to 1. While an encoder transmits nothing, a decoder always reproduces $\mathbf{o} = (0, 0, \ldots, 0)^T$ for all source block $\boldsymbol{x}$. In this case, the average distortion is equal to variance of the source and therefore the rate-distortion function is achievable. In the following sections characterization of the asymptotically optimal encoding scheme for an arbitrarily fixed distortion level $\Delta \in (0,1)$ as a non-trivial case.

2.3 Analysis of Asymptotic Behavior

This section is devoted to characterization of the asymptotically optimal code with increasing the blocklength. For guaranteeing the asymptotic optimality of the mapping defined in (2.7), (2.8) and (2.9), certain conditions are imposed on $\mathcal{A}$ and $\mathcal{C}$ in Section 2.3.1. In Section 2.3.2 three basic lemmata based on properties of Euclidean space are shown in order to facilitate the proof of theorems given in the subsequent section. In Section 2.3.3 the source coding theorem for memoryless gaussian sources under the squared-error criterion is directly proved. As a result, a structure that the asymptotically optimal code should have is obtained.

2.3.1 Conditions for the Scalar Quantizer

From the probability density function (2.1) for $\boldsymbol{x} \in \boldsymbol{R}^n$, it is easy to see that the probability density function of $\|\boldsymbol{x}\|_2$ is dependent on n. Therefore, it is necessary to choose $\mathcal{A}$ and $\mathcal{C}$ as a function of n in order to keep the average distortion small. The following three conditions are imposed on the scalar quantizer described by $\mathcal{A}$ and $\mathcal{C}$:

C1) For each blocklength n, the elements of $\mathcal{A}$ satisfy $0 = a_1 < a_2 < \cdots < a_L$, and $a_L = c_L = n^{\frac{1}{2}+\alpha}$, where $\alpha > 0$ is an arbitrary, but fixed, constant.

C2) The maximal quantization error $\zeta = \max_{1 \leq l \leq L-1} \max\{c_l - a_l, a_{l+1} - c_l\}$ caused by the scalar quantizer for inputs not greater than a_L satisfies

$$\lim_{n \to \infty} \frac{\zeta^2}{n} = 0,$$

where $\max\{x, y\} = y$ if $x \leq y$ and x otherwise.

C3) The number of quantization levels L is of polynomial order of n. That is, L satisfies

$$\lim_{n \to \infty} \frac{\log_2 L}{n} = 0.$$

There are many choices of $\mathcal{A}$ and $\mathcal{C}$. For example, if $L, \mathcal{A}$ and $\mathcal{C}$ are chosen such as

$$L = n^2 + 1, \qquad a_l = c_l = \frac{l-1}{\sqrt{n}} \quad \text{for all} \quad l = 1, 2, \ldots, L, \tag{2.18}$$

it is easy to check that these parameters satisfy C1), C2), C3) and (2.6).

Of course, L, $\mathcal{A}$ and $\mathcal{C}$ should be optimized to minimize the average distortion for encoding of finite blocklength. However, it is shown that $\mathcal{A}$ and $\mathcal{C}$ satisfying C1) $\sim$ C3) and (2.6) can be arbitrarily chosen so as to establish the asymptotic optimality of the block coding scheme.

2.3.2 Basic Lemmata

Let $\Delta \in (0,1)$ be a distortion level arbitrarily fixed and choose $\mathcal{A}$ and $\mathcal{C}$ satisfying C1) $\sim$ C3) and (2.6) arbitrarily. Before proving the asymptotic optimality of the encoding scheme defined in Section 2.2, three lemmata are given so as to facilitate the proof. These lemmata characterize properties of n-dimensional Euclidean space.

Define quantization regions D_{lm} by

$$D_{lm} = \{\boldsymbol{x} \in \boldsymbol{R}^n \mid \varphi(\boldsymbol{x}) = c_l \boldsymbol{y}_m\} \tag{2.19}$$

for all $l = 1, 2, \ldots, L$ and $m = 1, 2, \ldots, M$. The following lemma indicates that the distortion between an arbitrary $\boldsymbol{x} \in \boldsymbol{R}^n - \{\mathrm{o}\}$ and $\varphi(\boldsymbol{x})$ can be divided into two parts.

Lemma 2.1 *Let $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ be an arbitrary set whose all elements belong to S^{n-1}. For any $\boldsymbol{x} \in \boldsymbol{R}^n - \{\mathrm{o}\}$, if $\boldsymbol{x} \in D_{lm}$ for some $l = 1, 2, \ldots, L$ and $m = 1, 2, \ldots, M$, the squared-error between $\boldsymbol{x}$ and $\varphi(\boldsymbol{x})$ is divided into two parts as follows:*

$$\|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2 = \|\boldsymbol{x}\|_2^2 \|\tilde{\boldsymbol{x}} - \langle \tilde{\boldsymbol{x}}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m\|_2^2 + (\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - c_l)^2, \tag{2.20}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|_2$.

Proof: If $\boldsymbol{x} \in D_{lm}$, $\boldsymbol{x}$ is mapped to $\varphi(\boldsymbol{x}) = c_l \boldsymbol{y}_m$ from the definition of D_{lm}. Then, the left hand side of (2.20) is calculated in the following manner:

$$\begin{aligned} \|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2 &= \|\boldsymbol{x} - c_l \boldsymbol{y}_m\|_2^2 \\ &= \|\boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m + \langle \boldsymbol{x}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m - c_l \boldsymbol{y}_m\|_2^2 \\ &= \|\boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m\|_2^2 + \|\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m - c_l \boldsymbol{y}_m\|_2^2 \\ &= \|\boldsymbol{x}\|_2^2 \|\tilde{\boldsymbol{x}} - \langle \tilde{\boldsymbol{x}}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m\|_2^2 + (\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - c_l)^2, \end{aligned} \tag{2.21}$$

where the third equality in (2.21) follows since $\boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m$ is orthogonal to $\boldsymbol{y}_m$. □

Note that $\|\tilde{\boldsymbol{x}} - \langle \tilde{\boldsymbol{x}}, \boldsymbol{y}_m \rangle \boldsymbol{y}_m\|_2^2$ in (2.20) only depends on elements on S^{n-1}. Its multiplication by $\|\boldsymbol{x}\|_2^2$ corresponds to the first term in (2.20). Hereafter, the term is called distortion *in shape*. On the other hand, the second term in (2.21) corresponds to the squared-error caused by the scalar quantizer and is called distortion *in gain*.

For the fixed distortion level $\Delta \in (0,1)$ define M by

$$M = \left(\frac{1}{\Delta}\right)^{\frac{n}{2}}, \tag{2.22}$$

which, according to (2.17), means $\frac{1}{n}\log_2 M = R(\Delta)$. Since M denotes the cardinality of $\mathcal{Y}$, M in (2.22) is interpreted as $\lfloor\left(\frac{1}{\Delta}\right)^{\frac{n}{2}}\rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x. Though this notation seems rude, it makes an essential part of the following two lemmata clear. In fact, it is unnecessary to consider the rounding error when asymptotic properties with increasing n is of interest.

The following two lemmata show an interesting property of the rate-distortion function (2.17). For any $\Delta \in (0,1)$ and $\boldsymbol{z} \in S^{n-1}$ they indicate that, if M elements are chosen randomly and independently from S^{n-1}, there exists at least one element $\boldsymbol{y} \in \mathcal{Y}$ satisfying $\langle \boldsymbol{z}, \boldsymbol{y} \rangle$ is nearly equal to $\sqrt{1-\Delta}$ with probability arbitrarily close to 1 for sufficiently large n.

Lemma 2.2 *Let $\Delta \in (0,1)$ be arbitrarily fixed and define M by (2.22). Fix any η satisfying $0 < \eta < 1 - \Delta$. If $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M\}$ is generated by choosing all of the elements of $\mathcal{Y}$ randomly with uniform distribution and independently from the points on S^{n-1}, then for any $0 < \lambda < 1$ and $\boldsymbol{z} \in S^{n-1}$ there exists an integer $n_0 = n_0(\eta, \lambda)$ satisfying*

$$\Pr\left(\max_{1 \le m \le M} \langle \boldsymbol{z}, \boldsymbol{y}_m \rangle < \sqrt{1-(\Delta+\eta)}\right) < \lambda \tag{2.23}$$

for all $n > n_0$, where $\Pr(\cdot)$ denotes the probability with respect to the choice of $\mathcal{Y}$.

Proof: Fix $\boldsymbol{z} \in S^{n-1}$ arbitrarily and consider the probability of the event that $\max_{1 \le m \le M} \langle \boldsymbol{z}, \boldsymbol{y}_m \rangle < \sqrt{1-(\Delta+\eta)}$ with respect to the choice of $\mathcal{Y}$. For any $\boldsymbol{z}' \in S^{n-1}$ and $\rho \in (0,1)$ define $T_n(\boldsymbol{z}', \rho)$ by

$$T_n(\boldsymbol{z}', \rho) = \left\{\boldsymbol{y} \in S^{n-1} \mid \langle \boldsymbol{z}', \boldsymbol{y} \rangle \ge \sqrt{1-\rho}\right\}. \tag{2.24}$$

Notice that the left hand side of (2.23) means the probability that all elements of $\mathcal{Y}$ belong to $S^{n-1} - T_n(\boldsymbol{z}, \Delta+\eta)$. Therefore it is calculated as follows:

$$\Pr\left(\max_{1 \le m \le M} \langle \boldsymbol{z}, \boldsymbol{y}_m \rangle < \sqrt{1-(\Delta+\eta)}\right) = \left[1 - \frac{|T_n(\boldsymbol{z}, \Delta+\eta)|}{|S^{n-1}|}\right]^M, \tag{2.25}$$

where $|\cdot|$ denotes its surface area. Symmetry on S^{n-1} enables to choose $\boldsymbol{z} = \boldsymbol{z}_0 = (1, 0, \ldots, 0)^T$. By transforming the rectangular coordinates $(x_1, x_2, \ldots, x_n)^T$ into the polar coordinates $(r, \theta_1, \ldots, \theta_{n-1})$, $|T_n(\boldsymbol{z}_0, \Delta + \eta)|$ is lower-bounded in the following manner:

$$\begin{aligned} |T_n(\boldsymbol{z}_0, \Delta + \eta)| &= |S^{n-2}| \int_0^\gamma \sin^{n-2}\theta_1 d\theta_1 \\ &> |S^{n-2}| \int_0^\gamma \sin^{n-2}\theta_1 \cos\theta_1 d\theta_1 \\ &= \frac{|S^{n-2}|}{n-1}(\Delta + \eta)^{\frac{n-1}{2}}, \end{aligned} \tag{2.26}$$

where $\gamma = \cos^{-1}\sqrt{1 - (\Delta + \eta)}$. The inequality $(1-t)^m < \exp[-tm]$ for $0 < t < 1$ and $m > 0$ and (2.26) imply that the right hand side of (2.23) is evaluated in the following way:

$$\begin{aligned} &\Pr\left(\max_{1 \le m \le M} \langle \boldsymbol{z}, \boldsymbol{y}_m \rangle < \sqrt{1 - (\Delta + \eta)}\right) \\ &< \exp\left[-\left(\frac{1}{\Delta}\right)^{\frac{n}{2}} \frac{|S^{n-2}|}{(n-1)|S^{n-1}|}(\Delta + \eta)^{\frac{n-1}{2}}\right] \\ &= \exp\left[-\frac{|S^{n-2}|}{\sqrt{\Delta + \eta}(n-1)|S^{n-1}|} \exp\left[\frac{n}{2}\ln\left(1 + \frac{\eta}{\Delta}\right)\right]\right]. \end{aligned} \tag{2.27}$$

The remaining work is to evaluate $|S^{n-2}|/|S^{n-1}|$. By using a well-known fact that $|S^{n-1}| = n\pi^{\frac{n}{2}}/\Gamma(\frac{n}{2}+1)$, it is easy to verify

$$\frac{|S^{n-2}|}{|S^{n-1}|} = \begin{cases} \dfrac{1}{\pi}\dfrac{(2m-2)!!}{(2m-3)!!}, & \text{if} \quad n = 2m, \\ \dfrac{1}{2}\dfrac{(2m-1)!!}{(2m-2)!!}, & \text{if} \quad n = 2m+1, \end{cases} \tag{2.28}$$

where $m = 1, 2, \ldots$ and $k!! = \prod_{j=0}^{\lfloor \frac{k-1}{2} \rfloor} (k - 2j)$. Equation (2.28) implies that

$$\frac{1}{\pi} \le \frac{|S^{n-2}|}{|S^{n-1}|} \le \frac{n}{2} \tag{2.29}$$

for all n. Inequalities (2.27) and (2.29) yield

$$\begin{aligned} &\Pr\left(\max_{1 \le m \le M} \langle \boldsymbol{z}, \boldsymbol{y}_m \rangle < \sqrt{1 - (\Delta + \eta)}\right) \\ &< \exp\left[-\frac{1}{\pi\sqrt{\Delta + \eta}(n-1)} \exp\left[\frac{n}{2}\ln\left(1 + \frac{\eta}{\Delta}\right)\right]\right], \end{aligned} \tag{2.30}$$

which converges to 0 of double-exponential order of n. Consequently, (2.30) guarantees the existence of n_0 that satisfies the right hand side of (2.30) being less than λ for all $n > n_0$. □

Lemma 2.3 *Let $\Delta \in (0,1)$ be arbitrarily fixed and define M by (2.22). Fix η satisfying $0 < \eta < \Delta$ arbitrarily. If $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ is generated by choosing all of the elements of $\mathcal{Y}$ randomly and independently from the points on S^{n-1}, then for any $0 < \lambda < 1$ and $z \in S^{n-1}$ there exists an integer $n_0 = n_0(\eta, \lambda)$ satisfying*

$$\Pr\left(\max_{1\le m\le M} \langle z, y_m\rangle < \sqrt{1-(\Delta-\eta)}\right) > 1-\lambda \tag{2.31}$$

for all $n > n_0$, where $\Pr(\cdot)$ denotes the probability with respect to the choice of $\mathcal{Y}$.

Proof: This lemma is proved by the same argument used in the proof of Lemma 2.2. As is shown in (2.25), the probability of the event that $\max_{1\le m\le M} \langle z, y_m\rangle < \sqrt{1-(\Delta-\eta)}$ can be written as

$$\Pr\left(\max_{1\le m\le M} \langle z, y_m\rangle < \sqrt{1-(\Delta-\eta)}\right) = \left[1 - \frac{|T_n(z_0, \Delta-\eta)|}{|S^{n-1}|}\right]^M, \tag{2.32}$$

where $z_0 = (1, 0, \ldots, 0)^T$. To find a lower-bound of this probability, the inequality $(1-t)^m > \exp[-\frac{mt}{1-t}]$ for $0 < t < 1$ and $m > 0$ is applied to the right hand side of (2.32). Terms corresponding to t and mt in the right hand side of (2.32) are evaluated in the following manner:

$$\begin{aligned}
\frac{|T_n(z_0, \Delta-\eta)|}{|S^{n-1}|} &= \frac{|S^{n-2}|}{|S^{n-1}|}\int_0^{\gamma'} \sin^{n-2}\theta_1 d\theta_1 \\
&< \frac{\gamma' n}{2}(\Delta-\eta)^{\frac{n-2}{2}} \to 0 \qquad (n\to\infty), &(2.33)\\
M\frac{|T_n(z_0, \Delta-\eta)|}{|S^{n-1}|} &= \left(\frac{1}{\Delta}\right)^{\frac{n}{2}} \frac{|S^{n-2}|}{|S^{n-1}|}\int_0^{\gamma'} \sin^{n-2}\theta_1 d\theta_1 \\
&< \left(\frac{1}{\Delta}\right)^{\frac{n}{2}} \frac{n}{2}\gamma'(\Delta-\eta)^{\frac{n-2}{2}} \\
&= \frac{\gamma' n}{2(\Delta-\eta)}\exp\left[\frac{n}{2}\ln\left(1-\frac{\eta}{\Delta}\right)\right], &(2.34)
\end{aligned}$$

where $\gamma' = \cos^{-1}\sqrt{1-(\Delta-\eta)}$, and (2.29) is used to obtain (2.33) and (2.34). Then, (2.33) and (2.34) imply that

$$\begin{aligned}\Pr\left(\max_{1\le m\le M}\langle z, y_m\rangle < \sqrt{1-(\Delta-\eta)}\right)\\ > \exp\left[-\frac{C_n\gamma' n}{2(\Delta-\eta)}\exp\left[\frac{n}{2}\ln\left(1-\frac{\eta}{\Delta}\right)\right]\right],\end{aligned} \tag{2.35}$$

where $\{C_n\}_{n=1}^{\infty}$ is a sequence satisfying $C_n > 1$ for all $n = 1, 2, \ldots$, and $C_n \to 1$ as $n \to \infty$. Since the right hand side of (2.35) converges to 1 as $n \to \infty$, the claim of this lemma follows. □

2.3.3 Asymptotic Optimality of the Encoding Scheme

The following theorem claims the asymptotic optimality of the mapping φ defined by (2.7), (2.8) and (2.9) for memoryless gaussian sources under the squared-error criterion. The theorem also provides a simple proof for the direct part of the source coding theorem for memoryless gaussian sources under the squared-error criterion.

Theorem 2.1 *Let $\Delta \in (0,1)$ be a distortion level arbitrarily fixed. Let $\mathcal{A}$ and $\mathcal{C}$ be arbitrary sets satisfying* C1) $\sim$ C3) *and* (2.6). *If all of elements of $\mathcal{Y}$ are chosen randomly and independently from the points on S^{n-1}, then for any $\delta > 0$ there exists an integer $n_0 = n_0(\mathcal{A}, \mathcal{C}, \delta)$ such that the rate R of the code satisfies*

$$R(\Delta) < R < R(\Delta) + \delta \tag{2.36}$$

and the average distortion $\bar{D}$ satisfies

$$E[\bar{D}] < \Delta \tag{2.37}$$

for all blocklength $n > n_0$, where $E[\bar{D}]$ denotes the expectation of $\bar{D}$ with respect to the choice of $\mathcal{Y}$.

Proof: Choose $\mathcal{A}$ and $\mathcal{C}$ satisfying C1) $\sim$ C3) and (2.6) arbitrarily and fix any $\delta > 0$. If R satisfies (2.36), monotone decrease of the rate-distortion function with respect to D leads to the existence of $\eta > 0$ satisfying

$$R(\Delta) < R(\Delta - 2\eta) < R(\Delta) + \delta. \tag{2.38}$$

Define M by

$$M = \left(\frac{1}{\Delta - 2\eta}\right)^{\frac{n}{2}}, \tag{2.39}$$

where M is interpreted as the greatest integer not greater than the right hand side of (2.39). Then, rate R required for this encoding scheme becomes

$$R = R(\Delta - 2\eta) + \frac{1}{n}\log_2 L. \tag{2.40}$$

Condition C3) and definition of η guarantees the existence of an integer n_1 satisfying the rate of code satisfies (2.38) for all $n > n_1$.

By applying Lemma 2.1 to the average distortion defined in (2.13), the expectation of the average distortion can be divided into two parts as follows:

$$\begin{aligned} E[\bar{D}] &= E\left[\frac{1}{n}\int_{\boldsymbol{R}^n} \|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2 p(\boldsymbol{x})d\boldsymbol{x}\right] \\ &= E\left[\frac{1}{n}\sum_{m=1}^{M}\sum_{l=1}^{L}\int_{D_{lm}} \|\boldsymbol{x}\|_2^2\|\tilde{\boldsymbol{x}} - \langle\tilde{\boldsymbol{x}}, \boldsymbol{y}_m\rangle \boldsymbol{y}_m\|_2^2 p(\boldsymbol{x})d\boldsymbol{x}\right] \\ &\quad + E\left[\frac{1}{n}\sum_{m=1}^{M}\sum_{l=1}^{L}\int_{D_{lm}} (\langle\boldsymbol{x}, \boldsymbol{y}_m\rangle - c_l)^2 p(\boldsymbol{x})d\boldsymbol{x}\right] \\ &\stackrel{\text{def}}{=} I_1 + I_2, \end{aligned} \tag{2.41}$$

where

$$I_1 = E\left[\frac{1}{n}\sum_{m=1}^{M}\sum_{l=1}^{L}\int_{D_{lm}} \|\boldsymbol{x}\|_2^2\|\tilde{\boldsymbol{x}} - \langle\tilde{\boldsymbol{x}}, \boldsymbol{y}_m\rangle \boldsymbol{y}_m\|_2^2 p(\boldsymbol{x})d\boldsymbol{x}\right], \tag{2.42}$$

$$I_2 = E\left[\frac{1}{n}\sum_{m=1}^{M}\sum_{l=1}^{L}\int_{D_{lm}} (\langle\boldsymbol{x}, \boldsymbol{y}_m\rangle - c_l)^2 p(\boldsymbol{x})d\boldsymbol{x}\right]. \tag{2.43}$$

Roughly speaking, I_1 and I_2 correspond to the average distortion in shape and in gain, respectively.

Firstly, it is shown that there exists an integer n_2 satisfying $I_1 < \Delta - \frac{\eta}{2}$ for all $n > n_2$. Since the expectation operator and the integral in (2.42) are commutable, I_1 can be written as

$$I_1 = \frac{1}{n}\int_{\boldsymbol{R}^n} \|\boldsymbol{x}\|_2^2 E\left[1 - \left(\max_{1\le m\le M}\langle\tilde{\boldsymbol{x}}, \boldsymbol{y}_m\rangle\right)^2\right] p(\boldsymbol{x})d\boldsymbol{x}. \tag{2.44}$$

By substituting Δ in (2.23) into $\Delta - 2\eta$, for any $\lambda \in (0,1)$ and $\tilde{\boldsymbol{x}} \in S^{n-1}$ Lemma 2.2 guarantees the existence of an integer n_2 satisfying

$$\left(\max_{1\le m\le M}\langle\tilde{\boldsymbol{x}}, \boldsymbol{y}_m\rangle\right)^2 < 1 - (\Delta - \eta) \quad \text{with probability at most } \lambda \tag{2.45}$$

for all $n > n_2$. Then, by setting $\lambda = \frac{\eta}{2}$, the expectation in (2.44) is upper-bounded by

$$\begin{aligned} E\left[1 - \left(\max_{1 \le m \le M} \langle \tilde{\boldsymbol{x}}, \boldsymbol{y}_m \rangle\right)^2\right] &< (\Delta - \eta) + 1 \cdot \frac{\eta}{2} \\ &= \Delta - \frac{\eta}{2}, \end{aligned} \tag{2.46}$$

for all $n > n_2$. Combining (2.44) with (2.46) yields

$$\begin{aligned} I_1 &< \frac{1}{n} \int_{\boldsymbol{R}^n} \|\boldsymbol{x}\|_2^2 p(\boldsymbol{x}) d\boldsymbol{x} \cdot \left(\Delta - \frac{\eta}{2}\right) \\ &= \Delta - \frac{\eta}{2} \end{aligned} \tag{2.47}$$

for all $n > n_2$.

Secondly, it is shown that there exists an integer n_3 satisfying $I_2 < \frac{\eta}{2}$ for all $n > n_3$. It is convenient to write I_2 as follows:

$$\begin{aligned} I_2 &= E\left[\frac{1}{n} \sum_{m=1}^{M} \sum_{l=1}^{L-1} \int_{D_{lm}} (\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - c_l)^2 p(\boldsymbol{x}) d\boldsymbol{x}\right] \\ &\quad + E\left[\frac{1}{n} \sum_{m=1}^{M} \int_{D_{Lm}} (\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - a_L)^2 p(\boldsymbol{x}) d\boldsymbol{x}\right], \end{aligned} \tag{2.48}$$

where $c_L = a_L$ from the condition C1). Since the definitions of D_{lm} and ζ imply that

$$(\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - c_l)^2 < \zeta^2 \tag{2.49}$$

for all $l = 1, 2, \ldots, L-1$ and $m = 1, 2, \ldots, M$, the first term in the right hand side of (2.48) is upper-bounded in the following manner:

$$\begin{aligned} E\left[\frac{1}{n} \sum_{m=1}^{M} \sum_{l=1}^{L-1} \int_{D_{lm}} (\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle - c_l)^2 p(\boldsymbol{x}) d\boldsymbol{x}\right] &< \frac{\zeta^2}{n} \int_{\boldsymbol{R}^n} p(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{\zeta^2}{n}, \end{aligned} \tag{2.50}$$

and therefore from the condition C2) there exists an integer n_3 such that the right hand side of (2.50) becomes less than $\frac{\eta}{4}$ for all $n > n_3$. The second term in (2.48) converges to 0 of exponential order of n as is shown in Appendix A, and accordingly there exists an integer n_4 such that it becomes less than $\frac{\eta}{4}$. Then for all $n > \max\{n_3, n_4\}$,

$$I_2 < \frac{\eta}{4} + \frac{\eta}{4} = \frac{\eta}{2}. \tag{2.51}$$

Inequalities (2.47) and (2.51) immediately yield

$$E[\bar{D}] < \left(\Delta - \frac{\eta}{2}\right) + \frac{\eta}{2} = \Delta \tag{2.52}$$

for all $n > \max\{n_2, n_3, n_4\}$. By setting $n_0 = \max\{n_1, n_2, n_3, n_4\}$, the claim of this theorem follows. □

In the proof of Theorem 2.1, it is easy to see that the most dominant term tending to zero are $\frac{1}{n}\log_2 L$ in rate and $\frac{\zeta^2}{n}$ in average distortion. Moreover, (2.37) guarantees the existence of $\mathcal{Y}^*$ that satisfies $\bar{D} < \Delta$ instead of (2.37). For any distortion level $\Delta \in (0,1)$ if such $\mathcal{Y}^*$ is selected, performance of the rate and the average distortion is written in the following form:

$$R \; < \; R(\Delta) + \frac{\log_2 L}{n} + o\left(\frac{\log_2 L}{n}\right), \tag{2.53}$$

$$\bar{D} \; < \; \Delta + \frac{\zeta^2}{n} + o\left(\frac{\zeta^2}{n}\right). \tag{2.54}$$

It is remarkable that a trade-off relationship between the rate and the average distortion in (2.53) and (2.54) is concentrated only on the choice of the scalar quantizer of block gain. If $\mathcal{A}$ and $\mathcal{C}$ defined in (2.18) are used, R and $\bar{D}$ converge of order less than $O(\frac{\log_2 n}{n})$ and $O(\frac{1}{n^2})$, respectively. Sakrison [9] obtained upper-bounds of order $O(\frac{\log_2 n}{n})$ in the rate and $O(\frac{1}{n})$ in the average distortion. Results in (2.53) and (2.54) are not only tighter upper-bounds than the ones Sakrison obtained but also describe the trade-off relationship more generally.

Theorem 2.1 also gives an intuition for constructing block codes of finite blocklength for memoryless gaussian sources. As is shown in (2.46) and (2.47), if η is chosen sufficiently small, the contribution of $\boldsymbol{x}$ satisfying $(\max_{1\le m\le M}\langle\tilde{\boldsymbol{x}}, \boldsymbol{y}_m\rangle)^2 < 1-(\Delta-\eta)$ to the average distortion can be arbitrarily small for sufficiently large n from the property of Lemma 2.2. Hence, for constructing $\mathcal{Y}$ of finite blocklength for an arbitrarily distortion level $\Delta \in (0,1)$ it is desirable to choose a set $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K\}$ of minimal cardinality satisfying $\|\boldsymbol{s}_k\|_2 = 1$ for all $k = 1, 2, \ldots, K$ and

$$\max_{1\le k\le K}\langle \boldsymbol{z}, \boldsymbol{s}_k\rangle \ge \sqrt{1-(\Delta-\eta)} \tag{2.55}$$

for any $z \in S^{n-1}$. However, by using the reciprocal form of (2.33), the rate of $\mathcal{S}$ is lower-bounded in the following way:

$$\begin{aligned}\frac{1}{n}\log_2 K &> \frac{1}{n}\log_2\left[\frac{|S^{n-1}|}{|T_n(z_0,\Delta-\eta)|}\right] \\ &> \frac{1}{n}\log_2\left[\frac{2\gamma'}{n}(\Delta-\eta)^{-\frac{n-2}{2}}\right] \\ &= R(\Delta-\eta)+O\left(\frac{\log_2 n}{n}\right), \end{aligned} \tag{2.56}$$

where $z_0 = (1, 0, \ldots, 0)^T$ and $\gamma' = \cos^{-1}\sqrt{1-(\Delta-\eta)}$. Inequality (2.56) means that, even if $\mathcal{S}$ of minimal cardinality satisfying (2.55) could be constructed explicitly, it is impossible to make the rate of the code less than the rate-distortion bound in an asymptotic sense.

It is also interesting to consider the distortion between an arbitrary $x \in \boldsymbol{R}^n$ and $\varphi(x)$. The following theorem suggests that, for any given distortion level Δ and sufficiently large n, the distortion of source outputs of blocklength n is nearly equal to $\Delta\|x\|_2^2/n$ in the proposed block coding scheme.

Theorem 2.2 *Let $\Delta \in (0,1)$ be a distortion level arbitrarily given. Choose $\mathcal{A}$ and $\mathcal{C}$ satisfying* C1) $\sim$ C3) *and* (2.6). *If all of the elements of $\mathcal{Y}$ are chosen randomly and independently from the points on S^{n-1}, then for any $\varepsilon > 0$ and x satisfying $\|x\|_2 \le a_L$ there exists an integer $n_0 = n_0(\mathcal{A}, \mathcal{C}, \varepsilon)$ such that the rate of the code satisfies* (2.36) *and*

$$\left|\frac{1}{n}E\left[\|x-\varphi(x)\|_2^2\right] - \Delta\frac{\|x\|_2^2}{n}\right| < \varepsilon \tag{2.57}$$

for all $n > n_0$, where $E[\cdot]$ denotes the expectation with respect to the choice of $\mathcal{Y}$.

Proof: From the definition of R there exists a positive number η satisfying (2.38). Define M by (2.39) and fix any x satisfying $\|x\|_2 \le a_L$. Let $\varphi_1(x) = y_m$ and $\varphi_2(x) = c_l$. Note that x satisfies $(\langle x, y_m\rangle - c_l)^2 \le \zeta^2$ since $\langle x, y_m\rangle \le \|x\|_2 \le a_L$. By using Lemma 2.1, the left hand side of (2.57) is evaluated as follows:

$$\frac{1}{n}E[\|x-\varphi(x)\|_2^2] = \frac{1}{n}\|x\|_2^2 E\left[1-\left(\max_{1\le m\le M}\langle\tilde{x}, y_m\rangle\right)^2\right] + \frac{1}{n}(\langle x, y_m\rangle - c_l)^2. \tag{2.58}$$

It is necessary to evaluate lower and upper bounds of $E[1 - (\max_{1 \le m \le M} \langle \tilde{x}, y_m \rangle)^2]$ for the proof of this theorem. The upper one can be found in (2.46), which implies the existence of an integer n_1 satisfying

$$E\left[1 - \left(\max_{1 \le m \le M} \langle \tilde{x}, y_m \rangle\right)^2\right] < \Delta - \frac{\eta}{2} \tag{2.59}$$

for all $n > n_1$.

Now, the lower one is evaluated. By substituting Δ in (2.31) into $\Delta - 2\eta$, for any $\lambda' \in (0, 1)$ and $\tilde{x} \in S^{n-1}$, Lemma 2.3 guarantees the existence of integer n_2 such that

$$\left(\max_{1 \le m \le M} \langle \tilde{x}, y_m \rangle\right)^2 < 1 - (\Delta - 3\eta) \quad \text{with probability at least } 1 - \lambda' \tag{2.60}$$

for all $n > n_2$. By setting $\lambda' = \frac{2\eta}{\Delta - 3\eta}$ the left hand side of (2.59) is lower-bounded as

$$\begin{aligned} E\left[1 - \left(\max_{1 \le m \le M} \langle \tilde{x}, y_m \rangle\right)^2\right] &> (\Delta - 3\eta)(1 - \lambda') \\ &= \Delta - \eta \end{aligned} \tag{2.61}$$

for all $n > n_2$. Inequalities (2.59) and (2.61) yield

$$\Delta - \eta < E\left[1 - \left(\max_{1 \le m \le M} \langle \tilde{x}, y_m \rangle\right)^2\right] < \Delta - \frac{\eta}{2} \tag{2.62}$$

for all $n > \max\{n_1, n_2\}$.

By using (2.58) and (2.62), the left hand side of (2.57) is evaluated in the following form:

$$\left| \frac{1}{n} E\left[\|x - \varphi(x)\|_2^2\right] - \Delta \frac{\|x\|_2^2}{n} \right| < \frac{\|x\|_2^2}{n} \eta + \frac{\zeta^2}{n}. \tag{2.63}$$

Since $\frac{\zeta^2}{n}$ becomes arbitrarily small and η can be chosen of order $o(n^{-2\alpha})$ there exists an integer n_3 such that $(\zeta^2 + \|x\|_2^2 \eta)/n < \varepsilon$ for all $n > n_3$. Setting $n_0 = \max\{n_1, n_2, n_3\}$ completes the proof of the theorem. □

2.4 A Universal Data Compression Algorithm

The asymptotic optimality of the block coding scheme proposed in Section 2.2 is proved in the preceding section for memoryless gaussian sources with known mean and variance. In case of zero mean and unit variance, source blocks of blocklength n are encoded in gain by a scalar quantizer and in shape by a set of randomly and independently chosen points on S^{n-1}. The expectation of the average distortion caused by this encoding scheme is proved to be less than Δ for any fixed $\Delta \in (0,1)$. In this section extension of this algorithm to encoding of memoryless gaussian sources of unknown mean or variance is discussed.

It is assumed that the unknown parameters belong to a known and closed set. [*] An encoder described in this section consists of two stages. For a source block $\boldsymbol{x} \in \boldsymbol{R}^n$ it estimates the unknown parameters first by the maximum-likelihood estimation and transmits the estimated parameters in a quantized form to a decoder. Secondly, it determines codewords for the source block by using the estimated and quantized parameters. In this two-stage encoding strategy precision of the transmitted parameters should be determined in order to make effects caused by a gap between true parameters and the transmitted ones negligible subject to a constraint on rate.

In Section 2.4.1 the encoding scheme proposed in Section 2.2 is slightly modified for encoding of memoryless gaussian sources with known mean μ and variance σ^2. Another aspect of asymptotic properties of the scheme is deduced from a viewpoint to evaluate the probability that source blocks are not encoded within an acceptable distortion. Extension of the encoding scheme for memoryless gaussian sources with unknown parameters is discussed in Section 2.4.2. Asymptotic properties of the universal data compression scheme are competitive with the one for the sources with known mean and variance, though extra $\frac{1}{2}\log_2 n + Const.$ bits are required per unknown parameter for encoding a source block of blocklength n.

2.4.1 Modified Encoding Algorithm for the Sources with Known Parameters

The mapping defined in (2.7), (2.8) and (2.9) is intended to encode outputs of the memoryless gaussian source of zero mean and unit variance. They should be modified

[*] As far as encoding of memoryless gaussian sources is considered, there are at most two unknown parameters. Since there is only one unknown parameter in the case that only mean or variance is unknown, expression such as "unknown parameter(s)" should be used. For simplicity, however, terminology of "unknown parameters" is used for sources with one unknown parameter.

in a form applicable to the memoryless gaussian sources of mean μ and variance σ^2. The rate-distortion function for the source is written as

$$R(D) = \frac{1}{2}\log_2 \frac{\sigma^2}{D}, \qquad D \in (0, \sigma^2] \tag{2.64}$$

[2, 4, 5]. Fix any distortion level $\Delta \in (0, \sigma^2)$. Since the probability that source blocks are encoded within the distortion level is of interest, another form of asymptotic optimality of the proposed scheme is claimed in this section.

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_L\}$ and $\mathcal{C} = \{c_1, c_2, \ldots, c_L\}$ be two sets of non-negative numbers satisfying the following two conditions:

C4) The elements of $\mathcal{A}$ and $\mathcal{C}$ satisfy

$$0 = a_1 \leq c_1 < a_2 \leq c_2 < \cdots < a_L \leq c_L \tag{2.65}$$

C5) The element a_L satisfies $a_L \geq n^{\frac{1}{2}+\alpha}$ for a fixed and positive constant α.

A scalar quantizer of the block gain is characterized by $\mathcal{A}$ and $\mathcal{C}$. It is easy to see that conditions C4) and C5) are weaker than (2.6) and C1), respectively. Define ζ by

$$\zeta = \max_{1 \leq l \leq L-1} \max\{c_l - a_l,\ a_{l+1} - c_l\}, \tag{2.66}$$

where $\max\{x, y\} = y$ if $x \leq y$ and x otherwise. Conditions such as C2) and C3) are not imposed here since they are rather artificial conditions to guarantee the asymptotic optimality in a sense of the source coding theorem.

The shape of source blocks is quantized by $\mathcal{Y}^* = \{y_1, y_2, \ldots, y_M\}$ satisfying the following three conditions:

- $\|y_m\|_2 = 1$ for all $m = 1, 2, \ldots, M$.
- The cardinality of $\mathcal{Y}^*$ is

$$M = \left(\frac{\sigma^2}{\Delta}\right)^{\frac{n}{2}}, \tag{2.67}$$

  i.e., $\frac{1}{n}\log_2 M = R(\Delta)$,

- For any $z \in S^{n-1}$ and $\delta \in (0, \sigma^2 - \Delta)$ define $\chi(z, \mathcal{Y}^*)$ by

$$\chi(z, \mathcal{Y}^*) = \begin{cases} 0, & \text{if there exists } y \in \mathcal{Y}^* \text{ satisfying } \langle x, y \rangle \geq \sqrt{1 - \frac{\Delta+\delta}{\sigma^2}}, \\ 1, & \text{otherwise.} \end{cases} \tag{2.68}$$

Then, there exists $\beta = \beta(\delta) > 0$ satisfying

$$\frac{1}{|S^{n-1}|} \int_{S^{n-1}} \chi(z, \mathcal{Y}^*) dz \leq \exp[-\exp(n\beta)], \tag{2.69}$$

where the integral in the left hand side of (2.69) denotes the surface integral on S^{n-1}.

Existence of $\mathcal{Y}^*$ with all of these three properties is established by the same way used for obtaining (2.30) in Lemma 2.2. Inequality (2.69) means that the ratio of the surface area on S^{n-1} not satisfying $\max_{1 \leq m \leq M} \langle z, y_m \rangle \geq \sqrt{1 - \frac{\Delta+\delta}{\sigma^2}}$ to the total surface area $|S^{n-1}|$ converges to 0 of double-exponential order of n.

Now, the mappings (2.7),(2.8) and (2.9) are modified to the sources with mean μ and variance σ^2 as follows:

$$\varphi_1(x) = \arg\max_{y \in \mathcal{Y}^*} \langle x - \mu, y \rangle, \tag{2.70}$$

$$\varphi_2(x) = c_l \quad \text{such that} \quad a_l \leq \left\langle \frac{x - \mu}{\sigma}, \varphi_1(x) \right\rangle < a_{l+1}, \tag{2.71}$$

$$\varphi(x) = \mu + \sigma \varphi_2(x) \varphi_1(x), \tag{2.72}$$

where $\mu = (\mu, \mu, \ldots, \mu)^T$, a_{L+1} is interpreted as infinity and arg max in (2.70) means the argument $y \in \mathcal{Y}^*$ maximizing the inner-product $\langle x - \mu, y \rangle$. Encoding is described by three mappings $\varphi_1 : R^n \to \mathcal{Y}^*$, $\varphi_2 : R^n \to \mathcal{C}$ and $\varphi : R^n \to \mathcal{C} \times \mathcal{Y}^*$.

Rate required for transmitting a source block $x \in R^n$ is

$$R_n = \frac{n}{2} \log_2 \frac{\sigma^2}{\Delta} + \log_2 L. \tag{2.73}$$

The first term and the second term in (2.73) are the rate for transmitting indices of $\varphi_1(x)$ and $\varphi_2(x)$, respectively. Distortion between x and $\varphi(x)$ is defined by

$$D_n(x, \varphi(x)) = \|x - \varphi(x)\|_2^2. \tag{2.74}$$

Notice that both (2.73) and (2.74) are defined not per symbol but per source block, while in the preceding sections rate and distortion per source symbol are discussed.

As in the same manner used in the proof of Lemma 2.1, $D_n(\boldsymbol{x}, \varphi(\boldsymbol{x}))$ is represented in the following form:

$$
\begin{aligned}
D_n(\boldsymbol{x}, \varphi(\boldsymbol{x})) &= \|\boldsymbol{x} - \varphi(\boldsymbol{x})\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{\mu} - \sigma\varphi_2(\boldsymbol{x})\varphi_1(\boldsymbol{x})\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{\mu} - \langle \boldsymbol{x} - \boldsymbol{\mu}, \varphi_1(\boldsymbol{x})\rangle \varphi_1(\boldsymbol{x}) + \langle \boldsymbol{x} - \boldsymbol{\mu}, \varphi_1(\boldsymbol{x})\rangle \varphi_1(\boldsymbol{x}) - \sigma\varphi_2(\boldsymbol{x})\varphi_1(\boldsymbol{x})\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{\mu} - \langle \boldsymbol{x} - \boldsymbol{\mu}, \varphi_1(\boldsymbol{x})\rangle \varphi_1(\boldsymbol{x})\|_2^2 \\
&\qquad + \|\langle \boldsymbol{x} - \boldsymbol{\mu}, \varphi_1(\boldsymbol{x})\rangle \varphi_1(\boldsymbol{x}) - \sigma\varphi_2(\boldsymbol{x})\varphi_1(\boldsymbol{x})\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \left[1 - \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2}, \varphi_1(\boldsymbol{x}) \right\rangle^2\right] + \sigma^2 \left| \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\sigma}, \varphi_1(\boldsymbol{x}) \right\rangle - \varphi_2(\boldsymbol{x}) \right|^2 \\
&\stackrel{\text{def}}{=} D_{shape} + D_{gain}, \qquad (2.75)
\end{aligned}
$$

where

$$
D_{shape} = \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \left[1 - \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2}, \varphi_1(\boldsymbol{x}) \right\rangle^2\right] \qquad (2.76)
$$

and

$$
D_{gain} = \sigma^2 \left| \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\sigma}, \varphi_1(\boldsymbol{x}) \right\rangle - \varphi_2(\boldsymbol{x}) \right|^2. \qquad (2.77)
$$

The following theorem characterizes another aspect of asymptotic property of the mapping defined in (2.70), (2.71) and (2.72).

Theorem 2.3 *Let $\Delta \in (0, \sigma^2)$ be a distortion level arbitrarily fixed. Choose $\mathcal{A}$ and $\mathcal{C}$ satisfying* C4) *and* C5) *arbitrarily. Then, for any $\delta > 0$ and $\varepsilon \in (0,1)$ there exists an integer $n_0 = n_0(\delta, \varepsilon)$ satisfying*

$$
\Pr(D_{shape} \le n(\Delta + \delta)) > 1 - \varepsilon \qquad (2.78)
$$

and

$$
\Pr(D_{gain} \le \sigma^2\zeta^2) > 1 - \varepsilon \qquad (2.79)
$$

for all $n > n_0$.

Proof: Fix $\delta > 0$ and $\varepsilon \in (0, \frac{1}{2})$ arbitrarily. For $\delta_1 > 0$ and $\delta_2 \in (0, \sigma^2 - \Delta)$ define events E_1, E_2 and E_3 as follows:

$$
E_1 \ : \ \left| \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 - \sigma^2 \right| > \delta_1, \qquad (2.80)
$$

$$E_2 \ : \ \left\langle \frac{\boldsymbol{X}-\mu}{\|\boldsymbol{X}-\mu\|_2}, \varphi_1(\boldsymbol{X}) \right\rangle < \sqrt{1-\frac{\Delta+\delta_2}{\sigma^2}}, \quad (2.81)$$

$$E_3 \ : \ \|\boldsymbol{X}-\mu\|_2^2 > \sigma^2 n^{1+2\alpha}, \quad (2.82)$$

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T \in \boldsymbol{R}^n$ denotes a random vector that satisfies $X_i \sim N(\mu, \sigma^2)$ for all $i = 1, 2, \ldots, n$. Denote by E_i^c, $i = 1, 2, 3$ the complement of the events E_i. Note that E_1^c and E_2^c imply

$$\|\boldsymbol{X}-\mu\|_2^2 \leq n(\sigma^2+\delta_1) \quad (2.83)$$

and

$$\left[1-\left\langle \frac{\boldsymbol{X}-\mu}{\|\boldsymbol{X}-\mu\|_2}, \varphi_1(\boldsymbol{x}) \right\rangle^2\right] \leq \frac{\Delta+\delta_2}{\sigma^2}, \quad (2.84)$$

respectively. Hence, D_{shape} is evaluated under the events $E_1^c \cap E_2^c$ as follows:

$$D_{shape} \leq n(\sigma^2+\delta_1)\frac{\Delta+\delta_2}{\sigma^2}, \quad (2.85)$$

which becomes less than $n(\Delta+\delta)$ by choosing δ_1 and δ_2 that satisfy $\frac{\Delta}{\sigma^2}\delta_1 + \delta_2 + \frac{\delta_1\delta_2}{\sigma^2} \leq \delta$. On the other hand, event E_3^c means that D_{gain} becomes less than $\sigma^2\zeta^2$ from the definition of ζ.

Therefore, it is important to evaluate the probability of the events E_1, E_2 and E_3. By the weak law of large numbers, for any $\delta_1 > 0$ there exists an integer n_1 satisfying $\Pr(E_1) < \frac{\varepsilon}{2}$ for all $n > n_1$. Moreover, there exists an integer n_2 that satisfies $E_3 \subset E_1$ for all $n > n_2$ since α is a fixed and positive constant. By setting $n_3 = \max\{n_1, n_2\}$, $\Pr(E_3) < \frac{\varepsilon}{2} < \varepsilon$ for all $n > n_3$.

Now, $\Pr(E_2)$ is evaluated. The property (2.69) plays an important role. Probability $\Pr(E_2)$ is evaluated in the following manner:

$$\begin{aligned}
\Pr(E_2) &= \int_{\boldsymbol{R}^n} \chi\left(\frac{\boldsymbol{x}-\mu}{\|\boldsymbol{x}-\mu\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\mu\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \int_{\boldsymbol{R}^n} \chi\left(\frac{\boldsymbol{x}-\mu}{\|\boldsymbol{x}-\mu\|_2}, \mathcal{Y}^*\right) \exp\left[-\frac{\|\boldsymbol{x}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \int_0^\infty \exp\left[-\frac{r^2}{2\sigma^2}\right] dr \cdot \int_{\boldsymbol{z}\in S^{n-1}} \chi(\boldsymbol{z}, \mathcal{Y}^*) d\boldsymbol{z} \\
&\leq \exp[-\exp(n\beta)], \quad (2.86)
\end{aligned}$$

where (2.69) is used to obtain the last inequality in (2.86). Equation (2.86) implies the existence of an integer n_4 satisfying $\Pr(E_2) < \frac{\varepsilon}{2}$ for all $n > n_4$. Hence, the union bound leads to

$$\Pr(E_1^c \cap E_2^c) \geq 1 - \Pr(E_1) - \Pr(E_2) > 1 - \varepsilon \tag{2.87}$$

for all $n > \max\{n_1, n_4\}$. The proof of this theorem is completed by setting $n_0 = \max\{n_1, n_3, n_4\}$. □

2.4.2 Universal Encoding Algorithm for the Sources with Unknown Parameters

In this section a universal data compression algorithm for memoryless gaussian sources with unknown mean or variance is discussed. There are three cases as follows:

(a) only μ is unknown,

(b) only σ^2 is unknown,

(c) both μ and σ^2 are unknown.

In each case one of the following assumptions are imposed on the unknown parameters:

A1) the unknown parameters lie in a known and closed domain S,

A2) for any $\varepsilon > 0$ there exists an integer n_0 satisfying that the probability that the unknown parameters belong to a known and closed domain S is greater than $1-\varepsilon$ for all $n > n_0$.

Hereafter, A1) is assumed. However, all of the results can be easily extended to the case of A2) being assumed.

An encoder proposed here estimates them by using the maximum-likelihood estimators. The maximum likelihood estimators of μ and σ^2 are denoted by $\hat{\mu}_{ML}$ and $\hat{\sigma^2}_{ML}$, respectively. In case of (a),(b) and (c), the maximum-likelihood estimators are written in the following form:

(a)

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

(b)

$$\hat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2,$$

(c)

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}_{ML})^2,$$

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ denotes an n-dimensional random vector from the source. It is well-known that the maximum-likelihood estimator has asymptotic normality, i.e., the probability density function of $\hat{\mu}_{ML}$ and $\hat{\sigma^2}_{ML}$ converges *in distribution* to a normal distribution as the sample size n tends to infinity. For example,

$$\sqrt{n}(\mu - \hat{\mu}_{ML}) \sim N(0, \sigma^2) \tag{2.88}$$

in case (a), and

$$\sqrt{n}(\sigma^2 - \hat{\sigma^2}_{ML}) \sim N(0, 2\sigma^4) \tag{2.89}$$

in case (b) for sufficiently large n.

However, the encoder cannot transmit the estimated parameters obtained from the maximum-likelihood estimators since they generally take real values and infinitely large number of bits are required for transmitting them. The encoder can only transmit a quantized value of the estimated parameters. For going through this difficulty, a net that covers the known domain S is introduced. Denote by $\hat{\mu}$ and $\hat{\sigma^2}$ the estimated and quantized parameters. Figure 2.2 shows how the net covers the two-dimensional parameter space S in case of (c). Crosses of the net correspond to the candidates of the pair $(\hat{\mu}, \hat{\sigma^2})$.

Asymptotic normality of the maximum-likelihood estimators is crucial to determine a mesh of the net. Consider case (a), for instance. From the asymptotic normality of the maximum-likelihood estimators, for any $\varepsilon \in (0, 1)$ there exists $K > 0$ satisfying

$$\Pr\left(|\mu - \hat{\mu}_{ML}| > K\frac{\sigma}{\sqrt{n}}\right) < \varepsilon \tag{2.90}$$

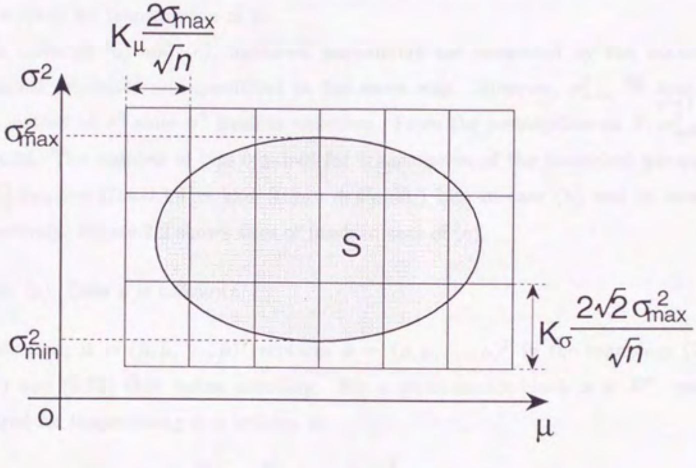


Figure 2.2 A net in the parameter space

for sufficient large n. Let the mesh be equal to $2K\frac{\sigma}{\sqrt{n}}$ and choose $\hat{\mu}$ which is the nearest to $\hat{\mu}_{ML}$ among the candidate of $\hat{\mu}$. It is obvious from the choice of $\hat{\mu}$ that $|\hat{\mu}-\hat{\mu}_{ML}| < 2K\frac{\sigma}{\sqrt{n}}$ is satisfied. Since the triangle inequality implies that

$$|\mu-\hat{\mu}| \leq |\mu-\hat{\mu}_{ML}| + |\hat{\mu}-\hat{\mu}_{ML}|, \tag{2.91}$$

(2.90) leads to

$$\Pr\left(|\mu-\hat{\mu}| > 3K\frac{\sigma}{\sqrt{n}}\right) < \varepsilon. \tag{2.92}$$

Since there are $O(\sqrt{n})$ candidates of the estimated parameters, $(\frac{1}{2}\log_2 n + Const.)$ bits are required for transmission of $\hat{\mu}$.

In cases of (b) and (c), unknown parameters are estimated by the maximum-likelihood estimator and quantized in the same way. However, $\sigma^2_{max} \stackrel{\text{def}}{=} \max_{\sigma^2\in S}\sigma^2$ is used instead of σ^2 since σ^2 itself is unknown. From the assumption on S, σ^2_{max} will be finite. The number of bits required for transmission of the estimated parameters are $(\frac{1}{2}\log_2 n + Const.)$ bits and $(\log_2 n + Const.)$ bits in case (b) and in case (c), respectively. Figure 2.2 shows sizes of mesh in case of (c).

- Case (a): Only μ is unknown

In this case, $\hat{\boldsymbol{\mu}} = (\hat{\mu},\hat{\mu},\ldots,\hat{\mu})^T$ replaces $\boldsymbol{\mu} = (\mu,\mu,\ldots,\mu)^T$ in the mappings (2.70), (2.71) and (2.72) that define encoding. For a given source block $\boldsymbol{x} \in \boldsymbol{R}^n$, rate R_n required for transmitting $\boldsymbol{x}$ is written as

$$R_n = \frac{n}{2}\log_2\frac{\sigma^2}{\Delta} + \log_2 L + \frac{1}{2}\log_2 n + Const., \tag{2.93}$$

where the third and the forth terms in (2.93) correspond to a cost for transmitting $\hat{\mu}$. As in the same manner that yields (2.75), $D(\boldsymbol{x},\varphi(\boldsymbol{x}))$ is evaluated in the following form:

$$\begin{aligned} D_n(\boldsymbol{x},\varphi(\boldsymbol{x})) &= \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2\left[1-\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2},\varphi_1(\boldsymbol{x})\right\rangle^2\right] \\ &\qquad + \sigma^2\left|\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\sigma},\varphi_1(\boldsymbol{x})\right\rangle - \varphi_2(\boldsymbol{x})\right|^2 \\ &\stackrel{\text{def}}{=} D_{shape} + D_{gain}, \end{aligned} \tag{2.94}$$

where

$$D_{shape} = \|\boldsymbol{x} - \hat{\boldsymbol{\mu}}\|_2^2 \left[1 - \left\langle \frac{\boldsymbol{x} - \hat{\boldsymbol{\mu}}}{\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}\|_2}, \varphi_1(\boldsymbol{x}) \right\rangle^2\right] \tag{2.95}$$

and

$$D_{gain} = \sigma^2 \left| \left\langle \frac{\boldsymbol{x} - \hat{\mu}}{\sigma}, \varphi_1(\boldsymbol{x}) \right\rangle - \varphi_2(\boldsymbol{x}) \right|^2. \tag{2.96}$$

The following theorem indicates an asymptotic property of D_{shape} and D_{gain}. The asymptotic property described in Theorem 2.3 and Theorem 2.4 are of the same form though extra $\frac{1}{2}\log_2 n + Const.$ bits are required in the latter case.

Theorem 2.4 *Let $\Delta \in (0, \sigma^2)$ be a distortion level arbitrarily fixed and choose $\mathcal{A}$ and $\mathcal{C}$ satisfying* C4) *and* C5)*. Then, for any $\delta > 0$ and $\varepsilon \in (0,1)$ there exists an integer $n_0 = n_0(\delta, \varepsilon)$ satisfying*

$$\Pr(D_{shape} \leq n(\Delta + \delta)) \geq 1 - \varepsilon \tag{2.97}$$

$$\Pr(D_{gain} \leq \sigma^2\zeta^2) \geq 1 - \varepsilon \tag{2.98}$$

for all $n > n_0$.

Proof: Fix $\delta > 0$ and $\varepsilon \in (0,1)$ arbitrarily and let $\varepsilon' \in (0, \varepsilon)$ be a real number satisfying $0 < \varepsilon' < \varepsilon$. For $\delta_1 > 0$ and $\delta_2 \in (0, \sigma^2 - \Delta)$ define four events as follows:

$$E_1 \; : \; \left| \frac{1}{n}\|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2^2 - \sigma^2 \right| > \delta_1, \tag{2.99}$$

$$E_2 \; : \; \left\langle \frac{\boldsymbol{X} - \hat{\boldsymbol{\mu}}}{\|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2}, \varphi_1(\boldsymbol{X}) \right\rangle < \sqrt{1 - \frac{\Delta + \delta_2}{\sigma^2}}, \tag{2.100}$$

$$E_3 \; : \; \|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2^2 > \sigma^2 n^{1+2\alpha}, \tag{2.101}$$

$$E_4 \; : \; |\mu - \hat{\mu}| > 3K\frac{\sigma}{\sqrt{n}}, \tag{2.102}$$

where $\boldsymbol{X}$ denotes an n-dimensional random vector from the source. The definition of K leads to the existence of an integer n_1 satisfying $\Pr(E_4^c) < \varepsilon'$ for all $n > n_1$. Under the event E_4^c, $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2^2$ is upper-bounded as

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2^2 \leq n\left(\frac{3K\sigma}{\sqrt{n}}\right)^2 = 9K\sigma^2, \tag{2.103}$$

which implies that $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2^2$ and $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2$ are of constant order. For proving this theorem, existence of integers n_2 and n_3 satisfying $\Pr(E_1|E_4^c) < \varepsilon'$ for all $n > n_2$ and

$\Pr(E_2|E_4^c) < \varepsilon'$ for all $n > n_3$ should be established. The proof on the existence of such n_3 is lengthy and given in Appendix B.

Now, $\Pr(E_1|E_4^c)$ is evaluated. Note that the triangle inequality yields

$$\|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2 - \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \|\boldsymbol{X} - \boldsymbol{\mu}\|_2 \leq \|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2 + \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2. \tag{2.104}$$

If for any fixed $\delta_1' \in (0, \sigma^2 - \Delta)$

$$\|\boldsymbol{X} - \boldsymbol{\mu}\|_2 < \sqrt{n(\sigma^2 - \delta_1')} \quad \text{or} \quad \sqrt{n(\sigma^2 + \delta_1')} < \|\boldsymbol{X} - \boldsymbol{\mu}\|_2 \tag{2.105}$$

is satisfied, under the event E_4^c (2.103) and (2.105) guarantee the existence of δ_1 satisfying

$$\|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2 < \sqrt{n(\sigma^2 - \delta_1)} \quad \text{or} \quad \sqrt{n(\sigma^2 + \delta_1)} < \|\boldsymbol{X} - \hat{\boldsymbol{\mu}}\|_2 \tag{2.106}$$

and $0 < \delta_1 < \delta_1'$ for sufficiently large n. Notice that (2.106) is equivalent to the event E_1. Since the weak law of large numbers guarantees the existence of an integer n_2 satisfying that the probability of the event (2.105) is less than ε', $\Pr(E_1) < \varepsilon'$ follows for all integers $n > n_2$.

Under the event $E_1^c \cap E_2^c \cap E_4^c$, D_{shape} is evaluated in the following manner:

$$\begin{aligned} D_{shape} &< n(\sigma^2 + \delta_1)\frac{\Delta + \delta_2}{\sigma^2} \\ &< n(\Delta + \delta), \end{aligned} \tag{2.107}$$

where the last inequality follows by choosing δ_1 and δ_2 satisfying $\frac{\Delta}{\sigma^2}\delta_1 + \delta_2 + \frac{\delta_1\delta_2}{\sigma^2} \leq \delta$. The probability of this event is lower-bounded in the following way:

$$\begin{aligned} \Pr(E_1^c \cap E_2^c \cap E_4^c) &= \Pr(E_4^c)\Pr(E_1^c \cap E_2^c|E_4^c) \\ &\geq \Pr(E_4^c)(1 - \Pr(E_1|E_4^c) - \Pr(E_2|E_4^c)) \\ &> (1 - \varepsilon')(1 - 2\varepsilon') \end{aligned} \tag{2.108}$$

where the last inequality follows for all $n > \max\{n_1, n_2, n_3\}$.

On the other hand, D_{gain} is not greater than $\sigma^2\zeta^2$ under the event $E_3^c \cap E_4^c$. The definition of α leads to the existence of an integer n_4 satisfying $E_3 \subset E_1$ for all $n > n_4$, and hence $\Pr(E_3|E_4^c) < \Pr(E_1|E_4^c) < \varepsilon$. Probability of the event $E_3^c \cap E_4^c$ is evaluated in the following way:

$$\begin{aligned}\Pr(E_3^c \cap E_4^c) &= \Pr(E_4^c)\Pr(E_3^c|E_4^c) \\ &> \Pr(E_4^c)(1-\Pr(E_3|E_4^c)) \\ &> (1-\varepsilon')^2. \end{aligned} \tag{2.109}$$

By choosing ε' satisfying $\max\{3\varepsilon' - 2\varepsilon'^2, 2\varepsilon' - \varepsilon'^2\} < \varepsilon$, (2.108) and (2.109) imply the claim of this theorem. □

• Case (b): Only σ^2 is unknown

In this case, $\sqrt{\hat{\sigma^2}}$ is used in the mapping (2.70), (2.71) and (2.72) instead of σ. For any $\varepsilon > 0$ the mesh K of a net in the parameter space satisfying

$$\Pr\left(|\sigma^2 - \hat{\sigma^2}| > \frac{3K\sigma_{max}^2}{\sqrt{n}}\right) < \varepsilon \tag{2.110}$$

is chosen. Define the cardinality of $\mathcal{Y}^*$ by

$$M = \left(\frac{\hat{\sigma^2}}{\Delta}\right)^{\frac{n}{2}} \tag{2.111}$$

and use $\mathcal{Y}^* = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ for quantizing a source output $\boldsymbol{x}$. Then, rate required for transmitting $\boldsymbol{x}$ is written as

$$R_n = \frac{n}{2}\log_2\frac{\hat{\sigma}^2}{\Delta} + \log_2 L + \frac{1}{2}\log_2 n + Const., \tag{2.112}$$

which is upper-bounded by

$$R < \frac{n}{2}\log_2\frac{\sigma^2}{\Delta} + \frac{3K\sigma_{max}^2}{2\sigma^2}\sqrt{n} + \log_2 L + \frac{1}{2}\log_2 n + Const. \tag{2.113}$$

with probability $1-\varepsilon$, i.e., whenever $|\sigma^2 - \hat{\sigma^2}| > \frac{3K\sigma_{max}^2}{\sqrt{n}}$ is satisfied. Inequality (2.113) clarifies the fact that extra cost of $O(\sqrt{n})$ bits is required by the unknown variance.

Distortion between $\boldsymbol{x}$ and $\varphi(\boldsymbol{x})$ is divided in order to represent distortion in shape and in gain as follows:

$$\begin{aligned} D_n(\boldsymbol{x}, \varphi(\boldsymbol{x})) &= \|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2\left[1 - \left\langle \frac{\boldsymbol{x}-\boldsymbol{\mu}}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}, \varphi_1(\boldsymbol{x})\right\rangle^2\right] \\ &\qquad + \hat{\sigma^2}\left|\left\langle \frac{\boldsymbol{x}-\boldsymbol{\mu}}{\sqrt{\hat{\sigma^2}}}, \varphi_1(\boldsymbol{x})\right\rangle - \varphi_2(\boldsymbol{x})\right|^2 \\ &\stackrel{\text{def}}{=} D_{shape} + D_{gain}, \end{aligned} \tag{2.114}$$

where

$$D_{shape} = \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 \left[1 - \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2}, \varphi_1(\boldsymbol{x}) \right\rangle^2\right] \tag{2.115}$$

and

$$D_{gain} = \hat{\sigma^2} \left| \left\langle \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\sqrt{\hat{\sigma^2}}}, \varphi_1(\boldsymbol{x}) \right\rangle - \varphi_2(\boldsymbol{x}) \right|^2. \tag{2.116}$$

The following theorem characterizes an asymptotic property of this encoding scheme with respect to the distortion, which is of the same form as Theorem 2.3.

Theorem 2.5 *Choose $\mathcal{A}$ and $\mathcal{C}$ satisfying* C4) *and* C5) *arbitrarily. For any fixed distortion level $\Delta \in (0, \sigma^2), \delta > 0$ and $\varepsilon \in (0,1)$ there exists an integer $n_0 = n_0(\mathcal{A}, \mathcal{C}, \delta, \varepsilon)$ that satisfies*

$$\Pr(D_{shape} \leq n(\Delta + \delta)) > 1 - \varepsilon \tag{2.117}$$

and

$$\Pr(D_{gain} \leq \sigma^2 \zeta^2) > 1 - \varepsilon \tag{2.118}$$

for all integers $n > n_0$.

Proof: This theorem is proved by using the same manner that has already developed in the proof of Theorem 2.3 and Theorem 2.4. Fix $\delta > 0$ and $\varepsilon \in (0,1)$ arbitrarily and for $\delta_1 > 0$ and $\delta_2 \in (0, \sigma^2 - \Delta)$ define four events as follows:

$$E_1 \ : \ \left| \frac{1}{n} \|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 - \hat{\sigma^2} \right| > \delta_1, \tag{2.119}$$

$$E_2 \ : \ \left\langle \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2}, \varphi_1(\boldsymbol{X}) \right\rangle < \sqrt{1 - \frac{\Delta + \delta_2}{\hat{\sigma^2}}}, \tag{2.120}$$

$$E_3 \ : \ \|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 > \sigma^2 n^{1+2\alpha}, \tag{2.121}$$

$$E_4 \ : \ |\sigma^2 - \hat{\sigma^2}| > 3K \frac{\sigma_{max}^2}{\sqrt{n}}. \tag{2.122}$$

For any $\varepsilon' > 0$ it is clear that there exist integers n_1 and n_2 that satisfy $\Pr(E_2) < \varepsilon'$ for all $n > n_1$ and $\Pr(E_4) < \varepsilon'$ for all $n > n_2$ by virtue of $\mathcal{Y}^*$ and the choice of K, respectively.

Since $E_3 \subset E_1$ for sufficiently large n and the gap between σ^2 and $\hat{\sigma^2}$ is of order $O(\frac{1}{\sqrt{n}})$, the weak law of large numbers guarantees the existence of an integer n_3 satisfying $\Pr(E_3|E_4^c) < \Pr(E_1|E_4^c) < \varepsilon'$ for all $n > n_3$. Under the event $E_1^c \cap E_2^c \cap E_4^c$, D_{shape}

is evaluated in the following manner:

$$\begin{aligned} D_{shape} &\leq n(\hat{\sigma^2}+\delta_1)\frac{\Delta+\delta_2}{\hat{\sigma^2}} \\ &\leq n(\Delta+\delta) \end{aligned} \quad (2.123)$$

by the appropriate choice of δ_1 and δ_2. Probability of the event $E_1^c \cap E_2^c \cap E_4^c$ is evaluated in the following way:

$$\begin{aligned} \Pr(E_1^c \cap E_2^c \cap E_4^c) &= \Pr(E_2^c)\Pr(E_4^c)(1-\Pr(E_1^c|E_4^c)) \\ &> (1-\varepsilon')^3, \end{aligned} \quad (2.124)$$

where independence of events E_2^c is used to deduce (2.124), which becomes greater than $1-\varepsilon$ by choosing ε' appropriately.

Evaluation on D_{gain} in (2.118) is essentially the same as already discussed in the proof of Theorem 2.4 and therefore omitted. □

• Case (c): Both μ and σ^2 are unknown

Now, a universal data compression scheme for memoryless gaussian sources with unknown mean and variance is discussed. In this case, $\hat{\mu}$ and $\sqrt{\hat{\sigma^2}}$ is used in the mappings (2.70), (2.71) and (2.72) instead of μ and σ, respectively. Let θ denote the pair (μ, σ^2) and $\hat{\theta}_{ML}$ denote the pair $(\hat{\mu}_{ML}, \hat{\sigma^2}_{ML})$. Then, the asymptotic normality of the maximum-likelihood estimators guarantees that $\sqrt{n}(\theta - \hat{\theta}_{ML}) \sim N(0, I(\theta)^{-1})$, where

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \quad (2.125)$$

is the Fisher information matrix. The form of $I(\theta)$ in (2.125) implies that $\hat{\mu}_{ML}$ and $\hat{\sigma^2}_{ML}$ are independently distributed in an asymptotic sense. Therefore, for any $\varepsilon \in (0,1)$ K_μ and K_σ with the following property can be chosen:

$$\Pr\left(|\mu - \hat{\mu}_{ML}| > \frac{K_\mu \sigma_{max}}{\sqrt{n}} \quad \text{or} \quad |\sigma^2 - \hat{\sigma^2}_{ML}| > \frac{K_\sigma \sigma_{max}^2}{\sqrt{n}}\right) < \varepsilon. \quad (2.126)$$

By introducing a net with mesh K_μ on the mean and K_σ on the variance, the estimated parameters $\hat{\mu}$ and $\hat{\sigma^2}$ satisfying the following property is obtained:

$$\Pr\left(|\mu - \hat{\mu}| > \frac{3K_\mu \sigma_{max}}{\sqrt{n}} \quad \text{or} \quad |\sigma^2 - \hat{\sigma^2}_{ML}| > \frac{3K_\sigma \sigma_{max}^2}{\sqrt{n}}\right) < \varepsilon. \quad (2.127)$$

Since there are two parameters to be transmitted, rate of the code can be written as

$$R_n = \frac{n}{2}\log_2\frac{\hat{\sigma^2}}{\Delta} + \log_2 L + \log_2 n + Const., \tag{2.128}$$

which is upper-bounded by

$$R_n = \frac{n}{2}\log_2\frac{\sigma^2}{\Delta} + \frac{3K\sigma_{max}^2}{2\sigma^2}\sqrt{n} + \log_2 L + \frac{1}{2}\log_2 n + Const. \tag{2.129}$$

with probability $1-\varepsilon$ from the property of (2.127). Notice that the coefficient of $\log_2 n$ in (2.128) becomes double compared with the one in (2.112).

For a given source block $\boldsymbol{x}$, $D(\boldsymbol{x}, \varphi(\boldsymbol{x}))$ can be separated in the following manner:

$$\begin{aligned} D_n(\boldsymbol{x}, \varphi(\boldsymbol{x})) &= \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2\left[1-\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \varphi_1(\boldsymbol{x})\right\rangle^2\right] \\ &\quad + \hat{\sigma^2}\left|\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\sqrt{\hat{\sigma^2}}}, \varphi_1(\boldsymbol{x})\right\rangle - \varphi_2(\boldsymbol{x})\right|^2 \\ &\stackrel{\text{def}}{=} D_{shape} + D_{gain}, \end{aligned} \tag{2.130}$$

where

$$D_{shape} = \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2\left[1-\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \varphi_1(\boldsymbol{x})\right\rangle^2\right] \tag{2.131}$$

and

$$D_{gain} = \hat{\sigma^2}\left|\left\langle\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\sqrt{\hat{\sigma^2}}}, \varphi_1(\boldsymbol{x})\right\rangle - \varphi_2(\boldsymbol{x})\right|^2. \tag{2.132}$$

The following theorem shows an asymptotic property of D_{shape} and D_{gain} with respect to the increasing blocklength, which is also in the same form as Theorem 2.3

Theorem 2.6 *For any fixed distortion level $\Delta \in (0, \sigma^2), \delta > 0$ and $\varepsilon \in (0,1)$ there exists an integer $n_0 = n_0(\delta, \varepsilon)$ that satisfies*

$$\Pr(D_{shape} \le n(\Delta+\delta)) > 1-\varepsilon \tag{2.133}$$

and

$$\Pr(D_{gain} \le \sigma^2\zeta^2) > 1-\varepsilon \tag{2.134}$$

for all integers $n > n_0$.

Proof: This theorem is proved by using the same method that establishes Theorem 2.4 and Theorem 2.5. Therefore, only outline is given here. Fix $\delta > 0$ and $\varepsilon \in (0,1)$ arbitrarily and for $\delta_1 > 0$ and $\delta_2 \in (0, \sigma - \Delta)$ define four events as follows:

$$E_1 \ : \ \left| \frac{1}{n} \|\boldsymbol{X} - \hat{\mu}\|_2^2 - \hat{\sigma^2} \right| > \delta_1, \tag{2.135}$$

$$E_2 \ : \ \left\langle \frac{\boldsymbol{X} - \hat{\mu}}{\|\boldsymbol{x} - \hat{\mu}\|_2^2}, \varphi_1(\boldsymbol{X}) \right\rangle < \sqrt{1 - \frac{\Delta + \delta_2}{\hat{\sigma^2}}}, \tag{2.136}$$

$$E_3 \ : \ \|\boldsymbol{X} - \hat{\mu}\|_2^2 > \sigma^2 n^{1+2\alpha}, \tag{2.137}$$

$$E_4 \ : \ |\mu - \hat{\mu}| > \frac{3K_\mu \sigma_{max}}{\sqrt{n}} \quad \text{or} \quad |\sigma^2 - \hat{\sigma^2}| > \frac{3K_\sigma \sigma_{max}^2}{\sqrt{n}} \tag{2.138}$$

For proving this theorem, it is necessary to show the existence of an integer n_1 satisfying $\Pr(E_4^c) < \varepsilon'$, $\Pr(E_1|E_4^c) < \varepsilon'$, $\Pr(E_2|E_4^c) < \varepsilon'$ and $\Pr(E_3|E_4^c) < \varepsilon'$ for all integers $n > n_1$. Then, under the event $E_1^c \cap E_2^c \cap E_4^c$ D_{shape} does not become greater than $n(\Delta + \delta)$ and the probability of this event is greater than $1 - \varepsilon$ if δ_1, δ_2 and ε' are appropriately chosen. On the other hand, D_{gain} becomes less than $\sigma^2 \zeta^2$ under the event $E_3^c \cap E_4^c$. The probability of this event is also greater than $1 - \varepsilon$ by choosing ε' judiciously. This completes the proof of this theorem. □

2.5 Encoding of Other Continuous Sources

Theorem 2.1 proved in Section 2.3.3 shows that the encoding scheme proposed in Section 2.2 has the asymptotic optimality in the sense of the source coding theorem for memoryless gaussian sources of known mean and variance under the squared-error criterion. Though the rate-distortion function defined by (2.14) is an infimum of the mutual information subject to a constraint, its form naturally appears in the analysis of (2.56) as a limit of covering bound. In this section encoding of other continuous sources is discussed. The rate-distortion function for the sources are also defined by (2.14). The source coding theorem guarantees the existence of an asymptotically optimal block code that achieves the rate-distortion bound for an arbitrarily distortion level. However, few codes are proved to achieve the bound. One of reasons why such codes are unknown is that the rate-distortion function is rarely expressed as an explicit function of distortion levels.

Memoryless Laplacian sources is one of the sources that the rate-distortion function is expressed as an explicit function of distortion levels. In Section 2.5.1 encoding for memoryless Laplacian sources is considered from a viewpoint of n-dimensional Euclidean geometry. A geometrical interpretation of the Shannon lower bound is given in Section 2.5.2, which is obtained as a byproduct from such geometrical approach.

2.5.1 Encoding of Memoryless Laplacian Sources

The memoryless Laplacian source with probability density function

$$p(x) = \frac{\lambda}{2}\exp[-\lambda|x|], \qquad E|x| = \frac{1}{\lambda} \tag{2.139}$$

is a source that the rate-distortion function is expressed explicitly as

$$R(D) = \log_2 \frac{1}{\lambda D}, \qquad D \in (0, \tfrac{1}{\lambda}] \tag{2.140}$$

under the magnitude-error criterion, that is, $d(x, y) = |x - y|$.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ be a random vector from the memoryless Laplacian source with probability distribution (2.139). It is easy to verify that

$$E\left[\frac{1}{n}\sum_{k=1}^{n}|X_i|\right] = \frac{1}{\lambda} \tag{2.141}$$

and

$$V\left[\frac{1}{n}\sum_{k=1}^{n}|X_i|\right] = \frac{1}{\lambda^2 n}. \tag{2.142}$$

Equations (2.141) and (2.142) suggest that source blocks of blocklength n will concentrate on neighborhood of the hypersphere $S_1(n, \frac{n}{\lambda})$, where for all $\nu \geq 1$ and $r \geq 0$

$$S_\nu(n, r) = \{\boldsymbol{x} \in R^n \mid \|\boldsymbol{x}\|_\nu = r\} \tag{2.143}$$

and $\|\boldsymbol{x}\|_\nu$ denote the norm defined in (2.2). Therefore, it is quite natural to use a code whose all elements belong to $S_1(\frac{n}{\lambda})$. There are many vector quantizers with a codebook based on such idea. [17, 18, 19, 20, 21]. In particular, Fischer's treatment [20] includes a discussion of the asymptotic optimality with increasing the blocklength. For any distortion level $\Delta \in (0, \frac{1}{\lambda})$ he shows that a codebook of size $(\frac{1}{\lambda\Delta})^n$, i.e., $\frac{1}{n}\log_2 M = R(\Delta)$, generated by choosing randomly and independently points on $S_1(n, \frac{n}{\lambda})$ has the asymptotic optimality with respect to the blocklength. However, his analysis includes a mistake as is shown in Appendix C. Therefore, finding an asymptotically optimal encoding scheme for memoryless Laplacian sources is still an open problem.

2.5.2 Geometrical Interpretation of the Shannon Lower Bound

The rate-distortion function for memoryless continuous sources is defined as an infimum of the mutual information subject to a constraint on expectation of single-letter distortion. However, it is rare to represent the infimum as an explicit function of the distortion levels even for simple memoryless sources and fidelity criterion. The Shannon lower bound, denoted by $R_L(\Delta)$, appears when the Lagrange multipliers are introduced for finding the infimum in (2.14)[2]. The Shannon lower bound is one of lower bounds of the rate-distortion function that is useful when a single-letter distortion measure $d(x, y)$ is a function of $x - y$. Though achievablity of them by block codes is not guaranteed, they are explicitly represented as a function of the distortion levels for a certain class of memoryless continuous sources. The Shannon lower bound for Laplacian memoryless sources of zero mean and the first-order absolute moment $\frac{1}{\lambda}$ and memoryless gaussian sources of zero mean and variance σ^2 are listed in the following table:

	magnitude-error criterion	squared-error criterion
Laplacian	$\log_2 \frac{1}{\lambda\Delta}$*	$\frac{1}{2}\log_2 \frac{2e}{\lambda^2\pi\Delta}$
Gaussian	$\frac{1}{2}\log_2 \frac{\pi\sigma^2}{2e\Delta^2}$	$\frac{1}{2}\log_2 \frac{\sigma^2}{\Delta}$*

where $d(x, y) = |x - y|$ and $d(x, y) = (x - y)^2$ under the magnitude-error criterion and the squared-error criterion, respectively. The Shannon lower bounds with an asterisk are coincident with the rate-distortion functions; they are asymptotically achievable by a block code.

Surprisingly, these four bounds are also obtained by computing the ratio of volume of n-dimensional hyperspheres. Fix any distortion level $\Delta > 0$ and let $V_\nu(n, r)$ be the

n-dimensional volume inside of $S_\nu(n,r)$. By a simple calculation, $V_1(n,r)$ and $V_2(n,r)$ are written as follows:

$$V_1(n,r) = \frac{2^n r^n}{\Gamma(n+1)}, \qquad V_2(n,r) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)}, \tag{2.144}$$

where $\Gamma(\cdot)$ denotes the gamma function. In case of the Laplacian sources, (2.141) and (2.142) suggests the importance of considering $S_1(n, \frac{n}{\lambda})$ and $V_1(n, \frac{n}{\lambda})$. In case of the gaussian sources, if $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ is a random vector from the source, it is easy to verify

$$E\left[\frac{1}{n}\sum_{k=1}^{n}|X_i|^2\right] = \sigma^2, \qquad V\left[\frac{1}{n}\sum_{k=1}^{n}|X_i|^2\right] = \frac{2\sigma^4}{n}, \tag{2.145}$$

which imply the significance of $S_2(n, \sqrt{n\sigma^2})$ and $V_2(n, \sqrt{n\sigma^2})$. For any $\boldsymbol{y} \in \boldsymbol{R}^n$ the n-dimensional volume of the region $\{\boldsymbol{x} \in \boldsymbol{R}^n \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}$ is equal to $V_1(n, n\Delta)$ if $d(x,y) = |x-y|$ and $V_2(n, \sqrt{n\Delta})$ if $d(x,y) = (x-y)^2$. How many $\boldsymbol{y}_m$, $m = 1, 2, \ldots M$ satisfying that $\{\boldsymbol{x} \in \boldsymbol{R}^n \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_m) \le \Delta\}$ does not overlap for all $m = 1, 2, \ldots M$ can be chosen in $S_1(n, \frac{\lambda}{n})$ or $S_2(n, \sqrt{n\sigma^2})$? The following theorem indicates another interpretation of the Shannon lower bound from the viewpoint of evaluating an upper-bound of M.

Theorem 2.7 *Let $\Delta \in (0, \frac{1}{\lambda})$ be a distortion level arbitrarily fixed. Then, the Shannon lower bounds for the memoryless Laplacian source of zero mean and first-order absolute moment $\frac{1}{\lambda}$ are obtained by the following calculation:*

$$\frac{1}{n}\log_2 \frac{V_1(n, \frac{n}{\lambda})}{V_1(n, n\Delta)} = \log_2 \frac{1}{\lambda\Delta}, \tag{2.146}$$

$$\frac{1}{n}\log_2 \frac{V_1(n, \frac{n}{\lambda})}{V_2(n, \sqrt{n\Delta})} = \frac{1}{2}\log_2 \frac{2e}{\lambda^2\pi\Delta} + O\left(\frac{\log_2 n}{n}\right). \tag{2.147}$$

For memoryless gaussian sources of zero mean and variance σ^2, the Shannon lower bounds are deduced by the following calculation:

$$\frac{1}{n}\log_2 \frac{V_2(n, \sqrt{n\sigma})}{V_1(n, n\Delta)} = \frac{1}{2}\log_2 \frac{\pi\sigma^2}{2e\Delta^2} + O\left(\frac{\log_2 n}{n}\right) \tag{2.148}$$

$$\frac{1}{n}\log_2 \frac{V_2(n, \sqrt{n\sigma})}{V_2(n, \sqrt{n\Delta})} = \frac{1}{2}\log_2 \frac{\sigma^2}{\Delta} \tag{2.149}$$

Proof: Both (2.146) and (2.149) immediately follow from (2.144). For establishing (2.147) and (2.148), the Stirling formula is used to evaluate the ratio of n-dimensional hyperspheres. However, the evaluation is not difficult and omitted. □

Note that n-dimensional hyperspheres are similar in the case of (2.146) and (2.147). By using the similarity, results in Theorem 2.7 are extended to the memoryless source with probability density function

$$p(x) = C_1 \exp[-C_2|x|^\nu] \tag{2.150}$$

under the single-letter fidelity criterion $d(x, y) = (x - y)^\nu$, where

$$C_1 = \frac{\nu}{2\Gamma(\frac{1}{\nu})}\sqrt[\nu]{\nu M_\nu}, \qquad C_2 = \sqrt[\nu]{\nu M_\nu} \tag{2.151}$$

and ν is an arbitrary constant satisfying $\nu \geq 1$. The two constants C_1 and C_2 are chosen so that (2.150) becomes the probability density function satisfying $E|X|^\nu = M_\nu$. The Shannon lower bound for the source with probability density function (2.150) can be computed as

$$R_L(D) = \frac{1}{\nu}\log_2\frac{M_\nu}{\Delta}, \qquad D \in (0, M_\nu] \tag{2.152}$$

by a conventional way using the Lagrange multipliers, while it is obvious that

$$\frac{1}{n}\log_2\frac{V_\nu(n, \sqrt[\nu]{nM_\nu})}{V_\nu(n, \sqrt[\nu]{n\Delta})} = R_L(\Delta) \tag{2.153}$$

for all $\Delta \in (0, M_\nu)$ from the similarity of $V_\nu(n, \sqrt[\nu]{nM_\nu})$ and $V_\nu(n, \sqrt[\nu]{n\Delta})$. Equations (2.146) and (2.149) directly follow by setting $\nu = 1$ and $\nu = 2$ in (2.153) though achievablity of the bound (2.152) is not guaranteed.

2.6 Conclusion

In this chapter rate-distortion behaviors of a block coding algorithm for compression of memoryless gaussian sources are analyzed in detail under the squared-error criterion. Geometric properties of the n-dimensional Euclidean space play an important role in the analysis.

First, the block encoding algorithm is proposed in Section 2.2 for memoryless gaussian sources with known mean and variance. Source blocks of blocklength n are treated as elements in n-dimensional Euclidean space and encoded by a scalar quantizer and a point of set on the n-dimensional unit hypersphere. In the case of zero mean and unit variance, for any source block $\boldsymbol{x}$ the scalar quantizer encodes $\|\boldsymbol{x}\|_2$ and the set of points encodes $\boldsymbol{x}/\|\boldsymbol{x}\|_2$, where $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm.

Rate-distortion properties of the algorithm are analyzed in Section 2.3. Given any distortion level, it is shown that the scheme achieves the rate-distortion bound with increasing the blocklength. Upper bounds for asymptotic behavior of rate and average distortion are given, which describe the trade-off relationship between the rate and the average distortion more tightly and clearly than Sakrison's results. The most dominant terms in the asymptotic behavior only depends on the scalar quantizer. The rate-distortion function is naturally deduced as an essential limit of data compression though it is originally defined as an infimum of the mutual information subject to a constraint on expected single-letter distortion.

In Section 2.4 the algorithm is extended to a form that is applicable to the memoryless gaussian sources with unknown mean or variance. Given an n-tuple of source block, encoding consists of two steps in the extended algorithm. Firstly, an encoder estimates the unknown parameters by using the maximum-likelihood estimators and transmits them to a decoder in a quantized form. Secondly, by using the quantized parameters it encodes the source block as if the parameters were known and transmits the codeword to the decoder. The extra cost on rate for transmitting the unknown parameters is $\frac{1}{2}\log_2 n + Const.$ per unknown parameter. The asymptotic normality of the maximum likelihood estimators leads to the cost, which means an application of the MDL(Minimum Description Length) criterion to data compression with fidelity criterion. It is shown that with probability arbitrarily close to 1 the source block of blocklength n is encoded within an acceptable distortion for sufficiently large blocklength whether or not parameters are known.

Applications of the proposed analyses to other continuous memoryless sources are discussed in Section 2.5. As a byproduct of the proposed analysis, a geometrical interpretation of the Shannon lower bound is clarified for memoryless gaussian sources and memoryless Laplacian sources under the magnitude-error criterion and the squared-error criterion. In such cases a ratio between n-dimensional volumes yields the Shannon lower bounds. A class of memoryless sources that the Shannon lower bound can be obtained in the same way is also obtained.

Chapter 3

Asymptotic Properties of Data Compression Algorithms with Fidelity Criterion Based on String Matching

3.1 Introduction

Development of data compression schemes with fidelity criterion is important for the sake of practical applications such as image and speech encodings. A theoretical basis for the data compression methods with fidelity criterion is described in detail by Berger [2]. Given any stationary and ergodic data sequences, it is shown that the rate-distortion bound is asymptotically achievable by a block code with increasing the blocklength under fidelity criterion. However, explicit construction of the asymptotically optimal block code remains unsolved even for simple memoryless sources.

On the other hand, many lossless data compression schemes, such as *compress* command in UNIX operating system, are now in use. In the case of lossless data compression if sources are stationary and ergodic, achievable compression efficiency is the entropy rate. In particular, the Lempel-Ziv codes [7, 22] are well-known as asymptotically optimal universal codes in a sense that the entropy rate is asymptotically achievable without prior knowledge of the sources. The encoding scheme described in [7] is based on string matching. A theoretical basis of the scheme from the viewpoint of evaluating a *recurrence time* of sequences is proposed by Willems [23] and Wyner and Ziv [24] and completed by Ornstein and Weiss [25]. Let $\boldsymbol{X} = \{X_t\}_{t=-\infty}^{\infty}$ be a sequence of random variable from a stationary and ergodic source with finite alphabet $\mathcal{A}$. Denote by $\boldsymbol{x} = \{x_t\}_{t=-\infty}^{\infty}$ a realization of $\boldsymbol{X}$ and define

$$M_n(\boldsymbol{x}) = \min\{M \geq 0 \,:\, x_1 x_2 \cdots x_n = x_{-M-(n-1)} x_{-M-(n-2)} \cdots x_{-M}\}. \tag{3.1}$$

Convergence *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X})$ to the entropy rate is shown in [24] and convergence of $\frac{1}{n}\log_2 M_n(\boldsymbol{X})$ *with probability one* is proved in [25].

Recent studies on data compression schemes with fidelity criterion [26, 27, 28, 29] attempt to realize the asymptotically optimal encoding scheme in a universal way while classical results on universal data compression with fidelity criterion [30, 31, 32] only formulate an achievable rates subject to a constraint on distortion. In particular, an extension of the Lempel-Ziv encoding scheme is discussed by Steinberg and Gutman [26]. For a stationary and ergodic sequence $\boldsymbol{x} = \{x_t\}_{t=-\infty}^{\infty}$, $x_t \in \mathcal{A}$ and appropriately given distortion level $\Delta > 0$ define

$$M_n(\boldsymbol{x}, \Delta) = \min\{M \geq 0 \,:\, d_n(x_1^n, x_{-n(M+1)+1}^{-nM}) \leq \Delta\}, \tag{3.2}$$

where x_i^j denotes $x_i x_{i+1} \cdots x_j$ for any $i \leq j$ and d_n is a distortion measure defined on $\mathcal{A}^n \times \mathcal{A}^n$ satisfying certain properties. It means a *recurrence time with fidelity criterion*. Reference [26] shows that for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta) > R\left(\frac{\Delta}{2}\right) + \varepsilon\right\} = 0 \tag{3.3}$$

for a class of totally ergodic sources, where $R(\cdot)$ is the rate-distortion function of the source and P denotes a probability measure on random variables $\boldsymbol{X} = \{X_t\}_{t=-\infty}^{\infty}$ from the source. Quite recently Kanaya and Muramatsu [33] shows that for any $\varepsilon > 0$

$$\frac{1}{n}\log_2 R_n(\boldsymbol{x}, \Delta) \leq R\left(\frac{\Delta}{2}\right) + \varepsilon \quad \textit{with probability one} \tag{3.4}$$

for a class of stationary ergodic sources, where

$$R_n(\boldsymbol{x}, \Delta) = \min\{m \geq n \,:\, d_n(x_1^n, x_{m+1}^{m+n}) \leq \Delta\}. \tag{3.5}$$

Though (3.3) and (3.5) shows that $R(\frac{\Delta}{2})$ becomes an upper-bound *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ and an upper-bound of $\frac{1}{n}\log_2 R_n(\boldsymbol{X},\Delta)$ *with probability one*, they are not equal to the rate-distortion bound at distortion level Δ. Therefore, these result does not directly imply the asymptotically optimal data compression scheme with fidelity criterion while results in [24, 25] are exactly bases of a lossless data compression scheme.

In this chapter tighter bounds of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ are deduced in the following two cases: (i) discrete memoryless sources with finite alphabet under single-letter fidelity criterion, and (ii) memoryless gaussian sources under the squared-error criterion. Case (i) is discussed in Section 3.2. First, an upper-bound *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ is deduced. For the source with probability distribution $\boldsymbol{p}$ it is shown that for any $\varepsilon > 0$ the probability $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ being greater than $R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \varepsilon$ goes to zero as n tends to infinity, where $R(\boldsymbol{p},\Delta)$ denotes the rate-distortion bound at distortion level Δ, $\boldsymbol{p}^*$ denotes output probability distribution of the test channel and $D(\boldsymbol{p}^*\|\boldsymbol{p})$ denotes the divergence from $\boldsymbol{p}$ to $\boldsymbol{p}^*$. The result leads to a sufficient condition that $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ is upper-bounded *in probability* by the rate-distortion bound itself when $\boldsymbol{p}$ is known. In fact, it converges *in probability* to $R(\boldsymbol{p},\Delta)$ under the probability measure $P^*_{(-\infty,0]} \times P_{[1,\infty)}$, where P and P^* denotes the probability measure on $\boldsymbol{X} = \{X_i\}_{i=-\infty}^{\infty}$ induced by $\boldsymbol{p}$ and $\boldsymbol{p}^*$, respectively and $P_{[i,j]}$ and $P^*_{[i,j]}$ denotes their restriction to $X_i X_{i+1} \cdots X_j$. The result not only shows the sufficient condition but also provide a simple proof of the source coding theorem. Behaviors of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ under another probability measure are also discussed.

Section 3.3 is devoted to analysis of asymptotic behavior of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ in case (ii). As is discussed in Chapter 2, the angle between a source block and a reproduced block closely relates with distortion levels when the two blocks are treated as elements of Euclidean space. If mean and variance of the sources are known, by introducing another way of matching based on the angle $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ converges *in probability* to the rate-distortion bound under the probability measure induced by the sources. Comparison case (i) and case (ii) leads to an essential difference between the asymptotically optimal encoding schemes for the discrete memoryless sources and memoryless gaussian sources. That is, the asymptotically optimal encoding is realized by using an output sequence of the same source of known mean and variance while it is quite difficult in case (ii) even if probability distribution of the sources are known.

Results obtained in the two sections are easily extended to a data compression algorithm with fidelity criterion. In the algorithm both an encoder and a decoder have a buffer of finite length whose contents are the same i.i.d. sequence. Given a source block of blocklength n, the encoder searches for the best-matched block in the buffer and transmits its index. The decoder reproduces the source block from the transmitted index. The obtained results are used for determining the length of the buffers or evaluating average distortion from a fixed buffer length. Results by computer simulation of the algorithm of finite n are also given in Section 3.3.

3.2 Data Compression for Discrete Memoryless Sources

This section is devoted to a data compression scheme with fidelity criterion for discrete memoryless sources of finite alphabet. In Section 3.2.1 a principle of encoding and decoding is given with introducing several notations. Main results are stated in Section 3.2.2. A practical implementation of the data compression scheme is discussed in Section 3.2.3.

3.2.1 Definitions

Let $n > 0$ be an arbitrary integer and $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ a source alphabet. The cardinality J is assumed to be finite. Denote by $\boldsymbol{p}$ a probability distribution on $\mathcal{A}$ and by $p(a_j)$ the probability that symbol a_j is generated. Let $\boldsymbol{X} = \{X_k\}_{k=-\infty}^{\infty}$ be a sequence of random variables from the source and $\boldsymbol{x} = \{x_k\}_{k=-\infty}^{\infty}$ a realization of $\boldsymbol{X}$. For any $i \leq j$ define x_i^j as a string $x_i x_{i+1} \ldots x_j$. The probability measure on $\boldsymbol{X}$ induced by $\boldsymbol{p}$ is denoted by P.

Let $d : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ be a single-letter fidelity criterion satisfying

$$d(a_j, a_k) \begin{cases} = 0, & \text{if } j = k, \\ > 0, & \text{otherwise,} \end{cases} \tag{3.6}$$

for all $j, k = 1, 2, \ldots, J$. Symmetry on d, i.e., $d(a_j, a_k) = d(a_k, a_j)$ for all $j, k = 1, 2, \ldots, J$ is not assumed. Distortion between $\boldsymbol{u} = u_1^n \in \mathcal{A}^n$ and $\boldsymbol{v} = v_1^n \in \mathcal{A}$ is defined by

$$d_n(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{n} \sum_{i=1}^{n} d(u_i, v_i). \tag{3.7}$$

The rate-distortion function is defined as

$$\begin{aligned} R(\boldsymbol{p}, D) &= \min_{W \in \mathcal{W}(\boldsymbol{p}, D)} I(\boldsymbol{p}; W) \\ &= \min_{W \in \mathcal{W}(\boldsymbol{p}, D)} \sum_{k=1}^{J} \sum_{j=1}^{J} p(a_j) W(a_k|a_j) \log_2 \frac{W(a_k|a_j)}{\sum_{l=1}^{J} p(a_l) W(a_j|a_l)}, \end{aligned} \tag{3.8}$$

where $I(\boldsymbol{p}; W)$ denotes the mutual information and

$$\mathcal{W}(\boldsymbol{p}, D) = \Big\{ W \text{ is a } J \times J \text{ stochastic matrix} \,|\, \sum_{j=1}^{J} \sum_{k=1}^{J} p(a_j) W(a_k|a_j) d(a_j, a_k) \leq D \Big\}. \tag{3.9}$$

The rate-distortion function $R(\boldsymbol{p}, \Delta)$ is the achievable lower bound in rate of data compression codes with fidelity criterion. It is positive for all $\Delta \in [0, D_{max})$, where

$$D_{max} = \min_{1 \leq k \leq J} \sum_{j=1}^{J} p(a_j) d(a_j, a_k). \tag{3.10}$$

Now, the problem temporally considered is described. For any realization $\boldsymbol{x}$ and distortion level $\Delta \in (0, D_{max})$ define $M_n(\boldsymbol{x}, \Delta)$ by

$$M_n(\boldsymbol{x}, \Delta) = \min\{M \geq 0 \,:\, x_{-n(M+1)+1}^{-nM} \in B_n(x_1^n, \Delta)\}, \tag{3.11}$$

where for any $\boldsymbol{u} \in \mathcal{A}^n$

$$B_n(\boldsymbol{u}, \Delta) = \{\boldsymbol{v} \in \mathcal{A}^n \,|\, d_n(\boldsymbol{u}, \boldsymbol{v}) \leq \Delta\}. \tag{3.12}$$

Roughly speaking, $\frac{1}{n} \log_2 M_n(\boldsymbol{x}, \Delta)$ bits per source symbol are required for transmission of the index $M_n(\boldsymbol{x}, \Delta)$. How does $\frac{1}{n} \log_2 M_n(\boldsymbol{x}, \Delta)$ behave asymptotically? If $\frac{1}{n} \log_2 M_n(\boldsymbol{x}, \Delta)$ converges *in probability* to $R(\boldsymbol{p}, \Delta)$, i.e., if for any $\delta > 0$

$$\lim_{n \to \infty} P \left\{ \left| \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) - R(\boldsymbol{p}, \Delta) \right| \geq \delta \right\} = 0, \tag{3.13}$$

is satisfied, then the following encoding scheme makes sense.

Let an arbitrarily double-infinite sequence $\boldsymbol{x}$ from the source and distortion level $\Delta \in (0, D_{max})$ be given. Suppose that both an encoder and a decoder have a buffer of length $L = n \cdot 2^{n[R(\boldsymbol{p},\Delta)+\delta]}$ containing x_{-L+1}^{0}. For encoding x_1^n the encoder searches for $M_n(\boldsymbol{x}, \Delta)$ and transmit $\lceil n[R(\boldsymbol{p}, \Delta) + \delta]\rceil$ bits to the decoder. For any $\delta > 0$ if n is sufficiently large, such $M_n(\boldsymbol{x}, \Delta)$ can be found in the buffer with probability arbitrarily close to 1. The block x_1^n is reproduced as $x_{-nM_n(\boldsymbol{x},\Delta)+1}^{-n(M_n(\boldsymbol{x},\Delta)-1)}$. While rate required for transmission of x_1^n is $R(\boldsymbol{p}, \Delta) + \delta$ bits per source symbol, and distortion between the original block and the reproduced block is at most Δ with probability close to 1.

From the viewpoint of evaluating an upper-bound of buffer length, convergence *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ seems redundant since only the upper-bound *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ is essential to the buffer length. However, lower-bounds of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ means the minimum length of the buffer and therefore leads to non-compressibility of data in rate less than the rate-distortion bound. Hence, it is also significant to find conditions for convergence *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ if it has such a property.

3.2.2 Analysis

In the case that a single-letter fidelity criterion d is the Hamming distance, that is, $d(a_j, a_k) = 0\,(j = k)$, $d(a_j, a_k) = 1\,(j \neq k)$, Steinberg and Gutman [26] shows that for any $\delta > 0$

$$\lim_{n\to\infty} P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta) \geq V(\boldsymbol{p}, \Delta) + \delta\right\} = 0, \tag{3.14}$$

where

$$V(\boldsymbol{p}, \Delta) = H(\boldsymbol{p}) - \sum_{j=1}^{J} p(a_j) h\left(\frac{\Delta}{Jp(a_j)}\right) - \Delta\log_2(J-1), \tag{3.15}$$

$$H(\boldsymbol{p}) = -\sum_{j=1}^{J} p(a_j)\log_2 p(a_j) \tag{3.16}$$

and $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$. Though $V(\boldsymbol{p}, \Delta)$ is coincident with $R(\boldsymbol{p}, \Delta)$ when $\boldsymbol{p} = (\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J})$, there is a gap between $V(\boldsymbol{p}, \Delta)$ and $R(\boldsymbol{p}, \Delta)$ for non-uniform sources that cannot be neglected.

The following theorem proposes a new upper-bound *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta)$ substitutable for $V(\boldsymbol{p},\Delta)$ under an arbitrarily single-letter fidelity criterion satisfying (3.6).

Theorem 3.1 *Let a single-letter fidelity criterion d satisfying* (3.6) *and a distortion level $\Delta \in (0, D_{max})$ be arbitrarily given. Then, for any $\delta > 0$ and $\varepsilon \in (0,1)$ there exists an integer $n_0 = n_0(\delta,\varepsilon)$ satisfying*

$$P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \delta\right\} < \varepsilon \tag{3.17}$$

for all $n > n_0$, where $D(\boldsymbol{p}^*\|\boldsymbol{p})$ denotes the divergence defined by

$$D(\boldsymbol{p}^*\|\boldsymbol{p}) = \sum_{j=1}^{J} p^*(a_j)\log_2\frac{p^*(a_j)}{p(a_j)}, \tag{3.18}$$

$\boldsymbol{p}^*$ means the probability distribution on $\mathcal{A}$ defined by

$$p^*(a_k) = \sum_{j=1}^{J} p(a_j)W^*(a_k|a_j), \quad k = 1,2,\ldots,J \tag{3.19}$$

and W^* implies the stochastic matrix satisfying $I(\boldsymbol{p};W^*) = R(\boldsymbol{p},\Delta)$ and $W^* \in \mathcal{W}(\boldsymbol{p},\Delta)$.

Proof: This theorem is proved by a technique similar to the one first introduced by Wyner and Ziv [24]. Fix $\delta > 0$ and $\varepsilon \in (0,1)$ arbitrarily. First, the left hand side of (3.17) is evaluated in the following way:

$$\begin{aligned}
&P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \delta\right\}\\
&= P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \delta \,\middle|\, X_1^n \in T_{[p]}^n\right\}\cdot P\{X_1^n \in T_{[p]}^n\}\\
&\quad + P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \delta \,\middle|\, X_1^n \notin T_{[p]}^n\right\}\cdot P\{X_1^n \notin T_{[p]}^n\}\\
&< P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta) + D(\boldsymbol{p}^*\|\boldsymbol{p}) + \delta \,\middle|\, X_1^n \in T_{[p]}^n\right\}\\
&\quad + P\{X_1^n \notin T_{[p]}^n\},
\end{aligned} \tag{3.20}$$

where $T_{[p]}^n$ denotes the set of strongly typical sequences of blocklength n. According to Lemma 1.2.12 in [3], there exists an integer $n_1(\varepsilon)$ satisfying $P\{X_1^n \notin T_{[p]}^n\} < \frac{\varepsilon}{2}$ for all $n > n_1$.

Now, the first term in the right hand side of (3.20) is evaluated. Assume that

$$P(B_n(x_1^n,\Delta)) \geq 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})]-o(n)} \tag{3.21}$$

for any $x_1^n \in T_{[p]}^n$. Then, the first term in the right hand side of (3.20) is upper-bounded in the following manner:

$$\begin{aligned}
&P\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \geq R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})+\delta \,|\, X_1^n = x_1^n\right\} \\
&\quad = P\{M_n(\boldsymbol{X},\Delta) \geq 2^{n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})+\delta]} \,|\, X_1^n = x_1^n\} \\
&\quad \overset{1)}{\leq} E_P[M_n(\boldsymbol{X},\Delta) \,|\, X_1^n = x_1^n] \cdot 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})+\delta]} \\
&\quad \overset{2)}{=} [P(B_n(x_1^n,\Delta))]^{-1} \cdot 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})+\delta]} \\
&\quad \overset{3)}{\leq} 2^{n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})]+o(n)} \cdot 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})+\delta]} \\
&\quad = 2^{-n\delta+o(n)},
\end{aligned} \tag{3.22}$$

where E_P denote the expectation with respect to the probability measure P, $o(n)$ denotes terms of order less than 1 whose coefficient of the highest term is positive, the Markov inequality implies inequality 1) in (3.22), the property that $\sum_{l=1}^{\infty} l(1-p)^{l-1}p = 1/p$ for any $p \in (0,1)$ to obtain equality 2), and inequality 3) follows from (3.21). Since δ is fixed and positive, there exists an integer $n_2 = n_2(\delta,\varepsilon)$ such that (3.22) is less than $\frac{\varepsilon}{2}$ for all $n > n_2$. Therefore, setting $n_0 = \max\{n_1, n_2\}$ yields the claim of this theorem under the assumption of (3.21).

For completing the proof (3.21) should be established. Inequality (3.21) is proved by using an argument of types. (See [3] for definition of the type, the joint type and the conditional type.) For any $\boldsymbol{u},\boldsymbol{v} \in \mathcal{A}^n$ denote the type of $\boldsymbol{u}$ by $q_{\boldsymbol{u}}$ and the joint type of the pair $(\boldsymbol{u},\boldsymbol{v})$ by $Q_{\boldsymbol{uv}}$, respectively. For given $\boldsymbol{u}$, let $W_{\boldsymbol{v}|\boldsymbol{u}}$ denote the conditional type of $\boldsymbol{v}$, and $T_{\boldsymbol{v}|\boldsymbol{u}}^W$ a set specified by the conditional type $W_{\boldsymbol{v}|\boldsymbol{u}}$, respectively. By setting $\boldsymbol{u} = x_1^n$, $P(B_n(\boldsymbol{u},\Delta))$ can be written in the following form:

$$\begin{aligned}
P(B_n(\boldsymbol{u},\Delta)) &= \sum_{\boldsymbol{v}:d_n(\boldsymbol{u},\boldsymbol{v})\leq\Delta} P\{\boldsymbol{v}\} \\
&= \sum_{W_{\boldsymbol{v}|\boldsymbol{u}}\in\tilde{\mathcal{W}}(q_{\boldsymbol{u}},\Delta)} P\{\boldsymbol{v}\}|T_{\boldsymbol{v}|\boldsymbol{u}}^W|,
\end{aligned} \tag{3.23}$$

where $|\cdot|$ denotes the cardinality of the set and

$$\tilde{\mathcal{W}}(q_{\boldsymbol{u}},\Delta) = \big\{W_{\boldsymbol{v}|\boldsymbol{u}} \text{ is a } J\times J \text{ conditional type}\,|$$

$$\sum_{k=1}^{J}\sum_{j=1}^{J} q_{\boldsymbol{u}}(a_j) W_{\boldsymbol{v}|\boldsymbol{u}}(a_k|a_j) d(a_j, a_k) \leq \Delta\Big\}. \tag{3.24}$$

Note that $P\{\boldsymbol{v}\}$ and $|T^{W}_{\boldsymbol{v}|\boldsymbol{u}}|$ are evaluated as follows [3, 5]:

$$P\{\boldsymbol{v}\} = 2^{-n[H(\boldsymbol{q_v})+D(\boldsymbol{q_v}||\boldsymbol{p})]}, \tag{3.25}$$

$$2^{nH(W_{\boldsymbol{v}|\boldsymbol{u}}|\boldsymbol{q_u})-o(n)} \leq |T^{W}_{\boldsymbol{v}|\boldsymbol{u}}| \leq 2^{nH(W_{\boldsymbol{v}|\boldsymbol{u}}|\boldsymbol{q_u})}, \tag{3.26}$$

where $H(W|\boldsymbol{p})$ denotes the conditional entropy of base 2. Since (3.23) is a sum with respect to conditional types, it becomes smaller by picking a conditional type $W_{\boldsymbol{v}|\boldsymbol{u}}$ satisfying $W_{\boldsymbol{v}|\boldsymbol{u}} \approx W^*$ and discarding other conditional types. Note that $\boldsymbol{q_u} \to \boldsymbol{p}$ and $\boldsymbol{q_v} \to \boldsymbol{p}^*$ as n tends to infinity since $\boldsymbol{u} \in T^n_{[p]}$. Then, (3.23), (3.25) and (3.26) yield

$$\begin{aligned} P(B_n(\boldsymbol{u}, \Delta)) &\geq 2^{-n[H(\boldsymbol{q_v})+D(\boldsymbol{q_v}||\boldsymbol{p})]} 2^{nH(W_{\boldsymbol{v}|\boldsymbol{u}}|\boldsymbol{q_u})-o(n)} \\ &= 2^{-n[H(\boldsymbol{p}^*)+D(\boldsymbol{p}^*||\boldsymbol{p})]} \cdot 2^{nH(W^*|\boldsymbol{p})-o(n)} \\ &= 2^{-n[I(\boldsymbol{p};W^*)+D(\boldsymbol{p}^*||\boldsymbol{p})]-o(n)} \\ &= 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{p})]-o(n)}, \end{aligned} \tag{3.27}$$

which establishes (3.21). □

Note that $D(\boldsymbol{p}^*||\boldsymbol{p}) = 0$ if $\boldsymbol{p}$ is uniform and symmetric single-letter fidelity criterion are used. Comparison of three bounds $R(\boldsymbol{p}, \Delta)$, $V(\boldsymbol{p}, \Delta)$ and $R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^*||\boldsymbol{p})$ for non-uniform binary memoryless source of alphabet $\mathcal{A} = \{0, 1\}$ with probability distribution $\boldsymbol{p} = (0.9, 0.1)$ is given in Fig. 3.1. Figure 3.2 shows the comparison for the source with probability distribution $\boldsymbol{p} = (0.7, 0.3)$. The Hamming distance is used as a single-letter fidelity criterion, otherwise $V(\boldsymbol{p}, \Delta)$ is meaningless. In the two figures the solid line, the broken line and the dotted line correspond to $R(\boldsymbol{p}, \Delta)$, $V(\boldsymbol{p}, \Delta)$ and $R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^*||\boldsymbol{p})$, respectively. In these cases, probability distribution $\boldsymbol{p}^*$ is calculated as $p^*(0) = (p(0) - \Delta)/(1 - 2\Delta)$ and $p^*(1) = (p(1) - \Delta)/(1 - 2\Delta)$ for each distortion level [5]. Generally speaking, $\boldsymbol{p}^*$ is approximately the same as $\boldsymbol{p}$ for small Δ and accordingly $D(\boldsymbol{p}^*||\boldsymbol{p})$ becomes negligible for any discrete memoryless sources as $\Delta \to 0$. In fact, in the settings of Fig. 3.1 and Fig. 3.2 $D(\boldsymbol{p}^*||\boldsymbol{p}) = O(\Delta)$ while $V(\boldsymbol{p}, \Delta) - R(\boldsymbol{p}, \Delta) = O(\Delta \log_2 \frac{1}{\Delta})$ as $\Delta \to 0$. In Fig. 3.1 and Fig. 3.2 $R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^*||\boldsymbol{p})$

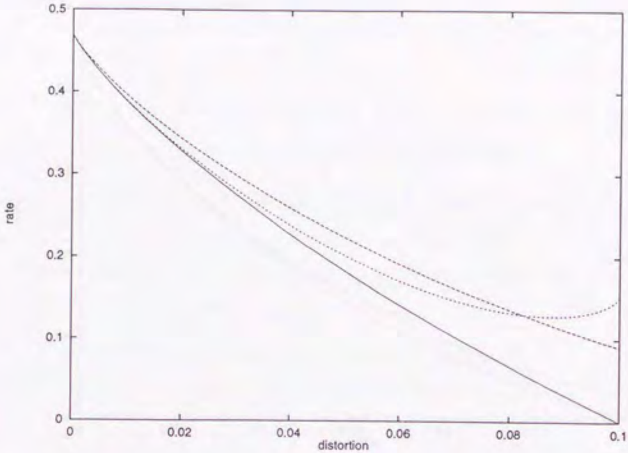


Figure 3.1 Comparison of Three Bounds for $\boldsymbol{p} = (0.9, 0.1)$

gives a tighter upper-bound *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ than $V(\boldsymbol{p}, \Delta)$ for all $\Delta \in (0, 0.082)$ and $\Delta \in (0, 0.135)$. However, these figures indicate that the divergence term becomes large as $\Delta \to D_{max}$, respectively.

The upper-bound given in Theorem 3.1 is not coincident with the rate-distortion function itself. How can the divergence term be eliminated? This problem can be solved by introducing a new probability measure $\hat{P}$ defined by

$$\hat{P} = P^*_{(-\infty,0]} \times P_{[1,\infty)}, \tag{3.28}$$

where P^* denotes the probability measure on $\boldsymbol{X}$ induced by $\boldsymbol{p}^*$ and $P_{[i,j]}$ and $P^*_{[i,j]}$ denote the restriction of P and P^* to the string $X_i X_{i+1} \cdots X_j$, respectively. If P is replaced by $\hat{P}$, not only the divergence term is eliminated but also the convergence *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ to the rate-distortion bound is obtained.

Theorem 3.2 *Let a single-letter fidelity criterion d satisfying* (3.6) *and a distortion level $\Delta \in (0, D_{max})$ is arbitrarily given. Then, for any $\delta > 0, \varepsilon \in (0,1)$ and probability*

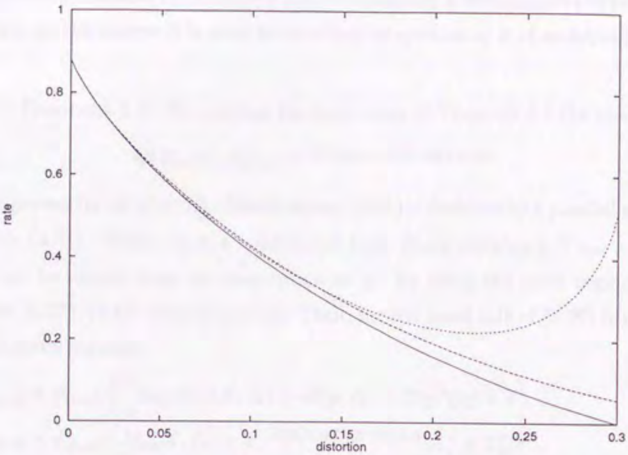


Figure 3.2 Comparison of Three Bounds for $\boldsymbol{p} = (0.7, 0.3)$

distribution $\boldsymbol{q} = (q(a_1), q(a_2), \ldots, q(a_J))$ *satisfying* $q(a_j) > 0$ *for all* $j = 1, 2, \ldots, J$ *there exists an integer* $n_0 = n_0(\delta, \varepsilon)$ *that satisfies*

$$Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) > R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^* \| \boldsymbol{q}) + \delta \right\} < \varepsilon \tag{3.29}$$

for all integers $n > n_0$, *where* Q *denotes the probability measure induced by* $\boldsymbol{q}$.

In particular, if $\boldsymbol{q}$ *is equal to* $\boldsymbol{p}^*$, *then* $\frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta)$ *converges in probability to* $R(\boldsymbol{p}, \Delta)$ *under the probability measure* $\hat{P} = P^*_{(-\infty,0]} \times P_{[1,\infty)}$, *that is, for any* $\delta > 0$ *and* $\varepsilon \in (0, 1)$ *there exists an integer* $n_0' = n_0'(\delta, \varepsilon)$ *satisfying*

$$\hat{P} \left\{ \left| \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) - R(\boldsymbol{p}, \Delta) \right| \geq \delta \right\} < \varepsilon \tag{3.30}$$

for all $n > n_0'$, *where* P^* *denotes the probability measure induced by* $\boldsymbol{p}^*$.

Remark: Though in Theorem 3.2 $\boldsymbol{q}$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ is assumed, the assumption can be weakened as $q(a_j) > 0$ for all j satisfying $p^*(a_j) > 0$. However, $\boldsymbol{q}$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ is more natural for the universal encoding scheme described in Section 3.2.3. Without any knowledge on the probability distribution on the sources it is quite hard to find all symbols $a_j \in \mathcal{A}$ satisfying $p^*(a_j) > 0$.

Proof of Theorem 3.2: For proving the first claim of Theorem 3.2 the inequality

$$Q(B_n(x_1^n, \Delta)) \geq 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{q})]-o(n)} \tag{3.31}$$

should be proved for all $x_1^n \in T_{[p]}^n$. Nevertheless, (3.31) is deduced by a parallel argument that yields (3.21). Notice that a conditional type $W_{\boldsymbol{v}|\boldsymbol{u}}$ satisfying $W_{\boldsymbol{v}|\boldsymbol{u}} \to W^*$ as $n \to \infty$ can be chosen from the assumption on $\boldsymbol{q}$. By using the same argument that establishes (3.27), (3.31) is easily proved. Then, the left hand side of (3.29) is evaluated in the following manner:

$$\begin{aligned}
&Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) \geq R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^*||\boldsymbol{q}) + \delta \right\} \\
&\leq Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{ M_n(\boldsymbol{X}, \Delta) \geq 2^{n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{q})+\delta]} | X_1^n \in T_{[p]}^n \right\} \\
&\qquad + Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{ X_1^n \notin T_{[p]}^n \right\} \\
&\stackrel{1)}{\leq} E_{Q\times P}[M_n(\boldsymbol{X}, \Delta) | X_1^n \in T_{[p]}^n] \cdot 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{q})+\delta]} + Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{ X_1^n \notin T_{[p]}^n \right\} \\
&\stackrel{2)}{=} [Q(B_n(\boldsymbol{u}, \Delta))]^{-1} \cdot 2^{-n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*||\boldsymbol{q})+\delta]} + P \left\{ X_1^n \notin T_{[p]}^n \right\} \\
&\stackrel{3)}{\leq} 2^{-n\delta+o(n)} + P \left\{ X_1^n \notin T_{[p]}^n \right\},
\end{aligned} \tag{3.32}$$

where $\boldsymbol{u} = X_1^n \in T_{[p]}^n$, $E_{Q\times P}$ denotes the expectation with respect to the probability measure $Q_{(-\infty,0]} \times P_{[1,\infty)}$, the Markov inequality is used to deduce inequality 1) in (3.32), $\sum_{l=1}^{\infty} l(1-p)^{l-1}p = 1/p$ for any $p \in (0,1)$ and definition of $Q_{(-\infty,0]} \times P_{[1,\infty)}$ imply equality 2), (3.31) leads to inequality 3). Since δ is fixed and positive, the first term in the left hand side of (3.32) tends to 0 as n tends to infinity. From the property of $T_{[p]}^n$, the second term in (3.32) also tends to 0 as n tend to infinity. This guarantees the existence of an integer $n_0 = n_0(\delta, \varepsilon)$ satisfying (3.29) for all $n > n_0$, which establishes the first claim of the theorem.

Now, convergence *in probability* of $\frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta)$ under the probability measure $\hat{P}$ is proved. For obtaining the convergence, the existence of an integer $n_0' = n_0'(\delta, \varepsilon)$ satisfying the following two inequalities for all $n > n_0'$:

$$\hat{P} \left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) \geq R(\boldsymbol{p}, \Delta) + \delta \right\} < \varepsilon, \tag{3.33}$$

$$\hat{P}\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \le R(\boldsymbol{p},\Delta)-\delta\right\} < \varepsilon. \tag{3.34}$$

Existence of an integer n_1 satisfying (3.33) for all $n > n_1$ is guaranteed by the first claim of the theorem by setting $\boldsymbol{q} = \boldsymbol{p}^*$. Therefore, for completing the proof it is sufficient to show the existence of an integer n_2 satisfying (3.34) for all $n > n_2$ and set $n_0' = \max\{n_1, n_2\}$.

Before establishing the existence of n_2, the inequality

$$P^*(B_n(\boldsymbol{u},\Delta)) \le 2^{-nR(\boldsymbol{p},\Delta)+o(n)} \tag{3.35}$$

for any $\boldsymbol{u} \in T^n_{[p]}$ should be proved. The right hand side of (3.35) is evaluated in the following way:

$$\begin{aligned} P^*(B_n(\boldsymbol{u},\Delta)) &= \sum_{W_{\boldsymbol{v}|\boldsymbol{u}}\in\hat{\mathcal{W}}(\boldsymbol{q}_{\boldsymbol{u}},\Delta)} P^*\{\boldsymbol{v}\}|T^W_{\boldsymbol{v}|\boldsymbol{u}}| \\ &\overset{4)}{\le} \sum_{W_{\boldsymbol{v}|\boldsymbol{u}}\in\hat{\mathcal{W}}(\boldsymbol{q}_{\boldsymbol{u}},\Delta)} 2^{-n[H(\boldsymbol{q}_{\boldsymbol{v}})+D(\boldsymbol{q}_{\boldsymbol{v}}\|\boldsymbol{p}^*)]}\cdot 2^{nH(W_{\boldsymbol{v}|\boldsymbol{u}}|\boldsymbol{q}_{\boldsymbol{u}})} \\ &\overset{5)}{\le} (n+1)^{J^2} 2^{-n[\min_{W_{\boldsymbol{v}|\boldsymbol{u}}\in\hat{\mathcal{W}}(\boldsymbol{q}_{\boldsymbol{u}},\Delta)}\{I(\boldsymbol{q}_{\boldsymbol{u}};W_{\boldsymbol{v}|\boldsymbol{u}})+D(\boldsymbol{q}_{\boldsymbol{v}}\|\boldsymbol{p}^*)\}]} \\ &\overset{6)}{\le} 2^{-nR(\boldsymbol{p},\Delta)+o(n)}, \end{aligned} \tag{3.36}$$

where (3.25) and (3.26) yield inequality 4), inequality 5) follows from the fact that the number of conditional types is at most $(n+1)^{J^2}$, and non-negativity of the divergence and $\boldsymbol{q}_{\boldsymbol{u}} \to \boldsymbol{p}\,(n\to\infty)$ implies inequality 6).

Then, the left hand side of (3.34) is evaluated in the following manner:

$$\begin{aligned} &\hat{P}\left\{\frac{1}{n}\log_2 M_n(\boldsymbol{X},\Delta) \le R(\boldsymbol{p},\Delta)-\delta\right\} \\ &< P^*\{M_n(\boldsymbol{X},\Delta) < 2^{n[R(\boldsymbol{p},\Delta)-\delta]}\,|\,X_1^n \in T^n_{[p]}\} + P\{X_1^n \notin T^n_{[p]}\} \\ &\overset{7)}{=} \sum_{k=1}^{\lfloor 2^{n[R(\boldsymbol{p},\Delta)-\delta]}\rfloor} [1-P^*(B_n(\boldsymbol{u},\Delta))]^{k-1}P^*(B_n(\boldsymbol{u},\Delta)) + P\{X_1^n \notin T^n_{[p]}\} \\ &= 1-[1-P^*(B_n(\boldsymbol{u},\Delta))]^{\lfloor 2^{n[R(\boldsymbol{p},\Delta)-\delta]}\rfloor} + P\{X_1^n \notin T^n_{[p]}\} \\ &\overset{8)}{\le} 1-\exp[-2^{-n\delta-o(n)}] + P\{X_1^n \notin T^n_{[p]}\}, \end{aligned} \tag{3.37}$$

where $\boldsymbol{u} = X_1^n \in T^n_{[p]}$, equality 7) in (3.37) follows from the definition of $\hat{P}$, inequality 8) follows from (3.35) and inequality $[1-t]^m \ge \exp[-\frac{mt}{1-t}]$ for any $t \in (0,1)$ and $m > 0$.

Since δ is fixed and positive, (3.37) means the existence of an integer $n_2 = n_2(\delta, \varepsilon)$ satisfying (3.34) for all $n > n_2$. Setting $n'_0 = \max\{n_1, n_2\}$ completes the proof of the second claim of the theorem. □

Theorem 3.1 and the first statement of Theorem 3.2 claim that convergence *in probability* of $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ in the form of (3.13) dose not generally hold for discrete memoryless sources; a gap between a "typical" sequence emitted from the source and an "optimal" sequence suitable for data compression with fidelity criterion prevents $\frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta)$ from converging *in probability* on its intrinsic probability measure. Finding the optimal sequences in a universal way is discussed in the subsequent chapter.

3.2.3 Encoding Scheme

A practical implementation of a data compression scheme with fidelity criterion is discussed in this section. Suppose that an arbitrary double-infinite sequence $\boldsymbol{x} = \{x_k\}_{k=-\infty}^{\infty}$ is given. If both an encoder and a decoder have a buffer of length L satisfying $L > n \cdot 2^{n[R(\boldsymbol{p},\Delta)+D(\boldsymbol{p}^*\|\boldsymbol{p})]}$ whose contents are x_{-L+1}^{0}, then with probability close to 1 x_1^n is transmitted within distortion Δ by transmitting $M_n(\boldsymbol{x}, \Delta)$. Blocks $x_{nk+1}^{n(k+1)}$ for $k = 1, 2, \ldots$ can be transmitted in the same way. After transmitting each block, there is no need to substitute the oldest n symbols in the two buffers for the latest block since the decoder cannot know x_1^n exactly. In this point, this encoding scheme is different from the Lempel-Ziv algorithm [7] in lossless data compression. The buffers of the encoder and the decoder should be the same at any time instants. Not only for x_1^n but also $x_{nk+1}^{n(k+1)}$ for $k = 1, 2, \ldots$ Theorem 3.1 still makes sense as long as the source is stationary and memoryless.

If another i.i.d sequence $\boldsymbol{y} = \{y_k\}_{k=1}^{L}$ of probability distribution $\boldsymbol{q}$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ is available to both the encoder and the decoder, it can substitute for x_{-L+1}^{0}. In the case that such a finite sequence $\boldsymbol{y}$ is used, the encoder transmits $M'_n(\boldsymbol{x}, \boldsymbol{y}, \Delta)$ to the decoder in a fixed-length binary form, where

$$M'_n(\boldsymbol{x}, \boldsymbol{y}, \Delta) = \min\{M \geq 0 \,:\, y_{nM+1}^{n(M+1)} \in B_n(x_1^n, \Delta)\}, \tag{3.38}$$

if such $y_{nM+1}^{n(M+1)}$ exists. Otherwise, $M'_n(\boldsymbol{x}, \boldsymbol{y}, \Delta)$ is defined as a pre-specified default value less than L/n. The encoder sends $\lceil \log_2(L/n) \rceil$ bits for transmitting $x_{nk+1}^{n(k+1)}$. If

$q = p^*$, Theorem 3.2 indicates that the length $L \approx n \cdot 2^{nR(p,\Delta)}$ of the buffer is enough for sufficiently large n. Hence, for any $\delta > 0$ rate R to transmit x_1^n satisfies

$$R < R(p, \Delta) + \delta \tag{3.39}$$

bits per source symbol if n is sufficiently large. On the other hands, for any $\varepsilon > 0$ average distortion $\bar{\Delta}$ of this encoding scheme is evaluated in the following manner:

$$\begin{aligned}\bar{\Delta} &< \Delta \cdot \Pr\left\{\frac{1}{n}\log_2 M_n'(X, Y, \Delta) < R(p, \Delta) + \delta\right\} \\ &\qquad + d_{max} \cdot \Pr\left\{\frac{1}{n}\log_2 M_n'(X, Y, \Delta) \geq R(p, \Delta) + \delta\right\} \\ &< \Delta + \varepsilon,\end{aligned} \tag{3.40}$$

where $d_{max} \stackrel{\text{def}}{=} \max_{1 \leq j,k \leq J} d(a_j, a_k) < \infty$ and Theorem 3.2 guarantees the last inequality for sufficiently large n. Since δ and ε are arbitrary, (3.39) and (3.40) claim that this encoding scheme achieves asymptotically the rate-distortion bound. This argument provides a simple proof on the direct part of the source coding theorem.

For any probability distribution p and q on $\mathcal{A}$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ and $\delta > 0$, Theorem 3.2 claims that $M_n(X, \Delta)$ satisfies

$$\lim_{n\to\infty} Q_{(-\infty,0]} \times P_{[1,\infty)} \left\{\frac{1}{n}\log_2 M_n(X, \Delta) > R(p, \Delta) + D(p^*\|q) + \delta\right\} = 0, \tag{3.41}$$

where Q denotes the probability measure induced by q. In this case, L satisfying $\frac{1}{n}\log_2 \frac{L}{n} > R(p, \Delta) + D(p^*\|q)$ should be chosen in order to make the average distortion close to Δ.

If $y = \{y_k\}_{k=1}^L$ is used instead of x_{-L+1}^0, both the encoder and the decoder need not have physical buffers. This property is one of the merits of the encoding scheme proposed here, though exhaust search for $M_n'(x, y, \Delta)$ is still inevitable. The encoder and the decoder only need to share a generation algorithm of $\{y_k\}_{k=1}^L$. In practice, y_k, $k = 1, 2, \ldots, L$, are obtained deterministically from pseudo-random numbers uniformly distributed in the unit interval $[0, 1]$ generated by, for example, a linear congruential method or sampling of maximum-length linearly recurring sequences. They use the same algorithm of pseudo-random generation with the same initial parameters and transformation for obtain y_k in order to share the same $\{y_k\}_{k=1}^L$.

3.3 Data Compression for Memoryless Gaussian Sources

In this section a sufficient condition for satisfying (3.13) is given for memoryless gaussian sources under the squared-error criterion. As is seen in the preceding section, (3.13) does not hold for discrete memoryless sources. In Section 3.3.1, however, convergence in the form of (3.13) is shown by modifying the way of matching, that is, the definition of $M_n(\boldsymbol{x}, \Delta)$. This property leads to an asymptotically optimal encoding scheme. Rate-distortion performance of the encoding scheme of finite blocklength is evaluated by computer simulation in Section 3.3.2.

3.3.1 Definitions and Analysis

Let $\boldsymbol{X} = \{X_k\}_{k=-\infty}^{\infty}$ be a sequence of random variables satisfying $X_k \sim N(\mu, \sigma^2)$ for all k from a memoryless gaussian sources, where μ and σ^2 denote the mean and the variance of the source, respectively. Let $\boldsymbol{x} = \{x_k\}_{k=-\infty}^{\infty}$ be a realization of $\boldsymbol{X}$. Since x_k takes real values, source alphabet $\mathcal{A}$ is equal to $\boldsymbol{R}$, the set of real numbers. The squared-error criterion is assumed here, so distortion between $u \in \mathcal{A}$ and $v \in \mathcal{A}$ is defined as $d(u, v) = (u - v)^2$. Distortion between $\boldsymbol{u} = u_1^n \in \mathcal{A}$ and $\boldsymbol{v} = v_1^n$ is defined by (3.7). The rate-distortion function for the source is denoted by $R(p, \Delta)$, which can be written as

$$R(p, \Delta) = \frac{1}{2} \log_2\left(\frac{\sigma^2}{\Delta}\right) \tag{3.42}$$

for $\Delta \in (0, \sigma^2]$ [2], where

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{3.43}$$

denotes the probability density function of the source. Let $M_n(\boldsymbol{x}, \Delta)$ be an integer defined by (3.11). For memoryless gaussian sources of $\mu = 0$, Steinberg and Gutman [26] shows that for any $\delta > 0$

$$\lim_{n\to\infty} P\left\{\frac{1}{n} \log M_n(\boldsymbol{X}, \Delta) > G(\Delta, \sigma) + \delta\right\} = 0, \tag{3.44}$$

where

$$G(\Delta, \sigma) = R(p, \Delta) - \frac{1}{2} \log_2\left(1 - \frac{\Delta}{4\sigma^2}\right). \tag{3.45}$$

Since conditions for convergence in a form of (3.13) are of interest, only memoryless gaussian sources of known mean and variance are considered. Without loss of generality, the memoryless gaussian source of zero-mean and unit-variance can be assumed. For any $\boldsymbol{u} \in \mathcal{A}$ define $B'_n(\boldsymbol{u}, \Delta)$ as

$$B'_n(\boldsymbol{u}, \Delta) = \left\{ \boldsymbol{v} \in \mathcal{A}^n \;\middle|\; \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2} \geq \sqrt{1 - \Delta} \right\} \tag{3.46}$$

and modify $M_n(\boldsymbol{x}, \Delta)$ as

$$M_n(\boldsymbol{x}, \Delta) = \min\{M \geq 0 \,:\, x_{-n(M+1)+1}^{-nM} \in B'_n(x_1^n, \Delta)\}, \tag{3.47}$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \stackrel{\text{def}}{=} \sum_{i=1}^n u_i v_i$ for any $\boldsymbol{u} = u_1^n \in \mathcal{A}^n$ and $\boldsymbol{v} = v_1^n \in \mathcal{A}^n$ and $\|\boldsymbol{u}\|_2 \stackrel{\text{def}}{=} \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}$. Convergence *in probability* in the form of (3.13) is guaranteed by the following theorem.

Theorem 3.3 *For any $\delta > 0$, $\varepsilon \in (0, 1)$ and distortion level $\Delta \in (0, 1)$ there exists an integer $n_0 = n_0(\delta, \varepsilon)$ satisfying*

$$P\left\{ \left| \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) - R(p, \Delta) \right| \geq \delta \right\} < \varepsilon \tag{3.48}$$

for all $n > n_0$, where P denotes the probability measure on $\boldsymbol{X}$ induced by the probability density function of the memoryless gaussian source of zero-mean and unit-variance. □

Proof: This theorem is proved by using similar argument already seen in the proof of Theorem 3.2. For proving the existence of an integer n_0, it is sufficient to show the existence of integers n_1 and n_2 satisfying

$$P\left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) \geq R(p, \Delta) + \delta \right\} < \varepsilon \tag{3.49}$$

for all $n > n_1$ and

$$P\left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) \leq R(p, \Delta) - \delta \right\} < \varepsilon \tag{3.50}$$

for all $n > n_2$. For any $\boldsymbol{u} \in \mathcal{A}^n$ if $P(B'_n(\boldsymbol{u}, \Delta))$ is evaluated as

$$P(B'_n(\boldsymbol{u}, \Delta)) \geq 2^{-nR(p,\Delta) - o(n)} \tag{3.51}$$

and

$$P(B'_n(\boldsymbol{u}, \Delta)) \leq 2^{-nR(p,\Delta) + o(n)}, \tag{3.52}$$

then (3.49) and (3.50) are proved by applying (3.51) and (3.52) in the same way as in obtaining (3.32) and (3.37), respectively.

Define $\tilde{\boldsymbol{u}} = \boldsymbol{u}/\|\boldsymbol{u}\|_2$ for $\boldsymbol{u} \in \boldsymbol{R}^n - \{\mathbf{o}\}$. Since $\tilde{\boldsymbol{u}}$ is uniformly distributed on the n-dimensional unit hypersphere, $P(B'_n(\boldsymbol{u},\Delta))$ can be written in the following form:

$$\begin{aligned} P(B'_n(\boldsymbol{u},\Delta)) = \int_{\boldsymbol{v}\in B'_n(\boldsymbol{u},\Delta)} p^n(\boldsymbol{v})d\boldsymbol{v} &= \frac{1}{|S^{n-1}|}\int_{\langle\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}}\rangle>\sqrt{1-\Delta}} d\tilde{\boldsymbol{v}} \\ &\overset{1)}{=} \frac{|S^{n-2}|}{|S^{n-1}|}\int_0^{\cos^{-1}\sqrt{1-\Delta}} \sin^{n-1}\theta d\theta, \end{aligned} \tag{3.53}$$

where

$$p^n(\boldsymbol{v}) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}(v_1^2+v_2^2+\cdots+v_n^2)\right] \tag{3.54}$$

is the probability density function for $\boldsymbol{v} = v_1^n \in \mathcal{A}^n$, $|S^{n-1}|$ and $|S^{n-2}|$ denote the surface area of n-dimensional and $(n-1)$-dimensional unit hypersphere, respectively and equality 1) follows by transforming the rectangular coordinates to the polar coordinates.

Since $|S^{n-1}| = n\pi^{\frac{n}{2}}/\Gamma(\frac{n}{2}+1)$ for any natural number n, it is easy to verify that

$$\frac{|S^{n-2}|}{|S^{n-1}|} = \begin{cases} \dfrac{1}{\pi}\dfrac{(2m-2)!!}{(2m-3)!!}, & \text{if} \quad n=2m, \\ \dfrac{1}{2}\dfrac{(2m-1)!!}{(2m-2)!!}, & \text{if} \quad n=2m+1, \end{cases} \tag{3.55}$$

and

$$\frac{1}{\pi} \le \frac{|S^{n-2}|}{|S^{n-1}|} \le \frac{n}{2} \tag{3.56}$$

for all n. Hence, (3.53) and (3.56) imply that

$$\begin{aligned} P(B'_n(\boldsymbol{u},\Delta)) &\le \frac{n}{2}\cos^{-1}\sqrt{1-\Delta}\cdot\Delta^{\frac{n-1}{2}} \\ &= 2^{-nR(p,\Delta)+o(n)} \end{aligned} \tag{3.57}$$

and

$$\begin{aligned} P(B'_n(\boldsymbol{u},\Delta)) &\ge \frac{1}{\pi}\int_0^{\cos^{-1}\sqrt{1-\Delta}} \sin^{n-1}\theta\cos\theta d\theta \\ &= \frac{1}{\pi}\cdot\Delta^{\frac{n}{2}} = 2^{-nR(p,\Delta)-o(n)}, \end{aligned} \tag{3.58}$$

which establish (3.51) and (3.52).

The left hands of (3.49) and (3.50) are evaluated directly without considering whether X_1^n is typical or not. Except for this point the remainder of the proof is parallel to the proof of Theorem 3.2 and therefore omitted. □

3.3.2 Simulation Results

The encoding scheme discussed in Section 3.2.3 also makes sense for memoryless gaussian sources. Theorem 3.3 proved in the preceding section implies that the buffer length L should satisfy $L \approx n \cdot 2^{nR(p,\Delta)}$ for sufficiently large n. In this section, rate-distortion performance of this encoding is evaluated by computer simulation.

Two i.i.d. gaussian sequences $\boldsymbol{x} = \{x_k\}_{k=1}^{nK}$ and $\boldsymbol{y} = \{y_k\}_{k=1}^{L}$ of zero-mean and unit-variance are generated by transforming pseudo-random numbers uniformly distributed in the unit interval. Sequence $\{x_k\}_{k=1}^{nK}$ is used as a data sequence. Both the encoder and the decoder are assumed to have buffers of length L containing $\{y_k\}_{k=1}^{L}$. After setting $\Delta = 0.9$ and $L = n \cdot 2^{nR(p,\Delta)}$, $K = 10,000$ source blocks are encoded for each $n = 60, 80, 100, 120, 140, 160$. Since rate-distortion performance of this encoding scheme is of interest, encoding strategy is slightly changed; the encoder finds an integer for $x_{nk+1}^{n(k+1)}$, $k = 0, 1, \ldots, K-1$, defined by

$$M_n(x_{nk+1}^{n(k+1)}, \boldsymbol{y}) = \arg \max_{0 \le M < L/n} \left\langle x_{nk+1}^{n(k+1)}, y_{nM+1}^{n(M+1)} \right\rangle, \tag{3.59}$$

where $\arg \max_{0 \le M < L/n} \left\langle x_{nk+1}^{n(k+1)}, y_{nM+1}^{n(M+1)} \right\rangle$ denotes the argument maximizing the inner-product. Rate required for transmission of $M_n(x_{nk+1}^{n(k+1)}, \boldsymbol{y})$ is $\lceil nR(p,\Delta) \rceil$ bits per source block.

The encoder includes a uniform scalar quantizer to encode the gain of source blocks. The range of the scalar quantizer is the interval $[0, n^{0.55}]$, which means to choose $\alpha = 0.05$ in Section 2.3.1. The number of quantization levels is $\lceil \sqrt{n} \rceil$. Quantization level c_i in the ith interval is the midpoint in the interval. For all $k = 1, 2, \ldots, K$, the scalar quantizer searches an integer i to which $\left\langle x_{nk+1}^{n(k+1)}, y_{nM_n^*+1}^{n(M_n^*+1)} \right\rangle$ belongs, where $M_n^* = M_n(x_{nk+1}^{n(k+1)}, \boldsymbol{y})$. It transmit i with rate $\lceil \frac{1}{2} \log_2 n \rceil$ bits.

In total, $\lceil nR(p,\Delta) \rceil + \lceil \frac{1}{2} \log_2 n \rceil$ bits are required to transmit $x_{nk+1}^{n(k+1)}$ for $k = 0, 1, \ldots, K-1$. Block $x_{nk+1}^{n(k+1)}$ is reproduced as $c_i y_{nM_n^*+1}^{n(M_n^*+1)}$, where c_i denotes the i-th quantization level. By using orthogonality, it is shown that the distortion between

Table 3.1 Performance of the Encoding Scheme

n	$\bar{D}_{shape}$	$V[\bar{D}_{shape}]$	$\bar{D}_{gain}$	$V[\bar{D}_{gain}]$
60	9.34×10^{-1}	2.60×10^{-3}	2.52×10^{-3}	2.12×10^{-5}
80	9.30×10^{-1}	1.03×10^{-3}	1.67×10^{-3}	2.06×10^{-5}
100	9.26×10^{-1}	1.29×10^{-3}	1.57×10^{-3}	1.67×10^{-5}
120	9.23×10^{-1}	9.31×10^{-4}	1.69×10^{-3}	9.31×10^{-6}
140	9.21×10^{-1}	9.86×10^{-4}	9.07×10^{-4}	9.13×10^{-6}
160	9.19×10^{-1}	8.20×10^{-4}	6.65×10^{-4}	9.06×10^{-6}

$x_{nk+1}^{n(k+1)}$ and its reproduced form can be separated as follows:

$$\begin{aligned} & d_n(x_{nk+1}^{n(k+1)}, c_i y_{nM_n^*+1}^{n(M_n^*+1)}) \\ &= \frac{1}{n}\|x_{nk+1}^{n(k+1)}\|_2^2 \left[1 - \left\langle x_{nk+1}^{n(k+1)}, y_{nM_n^*+1}^{n(M_n^*+1)} \right\rangle^2\right] + \frac{1}{n}\left|\left\langle x_{nk+1}^{n(k+1)}, y_{nM_n^*+1}^{n(M_n^*+1)} \right\rangle - c_i\right|^2. \quad (3.60) \end{aligned}$$

This kind of encoding scheme achieves asymptotically the rate distortion bounds if $y_{nM+1}^{n(M+1)}$, $M = 0, 1, \ldots, L/n - 1$ are judiciously chosen as is shown [34] or in Section 2.2 of this thesis.

Denote by D_{shape} and D_{gain} the first and the second terms in the right hand side of (3.60), respectively. Ten different pairs of $\boldsymbol{x} = \{x_k\}_{k=1}^{nK}$ and $\boldsymbol{y} = \{y_k\}_{k=1}^{L}$ are used to evaluate asymptotic behaviors of D_{shape} and D_{gain}. The average and the variance of them are listed in Table 3.3.2. It is obvious that D_{gain} is much smaller than D_{shape} for each n. Asymptotic performance of D_{shape} with respect to the blocklength is seen in Fig. 3.3. Figure 3.3 indicates that D_{shape} monotonously decreases as n increases. The term D_{shape} would converge to $\Delta = 0.9$ as n tends to infinity. Nevertheless, Fig. 3.3 also indicates that n greater than 160 should be chosen in order to make D_{shape} less than 0.919.

3.4 Conclusion

In this chapter a data compression scheme with fidelity criterion is proposed. The scheme is based on string matching, which is originated from Ziv and Lempel [7] in

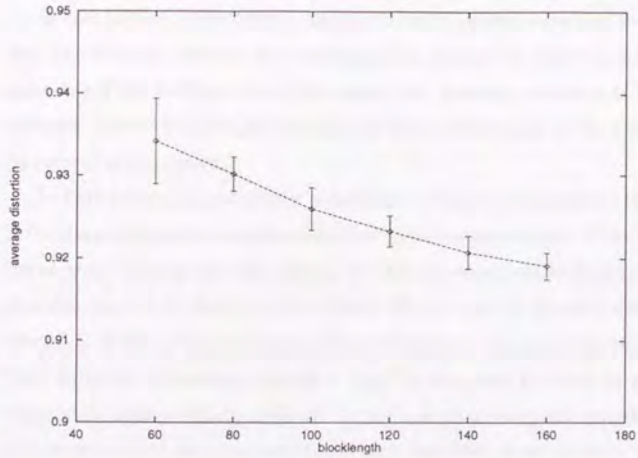


Figure 3.3 Asymptotic Rate-Distortion Performance

lossless case, and Steinberg and Gutman [26] in lossy case. In the proposed scheme both an encoder and a decoder have buffers of finite-length containing the same sequence. For each source block the encoder searches for a content in the buffer within a distortion level and transmits its index. The decoder reproduces the source blocks by the transmitted indices.

Theoretical basis of the algorithm is probabilistic behavior of the indices with increasing the blocklength. Upper-bounds *in probability* of the rate required for transmission of the indices is established in the following two cases: (i) discrete memoryless sources under a single-letter fidelity criterion and (ii) memoryless gaussian sources under the squared-error criterion.

In case (i) for the source of probability distribution $\boldsymbol{p}$ redundancy in rate is equal to $D(\boldsymbol{p}^*||\boldsymbol{q})$ when the two buffers contain an i.i.d. sequence of probability distribution $\boldsymbol{q}$ whose all elements are positive, where $D(\cdot||\cdot)$ denotes the divergence and $\boldsymbol{p}^*$ is the output probability distribution of the test channel. A sufficient condition that the redundancy goes to zero is deduced. The sufficient condition also provides a simple proof of the source coding theorem.

In case (ii) for memoryless gaussian sources of known mean and variance it is shown that the scheme achieves the rate-distortion bound by slight modification of index searching if the buffers contain the same i.i.d. gaussian sequence of known mean and variance. Rate-distortion performance of finite blocklength of the scheme is evaluated by computer simulation.

In both cases the probability distribution of sources is assumed to be known. How is the data compression scheme extended to unknown sources? If the two-stage encoding strategy is supposed, the answer for discrete memoryless sources and memoryless gaussian sources is found in this thesis. In the case of discrete memoryless sources extension of the scheme to compression of unknown sources is discussed in Chapter 4. Two universal estimation algorithm of $\boldsymbol{p}^*$ is proposed in order to make the scheme universally asymptotically optimal. In the case of memoryless gaussian sources of unknown mean and variance, additional task imposed on an encoder is estimating the unknown parameters and transmitting them to a decoder in a quantized form. Asymptotic properties of two-stage encoding of the unknown sources have already discussed in Section 2.4. However, the method that makes the data compression scheme universally asymptotically optimal in a single stage is still an open problem.

Chapter 4

Universal Estimation of the Optimal probability distribution for compression of Discrete Memoryless Source with Fidelity Criterion

4.1 Introduction

The rate-distortion function describes a basic bound of compression efficiency asymptotically achievable by data compression schemes with fidelity criterion. For discrete memoryless sources of finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ it is defined as a minimum of the mutual information as follows:

$$\begin{aligned} R(p, D) &= \min_{W \in \mathcal{W}(p,D)} I(p; W) \\ &= \min_{W \in \mathcal{W}(p,D)} \sum_{k=1}^{J} \sum_{j=1}^{J} p(a_j) W(a_k|a_j) \log_2 \frac{W(a_k|a_j)}{\sum_{l=1}^{J} p(a_l) W(a_j|a_l)}, \end{aligned} \quad (4.1)$$

where $\boldsymbol{p} = (p(a_1), p(a_2), \ldots, p(a_J))$ denotes probability distribution of the sources, $I(\boldsymbol{p}; W)$ denotes the mutual information,

$$\mathcal{W}(\boldsymbol{p}, D) = \Big\{W \text{ is a } J \times J \text{ stochastic matrix} \,|\, \sum_{j=1}^{J}\sum_{k=1}^{J} p(a_j)W(a_k|a_j)d(a_j, a_k) \leq D\Big\}. \tag{4.2}$$

and d is a single-letter fidelity criterion satisfying $d(a_j, a_k) = 0$ if $j = k$ and $0 < d(a_j, a_k) < \infty$ otherwise. The rate-distortion function is positive for all $D \in [0, D_{max})$, where $D_{max} \stackrel{\text{def}}{=} \min_{1 \leq k \leq J} \sum_{j=1}^{J} p(a_j)d(a_j, a_k)$. Denote by R rate of a block code and $\bar{D}$ average distortion caused by the scheme. For arbitrarily distortion level $\Delta \in (0, D_{max})$ converse part of the source coding theorem claims that there is no code satisfying $R < R(\boldsymbol{p}, \Delta)$ and $\bar{D} < \Delta$ and direct part of it guarantees the existence of a block code of sufficiently large blocklength satisfying $R \leq R(\boldsymbol{p}, \Delta) + \delta$ and $\bar{D} < \Delta + \varepsilon$ for any $\delta \geq 0$ and $\varepsilon \geq 0$ satisfying $\delta + \varepsilon > 0$. Proving the direct part is more complicated than showing the converse part. Though there are several ways to establish the direct part [1, 2, 3, 4, 5], using a random code drawn according to the probability distribution

$$p^*(a_k) = \sum_{j=1}^{J} p(a_j)W^*(a_k|a_j) \quad \text{for all} \quad k = 1, 2, \ldots, J \tag{4.3}$$

seems the most popular [2, 5], where W^* is a stochastic matrix achieving the minimum in (4.1). Therefore, knowledge on $\boldsymbol{p}^*$ enables to obtain deeper insight into the asymptotically optimal data compression schemes with fidelity criterion.

In this chapter two universal estimation algorithms of $\boldsymbol{p}^*$ are proposed. If a probability distribution $\boldsymbol{p}$ of discrete memoryless sources and $\frac{d}{dD}R(\boldsymbol{p}, D)|_{D=\Delta}$ are known, Blahut's iteration algorithm [35] for calculating the rate-distortion functions, which has dual relationship to Arimoto's method [36] for computing the capacity of memoryless channels, yields $\boldsymbol{p}^*$ as a byproduct. However, this algorithm is not appropriate for computing $\boldsymbol{p}^*$ in the following two points: (i) the assumption that $\frac{d}{dD}R(\boldsymbol{p}, D)|_{D=\Delta}$ is given is not realistic since $R(\boldsymbol{p}, D)$ is rarely expressed as an explicit function of D and (ii) the algorithm accumulates numerical errors caused by its iterative steps. The two universal estimation algorithms not only cause no numerical error but also require no knowledge of the source. They require two kinds of training sequences and output an estimate of $\boldsymbol{p}^*$ meeting a certain criterion of estimation.

The criterion of estimation discussed in Section 4.2 is deeply related to the data compression scheme proposed in Chapter 3. Let $\boldsymbol{X} = \{X_t\}_{t=-\infty}^{\infty}$ be a sequence of random variables from the source and for a realization $\boldsymbol{x} = \{x_t\}_{t=-\infty}^{\infty}$ of $\boldsymbol{X}$ define

$$M_n(\boldsymbol{x}, \Delta) = \min\{M \geq 0 \,:\, d_n(x_1^n, x_{-n(M+1)+1}^{-nM}) \leq \Delta\}, \tag{4.4}$$

where for any $i \leq j$ $x_i^j = x_i x_{i+1} \cdots x_j$ and d_n denotes the distortion measure between n-tuples defined by a single-letter fidelity criterion d. Then, Theorem 3.2 in Section 3.2.2 guarantees that for any $\zeta > 0$ $M_n(\boldsymbol{X}, \Delta)$ satisfies

$$\lim_{n\to\infty} P'_{(-\infty,0]} \times P_{[1,\infty)} \left\{ \frac{1}{n}\log_2 M_n(\boldsymbol{X}, \Delta) \geq R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^* || \boldsymbol{p}') + \zeta \right\} = 0, \tag{4.5}$$

where P denotes the probability measure on $\boldsymbol{X}$ induced by $\boldsymbol{p}$, P' denotes the probability measure on $\boldsymbol{X}$ induced by a probability distribution $\boldsymbol{p}' = (p'(a_1), p'(a_2), \ldots, p'(a_J))$ satisfying $p'(a_j) > 0$ for all $j = 1, 2, \ldots, J$, for any $i \leq j$ $P_{[i,j]}$ and $P'_{[i,j]}$ means their restriction to $X_i X_{i+1} \cdots X_j$ and $D(\cdot||\cdot)$ denotes the divergence. Equation (4.5) is directly connected to a data compression scheme with fidelity criterion based on the data-base drawn according to $\boldsymbol{p}'$. It implies that rate required by the scheme for making average distortion close to Δ is upper-bounded by $R(\boldsymbol{p}, \Delta) + D(\boldsymbol{p}^*||\boldsymbol{p}')$ per source symbol if n is sufficiently large. That is, the divergence term means a cost in rate of the scheme as $n \to \infty$. From the viewpoint to make the cost small, a good estimate of $\boldsymbol{p}^*$ should be used for generation of the data-base.

The criterion on $\hat{\boldsymbol{p}^*}$ introduced in Section 4.2 is as follows: let $\varepsilon > 0$ and $\delta \in (0, 1)$ be given arbitrarily and suppose that two kinds of training sequences are available. One of the training sequences is drawn from the source to be compressed and the other is drawn from another source that puts out all of symbols in the same alphabet. Universal estimation algorithms must output an estimate of $\boldsymbol{p}^*$, denoted by $\hat{\boldsymbol{p}^*}$, that satisfies

$$Prob\,(D(\boldsymbol{p}^*||\hat{\boldsymbol{p}^*}) > \varepsilon) < \delta, \tag{4.6}$$

for all integers $n > n_0$, where $Prob$ denotes the probability with respect to the two training sequences and n_0 is an integer dependent on ε and δ. The criterion (4.6) implies that with probability at least $1 - \delta$ the algorithm outputs $\hat{\boldsymbol{p}^*}$ satisfying $D(\boldsymbol{p}^*||\hat{\boldsymbol{p}^*}) \leq \varepsilon$.

Then, for any $\zeta > \varepsilon$ $M_n(\boldsymbol{X}, \Delta)$ satisfies

$$\lim_{n \to \infty} \hat{P^*}_{(-\infty,0]} \times P_{[1,\infty)} \left\{ \frac{1}{n} \log_2 M_n(\boldsymbol{X}, \Delta) > R(\boldsymbol{p}, \Delta) + \zeta \right\} = 0, \qquad (4.7)$$

where $\hat{P^*}$ denotes the probability measure on $\boldsymbol{X}$ induced by $\hat{\boldsymbol{p}^*}$. Section 4.2 is devoted to proposal of a universal estimation algorithm of $\boldsymbol{p}^*$ and evaluation of size of the two training sequences as a function of ε and δ required for meeting the criterion (4.6).

The criterion (4.6) is similar to the PAC (*Probably Approximately Correct*) learning models that often appear in the field of the computational learning theory. The PAC learning model first introduced by Valiant [37] is a criterion for learning of deterministic objects such as Boolean functions. Its extensions to learning of stochastic objects are discussed in [38][39][40]. Their theories, however, are not directly applicable to works in Shannon theoretic field. They are only concerned with identification problems of deterministic or stochastic targets in a fixed dimensional space while many results in the Shannon theory make sense with increasing the dimension. The algorithm proposed in Section 4.2 not only proposes a universal estimator of $\boldsymbol{p}^*$ under a feasible criterion but also the first attempt to throw a light on an area in the Shannon theory from a viewpoint of the PAC learning model.

The algorithm proposed in Section 4.2 aggressively permits estimation error in a sense of divergence in order to make sizes of required training sequences small. On the other hand, the criterion introduced in Section 4.3 measures a gap between $\boldsymbol{p}^*$ and $\hat{\boldsymbol{p}^*}$ by l_1-norm. The universal estimation algorithm proposed in the section also requires two kinds of training sequences. One of them is of finite size and drawn from a discrete memoryless source to be compressed. The other is an infinite sequence from another discrete memoryless source putting out all of symbols in the same alphabet. It is shown that the algorithm outputs $\hat{\boldsymbol{p}^*}$ arbitrarily close to $\boldsymbol{p}^*$ with high probability and stops *with probability one* if appropriate parameters are chosen in the algorithm.

4.2 A Universal Estimation Algorithm

In Section 4.2.1 a universal estimation algorithm of $\boldsymbol{p}^*$ is proposed after strict formulation of the estimation problem. Section 4.2.2 is devoted to comparison of this work with two conventional PAC learning models. Sizes of training sequences that are required by the algorithm for meeting the criterion (4.6) is evaluated in Section 4.2.3. If two kinds of training sequences of sufficient length are given, it is proved that the algorithm outputs $\hat{p^*}$ satisfying $D(p^*||\hat{p^*}) \leq \varepsilon$ with probability at least $1-\delta$.

4.2.1 Definition of the Algorithm

Throughout this chapter, only discrete memoryless sources with finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots . a_J\}$ are considered. Probability distribution of the source is denoted by $\boldsymbol{p} = (p(a_1), p(a_2), \ldots, p(a_J))$. Without loss of generality, $p(a_j) > 0$ for all $j = 1, 2, \ldots, J$ can be assumed. A single-letter fidelity criterion d is defined on $\mathcal{A} \times \mathcal{A}$ satisfying the following conditions:

1) $d(a_j, a_k) \geq 0$ for all $j, k = 1, 2, \ldots, J$,

2) For all $a_j \in \mathcal{A}$ there exists $a_k \in \mathcal{A}$ satisfying $d(a_j, a_k) = 0$,

3) $\max_{1 \leq j,k \leq J} d(a_j, a_k) < \infty$.

Symmetry on d, i.e., $d(a_j, a_k) = d(a_k, a_j)$, is not assumed here. Let n be a positive integer arbitrarily fixed. Distortion between $\boldsymbol{x} = x_1 x_2 \cdots x_n \in \mathcal{A}^n$ and $\boldsymbol{y} = y_1 y_2 \cdots y_n \in \mathcal{A}^n$ is defined by

$$d_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i). \tag{4.8}$$

The type of $\boldsymbol{x} \in \mathcal{A}^n$ is denoted by $\boldsymbol{t}(\boldsymbol{x})$, which is the empirical distribution on $\mathcal{A}$ defined by frequency of symbols in $\boldsymbol{x}$.

The rate-distortion function for the discrete memoryless sources with probability distribution $\boldsymbol{p}$ is defined by (4.1). It takes positive values for all $D \in [0, D_{max})$, where

$$D_{max} = \min_{1 \leq k \leq J} \sum_{j=1}^{J} p(a_j) d(a_j, a_k) \tag{4.9}$$

[2, 4, 5]. For an arbitrarily fixed $\Delta \in (0, D_{max})$ denote by W^* the $J \times J$ stochastic matrix that achieves the minimum in (4.1) and define $\boldsymbol{p}^* = (p^*(a_1), p^*(a_2), \ldots, p^*(a_J))$ by

$$p^*(a_k) = \sum_{j=1}^{J} p(a_j) W^*(a_k | a_j) \tag{4.10}$$

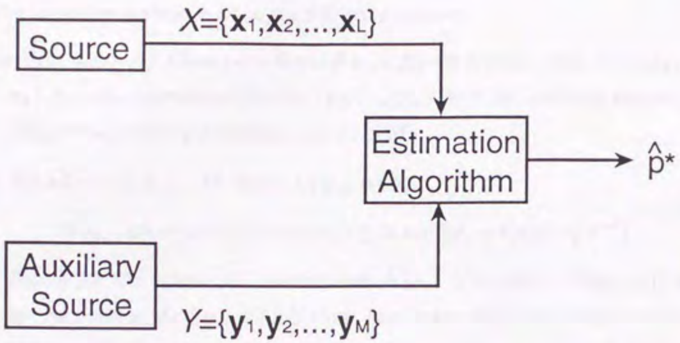


Figure 4.1 Block Diagram of the Universal Estimation System I

for all $k = 1, 2, \ldots, J$. It is assumed that $\boldsymbol{p}^*$ is unique.

Main objective of this section is developing a universal estimation algorithm of $\boldsymbol{p}^*$. An output of the algorithm is denoted by $\hat{\boldsymbol{p}^*}$. Suppose that another discrete memoryless source with the same alphabet and probability distribution $\boldsymbol{q} = (q(a_1), q(a_2), \ldots, q(a_J))$ is available to the estimation algorithm. It is assumed that $q(a_k) > 0$ for all $k = 1, 2, \ldots, J$. This source is called an *auxiliary source.* Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ are L n-tuples drawn independently from the source and $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ are M n-tuples drawn from the auxiliary source. Two sets $\mathcal{X}$ and $\mathcal{Y}$ are two kinds of training sequences available to the estimation algorithm. Figure 4.1 shows a block diagram of the estimation system.

It is assumed that the estimation algorithm can use an estimate of $\boldsymbol{p}$, denoted by $\boldsymbol{p}_e$, satisfying $\|\boldsymbol{p} - \boldsymbol{p}_e\|_1 = O(n^{-\beta_e})$, where

$$\|\boldsymbol{p} - \boldsymbol{p}_e\|_1 = \sum_{j=1}^{J} |p(a_j) - p_e(a_j)|, \tag{4.11}$$

$\beta_e \in (0, \frac{1}{2}]$ is a constant arbitrarily fixed and $O(n^{-\beta_e})$ is interpreted in a sense that $\lim_{n\to\infty} \|\boldsymbol{p} - \boldsymbol{p}_e\|_1 / n^{-\beta_e}$ is finite. This assumption is not so strong. Type $\boldsymbol{t}(\boldsymbol{x}_1)$ can be used instead of $\boldsymbol{p}_e$ since, roughly speaking, it is of precision of $O(n^{-\frac{1}{2}})$; expectation of each element of the type is equal to $\boldsymbol{p}$ and its variance is $O(n^{-\frac{1}{2}})$.

The algorithm estimates $\boldsymbol{p}^*$ in the following manner:

Algorithm 4.1 *1) Choose $\alpha > 0$ and $\beta \in (0, \beta_e)$ arbitrarily. Draw $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ from the source and $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ from the auxiliary source. Fix an integer m_0 arbitrary satisfying $1 \le m_0 \le M$.*

2) For all $m = 1, 2, \ldots, M$ define $\mathcal{N}(\boldsymbol{y}_m, \Delta)$ by

$$\mathcal{N}(\boldsymbol{y}_m, \Delta) = \{\boldsymbol{x} \in \mathcal{X} \mid d_n(\boldsymbol{x}, \boldsymbol{y}_m) \le \Delta \text{ and } \|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 \le n^{-\beta}\}. \tag{4.12}$$

Search for the integer m^* maximizing $|\mathcal{N}(\boldsymbol{y}_m, \Delta)|$, where $|\mathcal{N}(\boldsymbol{y}_m, \Delta)|$ denotes the cardinality of $\mathcal{N}(\boldsymbol{y}_m, \Delta)$. If there exist more than one integers maximizing $|\mathcal{N}(\boldsymbol{y}_m, \Delta)|$, choose one of them arbitrarily and define it as m^*. Define

$$N_{max} = |\mathcal{N}(\boldsymbol{y}_{m^*}, \Delta)|. \tag{4.13}$$

3) If $N_{max} \ge n^\alpha$, output $\boldsymbol{t}(\boldsymbol{y}_{m^*})$. Otherwise, output $\boldsymbol{t}(\boldsymbol{y}_{m_0})$. □

Notice that this algorithm does not use any knowledge on $\boldsymbol{p}$ as well as $\boldsymbol{q}$. For each $m = 1, 2, \ldots, M$ the algorithm counts the number of *typical* elements of $\mathcal{X}$ satisfying $d_n(\boldsymbol{x}_l, \boldsymbol{y}_m) \le \Delta$. After that, it searches for $\boldsymbol{y}_{m^*} \in \mathcal{Y}$ that has elements of $\mathcal{X}$ within distortion Δ as many as possible. For a chosen $\alpha > 0$ if $N_{max} \ge n^\alpha$, then the algorithm outputs the type of $\boldsymbol{y}_{m^*}$ as an estimate of $\boldsymbol{p}^*$. Otherwise, it does not recognize the type of $\boldsymbol{y}_{m^*}$ as an estimate of $\boldsymbol{p}^*$ and outputs a default type specified in Step 1).

The criterion imposed on this algorithm is meeting

$$Prob\,[D(\boldsymbol{p}^*\|\hat{\boldsymbol{p}^*}) > \varepsilon] < \delta \tag{4.14}$$

for sufficiently large n, where $Prob$ in (4.14) denotes probability with respect to the joint probability distribution on $\mathcal{X} \times \mathcal{Y}$. A gap between $\boldsymbol{p}^*$ and $\hat{\boldsymbol{p}^*}$ is measured in a form of the divergence. Appropriateness of introducing the divergence in (4.14) has already seen in the preceding section; it means redundancy of a data compression scheme with fidelity criterion. In the following section comparison the criterion (4.14) with the conventional PAC learning models is discussed. In Section 4.2.3 it is proved that Algorithm 4.1 actually outputs $\hat{\boldsymbol{p}^*}$ satisfying (4.14) under proper choices of L, M and n. Readers who have knowledge on the PAC learning models can skip Section 4.2.2.

4.2.2 Comparison with the PAC Learning Models

Though the criterion (4.14) in the form of divergence originated from the cost of a data compression scheme, it is similar to the PAC learning models that often appear in the computational learning theory. This section is devoted to brief survey of two PAC learning models for making difference between the PAC learning models and the criterion (4.6) clear.

The PAC learning model first proposed by Valiant [37] is summarized as follows: Let $\mathcal{F}$ be a class of functions from $\mathcal{R} = \prod_{i=1}^{k} \mathcal{R}_i$ to $\mathcal{S}$, where $\mathcal{R}_i$, $i = 1, 2, \ldots, k$ are discrete and finite sets. Denote by E^N a set of random samples $(\boldsymbol{r}_1, s_1), (\boldsymbol{r}_2, s_2), \ldots, (\boldsymbol{r}_N, s_N)$ satisfying $(\boldsymbol{r}_i, s_i) \in \mathcal{R} \times \mathcal{S}$ and $f^*(\boldsymbol{r}_i) = s_i$ for unknown function f^* for all $i = 1, 2, \ldots, N$. Samples $\boldsymbol{r}_i$, $i = 1, 2, \ldots, N$ are independently drawn according to an identical probability distribution $Q(\boldsymbol{R})$. An algorithm estimates f^* by using a partial order defined in a class $\mathcal{F}$ and outputs $\hat{f}_{[E^N]}$. For arbitrarily fixed $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$ and any $f^* \in \mathcal{F}$ and $Q(\boldsymbol{R})$ the size of random samples N required for meeting the criterion

$$Prob\,[Q\{\boldsymbol{r} \in \mathcal{R} \,|\, f^*(\boldsymbol{r}) \neq \hat{f}_{[E^N]}\} > \varepsilon] < \delta, \tag{4.15}$$

is evaluated, where $Prob$ in (4.15) denotes the probability with respect to the joint probability distribution on $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$. Usually, the sample size N required for meeting the criterion (4.15) is expressed as a function of $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and k.

Note that in Valiant's model s_i is uniquely determined by $\boldsymbol{r}_i$ while $\boldsymbol{r}_i$ is stochastically generated according to the probability distribution $Q(\boldsymbol{R})$. There are several extensions of the PAC model to learning of stochastic objects [38][39][40]. In particular, Yamanishi [40] extends the model to learning of conditional probability distributions. A class $\mathcal{P}$ of the conditional probability distribution $P(S|\boldsymbol{R})$ is the one to be learned by algorithms, where $S \in \mathcal{S}$, $\boldsymbol{R} \in \mathcal{R} = \prod_{i=1}^{k} \mathcal{R}_i$ and both $\mathcal{S}$ and $\mathcal{R}$ are discrete and finite sets. Given random samples $E^N = (\boldsymbol{r}_1, s_1), (\boldsymbol{r}_2, s_2), \ldots, (\boldsymbol{r}_N, s_N)$ generated independently according to a probability distribution $Q(\boldsymbol{R})P^*(S|\boldsymbol{R})$, for any unknown $Q(\boldsymbol{R})$ and $P^*(S|\boldsymbol{R}) \in \mathcal{P}$, arbitrarily fixed $\varepsilon > 0$ and $\delta \in (0, 1)$, by using an MDL-like strategy an algorithm outputs a conditional probability $\hat{P}_{[E^N]}$ meeting the criterion

$$Prob\,[d(P^*, \hat{P}_{[E^N]}) > \varepsilon] < \delta, \tag{4.16}$$

where $Prob$ in (4.16) denotes the probability with respect to samples E^N and d is a non-negative function satisfying $d(P^*, \hat{P}_{[E^N]}) = 0$ if and only if $P^* = \hat{P}_{[E^N]}$. The conditional divergence and the conditional Hellinger distance are candidates of d. Resemblance of P^* to $\hat{P}_{[E^N]}$ is measured by d. Sample sizes required for meeting the criterion (4.16) is also evaluated as a function of $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and k.

The criterion (4.14) resembles (4.16) in a sense that a gap between a target to be learned and an output of algorithms is measured by a non-negative function satisfying the reflexive property. The divergence used in (4.14), however, means a cost in rate as is seen in (4.5). Finding $\hat{p^*}$ that satisfies $D(p^*||\hat{p^*}) \leq \varepsilon$ asymptotically enables to encode a data sequence in rate $R(p, \Delta) + \varepsilon$ per source symbol. This property clarifies a difference between (4.14) and (4.16) since d in (4.16) is only a measure of resemblance of conditional probability distributions.

Under both learning criteria (4.15) and (4.16) required sample sizes are evaluated as a function of $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and k, where k denotes the dimension of $\mathcal{R}$ that is usually fixed and finite. On the other hand, in the estimation problem of p^* defined in the previous section random samples belong to $\mathcal{A}^n$. From the Shannon theoretical viewpoint, estimation of p^* should be treated with increasing n since many interesting results in the Shannon theory such as the channel coding theorem and the source coding theorem make sense for sufficiently large n. This property motivates to introduce another way to evaluate the number of random samples as a function of ε and δ defined in the following section.

4.2.3 Analysis of the Algorithm

After drawing two kinds of random samples $\mathcal{X} = \{x_1, x_2, \ldots, x_L\}$ from the source and $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ from the auxiliary source, Algorithm 4.1 outputs an estimate of p^*. Lower bounds on L and M required by the algorithm for meeting the criterion (4.14) are deduced in this section. Define $R_\mathcal{X}$ and $R_\mathcal{Y}$ as follows:

$$R_\mathcal{X} = \frac{1}{n} \log_2 L, \tag{4.17}$$

$$R_\mathcal{Y} = \frac{1}{n} \log_2 M. \tag{4.18}$$

In the following theorem lower bounds of $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$ that guarantee Algorithm 4.1 meeting the criterion (4.14) are expressed in a form of the mutual information and the divergence, respectively.

Theorem 4.1 *Fix $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfying $q(a_k) > 0$ for all $k = 1, 2, \ldots, K$ arbitrarily. Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ and $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$ be sets of n-tuples drawn independently from the source and the auxiliary source, respectively and define $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$ by* (4.17) *and* (4.18). *Denote by $\hat{\boldsymbol{p}}^*$ an output of Algorithm* 4.1. *Under the assumption of uniqueness of $\boldsymbol{p}^*$, for any fixed $\Delta \in (0, D_{max})$ if the two inequalities*

$$\min_{\boldsymbol{q}': D(\boldsymbol{p}^* \| \boldsymbol{q}') \le \varepsilon} \min_{V \in \mathcal{V}(\boldsymbol{p}, \boldsymbol{q}', \Delta)} I(\boldsymbol{q}'; V) < R_{\mathcal{X}} < R(\boldsymbol{p}, \Delta) \tag{4.19}$$

and

$$R_{\mathcal{Y}} > D(\boldsymbol{q}_\varepsilon \| \boldsymbol{q}) \tag{4.20}$$

are satisfied, then there exists an integer n_0 satisfying that Algorithm 4.1 *outputs $\hat{\boldsymbol{p}}^*$ meeting the criterion* (4.14) *for all $n > n_0$, where $\boldsymbol{p}^*$ is the probability distribution on $\mathcal{A}$ defined by* (4.3), *$I(\boldsymbol{q}'; V)$ denotes the mutual information,*

$$\mathcal{V}(\boldsymbol{p}, \boldsymbol{q}', \Delta) = \Big\{ V \text{ is a } J \times J \text{ stochastic matrix} \,\Big|\, \sum_{j=1}^{J} \sum_{k=1}^{J} q'(a_k) V(a_j|a_k) d(a_j, a_k) \le \Delta,$$
$$\sum_{k=1}^{J} q'(a_k) V(a_j|a_k) = p(a_j) \text{ for all } j = 1, 2, \ldots, J \Big\}, \tag{4.21}$$

and $\boldsymbol{q}_\varepsilon$ is a probability distribution on $\mathcal{A}$ that achieves the minimum in (4.19) *with a stochastic matrix $V \in \mathcal{V}(\boldsymbol{p}, \boldsymbol{q}_\varepsilon, \Delta)$.*

Note that the lower bound of $R_{\mathcal{X}}$ expressed in (4.19) is non-increasing function of ε. Before proving Theorem 4.1, three lemmata are shown in order to make the essential part of the proof of the theorem clear. Though Algorithm 4.1 does not know $\boldsymbol{p}$ itself, it knows $\boldsymbol{p}_e$ as its estimate. The following lemma suggests that a gap between $\boldsymbol{p}$ and $\boldsymbol{p}_e$ is negligible for sufficiently large n.

Lemma 4.1 *Let $\boldsymbol{p}_e$ be an estimate of $\boldsymbol{p}$ satisfying $\|\boldsymbol{p} - \boldsymbol{p}_e\|_1 = O(n^{-\beta_e})$ for an arbitrary $\beta_e \in (0, \frac{1}{2}]$. Fix $\beta \in (0, \beta_e)$ arbitrarily and define*

$$\mathcal{U}_\beta(\boldsymbol{p}_e) = \{\boldsymbol{x} \in \mathcal{A}^n \,|\, \|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 \le n^{-\beta}\}. \tag{4.22}$$

Then, for any β_1 and β_2 satisfying $0 < \beta_1 \le \beta < \beta_2 < \beta_e$ there exists an integer $n_0 = n_0(\beta_1, \beta_2)$ that satisfies

$$\mathcal{K}_{\beta_2}(\boldsymbol{p}) \subset \mathcal{U}_\beta(\boldsymbol{p}_e) \subset \mathcal{K}_{\beta_1}(\boldsymbol{p}) \tag{4.23}$$

for all integers $n > n_0$, where $\mathcal{K}_\nu(\boldsymbol{p})$ is a subset of $\mathcal{A}^n$ defined by

$$\mathcal{K}_\nu(\boldsymbol{p}) = \{\boldsymbol{x} \in \mathcal{A}^n \mid \|\boldsymbol{p} - \boldsymbol{t}(\boldsymbol{x})\|_1 = O(n^{-\nu})\}. \tag{4.24}$$

Proof: Fix β_1 and β_2 arbitrarily satisfying $0 < \beta_1 \le \beta < \beta_2 < \beta_e$. For proving this lemma, it is sufficient to show the existence of integers n_1 and n_2 satisfying $\mathcal{U}_\beta(\boldsymbol{p}_e) \subset \mathcal{K}_{\beta_1}(\boldsymbol{p})$ for all $n > n_1$ and $\mathcal{K}_{\beta_2}(\boldsymbol{p}) \subset \mathcal{U}_\beta(\boldsymbol{p}_e)$ for all $n > n_2$, respectively. Existence of the integer n_1 is established first.

If $\boldsymbol{x} \in \mathcal{U}_\beta(\boldsymbol{p}_e)$, $\|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 \le n^{-\beta}$. Then, the assumption on $\boldsymbol{p}_e$ and the triangle inequality yield

$$\begin{aligned} \|\boldsymbol{p} - \boldsymbol{t}(\boldsymbol{x})\|_1 &\le \|\boldsymbol{p} - \boldsymbol{p}_e\|_1 + \|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 \\ &\le O(n^{-\beta_e}) + n^{-\beta} = O(n^{-\beta}). \end{aligned} \tag{4.25}$$

Since $\beta_1 \le \beta$, i.e., $n^{-\beta} \le n^{-\beta_1}$, (4.25) guarantees the existence of integer n_1.

Next, existence of the integer n_2 is shown. If $\boldsymbol{x} \in \mathcal{K}_{\beta_2}(\boldsymbol{p})$, then $\|\boldsymbol{p} - \boldsymbol{t}(\boldsymbol{x})\|_1 = O(n^{-\beta_2})$. The triangle inequality implies that

$$\begin{aligned} \|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 &\le \|\boldsymbol{p}_e - \boldsymbol{p}\|_1 + \|\boldsymbol{p} - \boldsymbol{t}(\boldsymbol{x})\|_1 \\ &= O(n^{-\beta_e}) + O(n^{-\beta_2}) = O(n^{-\beta_2}). \end{aligned} \tag{4.26}$$

From the assumption on β_2, the term of $O(n^{-\beta_2})$ becomes less than $n^{-\beta}$ for all $\boldsymbol{x} \in \mathcal{K}_{\beta_2}(\boldsymbol{p})$ if n is sufficiently large. This establishes the existence of integer n_2. □

The following lemma characterizes an interesting property of a random variable N drawn according to a binomial distribution $B(L, \theta)$, that is, for all $r = 1, 2, \ldots, L$

$$\Pr(N = r) = \binom{L}{r} \theta^r (1 - \theta)^{L-r}. \tag{4.27}$$

If $L = 2^{nR_L}$ and $\theta = 2^{-nR_\theta}$ for positive numbers R_L and R_θ, expectation of N is equal to $2^{n(R_L - R_\theta)}$. It diverges to infinity of exponential order of n as $n \to \infty$ if $R_L > R_\theta$

and converges to zero of the same order if $R_L < R_\theta$. The following lemma claims on a relation the behavior of expectation to the probability of N satisfying $N < n^\alpha$ or $N \geq n^\alpha$ for arbitrarily chosen $\alpha > 0$.

Lemma 4.2 *Suppose a random variable N drawn according to the binomial distribution $B(L,\theta)$, where $L = 2^{nR_L}$ and $\theta = 2^{-nR_\theta}$ for any fixed positive numbers R_L and R_θ. Fix $\alpha > 0$ arbitrarily.*

If $R_L > R_\theta$, then there exists an integer n_0 satisfying

$$\Pr(N < n^\alpha) \leq \exp[-\exp_2[n(R_L - R_\theta) + o(n)]] \tag{4.28}$$

for all $n > n_0$, where $\exp_2[x] = 2^x$ and $o(n)$ denotes terms of order less than 1.

On the other hand, if $R_L < R_\theta$, then there exists an integer n_0' satisfying

$$\Pr(N \geq n^\alpha) \leq \exp[-n^{\alpha+1}(R_\theta - R_L) \cdot \ln 2 + nR_L \ln 2 + o(n)] \tag{4.29}$$

for all $n > n_0'$, where $\ln$ denotes the natural logarithm.

Proof: Fix $\alpha > 0$ arbitrarily and define

$$C_r = \binom{L}{r}\theta^r(1-\theta)^{L-r} \tag{4.30}$$

for all $r = 0, 1, \ldots, L$. By calculating $\frac{C_r}{C_{r+1}}$, it is easy to verify that C_r takes the maximum value when r is equal to $r^* = \lfloor \theta(L+1) \rfloor$. Notice that $r^* \to \infty$ as $n \to \infty$ of exponential order of n if $R_L > R_\theta$ and $r^* \to 0$ as $n \to \infty$ if $R_L < R_\theta$. Hence, there exists an integer n_0 satisfying $n^\alpha \leq r^*$ for all $n > n_0$. For all $n > n_0$ the left hand side of (4.28) is evaluated in the following way:

$$\begin{aligned}
\Pr(N < n^\alpha) &= \sum_{r=0}^{T} \binom{L}{r}\theta^r(1-\theta)^{L-r} \\
&\overset{1)}{\leq} (T+1)\binom{L}{T}\theta^T(1-\theta)^{L-T} \\
&\overset{2)}{\leq} (T+1)L^T(1-\theta)^{L-T} \\
&\overset{3)}{\leq} \exp[\ln(T+1) + T\ln L - (L-T)\theta] \\
&\overset{4)}{\leq} \exp\left[\ln(n^\alpha+1) + n^{\alpha+1}R_L \ln 2 - (2^{nR_L} - n^\alpha)2^{-nR_\theta}\right] \\
&= \exp\left[-\exp_2[n(R_L - R_\theta) + o(n)]\right],
\end{aligned} \tag{4.31}$$

where $T = \lfloor n^\alpha \rfloor$, the property of r^* leads to the inequality 1) in (4.31), inequality 2) follows from trivial inequalities $\binom{L}{T} \le L^T$ and $\theta^T \le 1$, inequality $\ln(1-\theta) < -\theta$ for $\theta \in (0,1)$ implies inequality 3) and the definitions of T, R_L and R_θ yield inequality 4). This establishes the first claim of this lemma.

In the case that $R_\theta > R_L$ it is obvious that there exists an integer n_0' satisfying $r^* < n^\alpha$ for all $n > n_0'$. For all $n > n_0'$ the left hand side of (4.29) is upper-bounded in the following manner:

$$
\begin{aligned}
\Pr(N \ge n^\alpha) &= \sum_{r=T'}^{L} \binom{L}{r} \theta^r (1-\theta)^{L-r} \\
&\overset{5)}{\le} (L - T' + 1) \binom{L}{T'} \theta^{T'} (1-\theta)^{L-T'} \\
&\overset{6)}{\le} L^{T'+1} \theta^{T'} \\
&\overset{7)}{=} \exp\left[(\lceil n^\alpha \rceil + 1) n R_L \ln 2 - \lceil n^\alpha \rceil n R_\theta \ln 2\right] \\
&= \exp[-n^{\alpha+1}(R_\theta - R_L) \cdot \ln 2 + n R_L \ln 2 + o(n)], \qquad (4.32)
\end{aligned}
$$

where $T' = \lceil n^\alpha \rceil$, the property of r^* implies inequality 5) in (4.32), inequality 6) follows from $(1-\theta)^{L-T'} \le 1$ and the definitions of T', R_L and R_θ yield inequality 7), which establishes the second claim of this lemma. □

For any given $\boldsymbol{y} \in \mathcal{A}^n$ and $\Delta \in (0, D_{max})$ it is important to evaluate the probability that $\boldsymbol{x}$ satisfying $d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta$ is generated from the source. In the following lemma this probability is evaluated with increasing n. The lemma shows that the probability converges to 0 of exponential order of n and describes its exponent in a form of minimum of the mutual information subject to a constraint.

Lemma 4.3 *Let $\Delta \in (0, D_{max})$ be an arbitrary distortion level and fix types $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{q}}$ of $\mathcal{A}^n$ arbitrarily. Then, for any $\boldsymbol{y} \in \mathcal{A}^n$ of type $\tilde{\boldsymbol{q}}$*

$$
\begin{cases}
\Pr(d_n(X^n, \boldsymbol{y}) \le \Delta \,|\, \boldsymbol{t}(X^n) = \tilde{\boldsymbol{p}}) \ge 2^{-n \min_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; \tilde{V}) - o(n)}, \\
\Pr(d_n(X^n, \boldsymbol{y}) \le \Delta \,|\, \boldsymbol{t}(X^n) = \tilde{\boldsymbol{p}}) \le 2^{-n \min_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; \tilde{V}) + o(n)},
\end{cases} \qquad (4.33)
$$

where

$$
\tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta) = \left\{ \tilde{V} \text{ is a } J \times J \text{ conditional type} \,\middle|\, \sum_{j=1}^{J} \sum_{k=1}^{J} \tilde{q}(a_k) \tilde{V}(a_j | a_k) d(a_j, a_k) \le \Delta, \right.
$$

$$\sum_{k=1}^{J} \tilde{q}(a_k)\tilde{V}(a_j|a_k) = \tilde{p}(a_j),\, j = 1, 2, \ldots, J\Big\}. \quad (4.34)$$

Moreover, for any $\nu \in (0, \frac{1}{2}]$ *and* $\boldsymbol{y} \in \mathcal{A}^n$ *of type* $\tilde{\boldsymbol{q}}$

$$\boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_\nu(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\} \le 2^{-n \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; \tilde{V}) + o(n)}, \quad (4.35)$$

where $\boldsymbol{p}^n$ *denotes the probability measure on* $\mathcal{A}^n$ *induced by* $\boldsymbol{p}$*,* $\mathcal{K}_\nu(\boldsymbol{p})$ *is the set defined by* (4.24) *and*

$$\mathcal{V}(\boldsymbol{p}, \tilde{q}, \Delta) = \Big\{\tilde{V} \text{ is a } J \times J \text{ stochastic matrix} \,\Big|\, \sum_{j=1}^{J}\sum_{k=1}^{J} \tilde{q}(a_k)V(a_j|a_k)d(a_j, a_k) \le \Delta$$

$$\sum_{k=1}^{J} \tilde{q}(a_k)V(a_j|a_k) = p(a_j),\, j = 1, 2, \ldots, J\Big\}. \quad (4.36)$$

Proof: Fix $\nu \in (0, \frac{1}{2}]$, types $\tilde{\boldsymbol{p}}, \tilde{q}$ and $\boldsymbol{y} \in \mathcal{A}^n$ satisfying $\boldsymbol{t}(\boldsymbol{y}) = \tilde{q}$ arbitrarily. Inequalities(4.33) is proved first. Since d_n is context-free,

$$\Pr(d_n(X^n, \boldsymbol{y}) \le \Delta \,|\, \boldsymbol{t}(\boldsymbol{y}) = \tilde{q}) = \frac{|\{\boldsymbol{x} \in \mathcal{A}^n \,|\, \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}|}{|\{\boldsymbol{x} \in \mathcal{A}^n \,|\, \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}}\}|}, \quad (4.37)$$

where $|\cdot|$ denotes the cardinality of the set. It is well-known that

$$|\{\boldsymbol{x} \in \mathcal{A}^n \,|\, \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}}\}| = 2^{nH(\tilde{\boldsymbol{p}}) - o(n)}, \quad (4.38)$$

where $H(\cdot)$ is the entropy function of base 2. (See e.g., [3, 5].) The numerator in the right hand side of (4.37) can be written as

$$|\{\boldsymbol{x} \in \mathcal{A}^n \,|\, \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}| = \sum_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} |T_{\tilde{V}}(\boldsymbol{y})|, \quad (4.39)$$

where $T_{\tilde{V}}(\boldsymbol{y})$ is a subset of $\mathcal{A}^n$ defined as

$$T_{\tilde{V}}(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathcal{A}^n \,|\, \text{the conditional type from } \boldsymbol{y} \text{ to } \boldsymbol{x} \text{ is equal to } \tilde{V}\} \quad (4.40)$$

and $|T_{\tilde{V}}(\boldsymbol{y})|$ means its cardinality. It is also well-known that

$$2^{nH(\tilde{V}|\tilde{\boldsymbol{q}}) - o(n)} \le |T_{\tilde{V}}(\boldsymbol{y})| \le 2^{nH(\tilde{V}|\tilde{\boldsymbol{q}})}, \quad (4.41)$$

where $H(\cdot|\cdot)$ denotes the conditional entropy function of base 2 [3, 5]. Note that the number of conditional types belonging to $\tilde{\mathcal{V}}(\tilde{p}, \tilde{q}, \Delta)$ is not greater than $(n+1)^{J^2}$.

Combining (4.39) and (4.41) yields

$$\begin{cases} |\{\boldsymbol{x} \in \mathcal{A}^n \mid \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}| \ge 2^{\max_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} H(\tilde{V}|\tilde{\boldsymbol{q}}) - o(n)} \\ |\{\boldsymbol{x} \in \mathcal{A}^n \mid \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}| \le (n+1)^{J^2} 2^{\max_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} H(\tilde{V}|\tilde{\boldsymbol{q}})}, \end{cases} \tag{4.42}$$

where the lower-bound is deduced by a choice of $\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)$ achieving the maximum of $H(\tilde{V}|\tilde{\boldsymbol{q}})$. Two inequalities in (4.42) directly imply

$$\begin{cases} |\{\boldsymbol{x} \in \mathcal{A}^n \mid \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}| \ge 2^{\max_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} H(\tilde{V}|\tilde{\boldsymbol{q}}) - o(n)} \\ |\{\boldsymbol{x} \in \mathcal{A}^n \mid \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}} \text{ and } d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta\}| \le 2^{\max_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} H(\tilde{V}|\tilde{\boldsymbol{q}}) + o(n)}. \end{cases} \tag{4.43}$$

Substituting (4.38) and (4.43) into (4.37) yields both of two inequalities in (4.33).

Secondly, (4.35) is established. Fix $\boldsymbol{y} \in \mathcal{A}^n$ of type $\tilde{\boldsymbol{q}}$, $\nu \in (0, \frac{1}{2}]$ and type $\tilde{\boldsymbol{p}}$ satisfying $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_1 = O(n^{-\nu})$ arbitrarily. Then,

$$\begin{aligned} & \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \mid d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta \text{ and } \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}}\} && (4.44) \\ & \qquad = \Pr(\boldsymbol{t}(X^n) = \tilde{\boldsymbol{p}}) \cdot \Pr(d_n(X^n, \boldsymbol{y}) \le \Delta \mid \boldsymbol{t}(X^n) = \tilde{\boldsymbol{p}}) \\ & \qquad \le 2^{-nD(\tilde{\boldsymbol{p}}\|\boldsymbol{p}) + o(n)} 2^{-n \min_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; \tilde{V}) + o(n)}, && (4.45) \end{aligned}$$

where $\Pr(\boldsymbol{t}(X^n) = \tilde{\boldsymbol{p}}) \le 2^{-nD(\tilde{\boldsymbol{p}}\|\boldsymbol{p}) + o(n)}$ for any type $\tilde{\boldsymbol{p}}$ [3, 5] and (4.33) imply the inequality. Note that $D(\tilde{\boldsymbol{p}}\|\boldsymbol{p})$ converges to 0 for all $\tilde{\boldsymbol{p}}$ satisfying $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_1 = O(n^{-\nu})$. Moreover,

$$n \min_{\tilde{V} \in \tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; \tilde{V}) = n \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) + o(n) \tag{4.46}$$

for such $\boldsymbol{p}$ since the two sets $\tilde{\mathcal{V}}(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, \Delta)$ and $\mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)$ asymptotically become the same as $n \to \infty$. Hence,

$$\boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \mid d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta \text{ and } \boldsymbol{t}(\boldsymbol{x}) = \tilde{\boldsymbol{p}}\} \le 2^{-n \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) + o(n)} \tag{4.47}$$

for all $\tilde{\boldsymbol{p}}$ satisfying $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_1 = O(n^{-\nu})$. Notice that the right hand side of (4.47) does not depend on $\tilde{\boldsymbol{p}}$. Inequality (4.35) is obtained by adding the both hand side of (4.47) with respect to $\tilde{\boldsymbol{p}}$ and using the fact that the number of such types is not greater than $(n+1)^J$. □

Now, it is ready to prove Theorem 4.1. It is shown that with probability close to one the algorithm selects $\boldsymbol{y} \in \mathcal{Y}$ that has a type in a neighborhood of $\boldsymbol{q}_\varepsilon$. The algorithm causes two kinds of errors similar to the type-I error and the type-II error in statistical tests. By virtue of Lemma 4.2 and Lemma 4.3 both of the two errors goes to zero as $n \to \infty$.

Proof of Theorem 4.1: Choose $\Delta \in (0, D_{max})$ arbitrarily and define q_ε as the probability distribution on $\mathcal{A}$ achieving the minimum of (4.19) with a stochastic matrix $V \in \mathcal{V}(p, q_\varepsilon, \Delta)$. Continuity of the divergence and the mutual information leads to the existence of $\xi > 0$ satisfying

$$R_{\mathcal{X}} > \max_{q' \in \mathcal{Q}_\xi(q_\varepsilon)} \min_{V \in \mathcal{V}(p, q', \Delta)} I(q'; V) \tag{4.48}$$

and

$$R_{\mathcal{Y}} > \max_{q' \in \mathcal{Q}_\xi(q_\varepsilon)} D(q' \| q), \tag{4.49}$$

where

$$\mathcal{Q}_\xi(q_\varepsilon) = \{q' \text{ is a probability distribution on } \mathcal{A} \mid D(p^* \| q') \le \varepsilon \text{ and } \|q_\varepsilon - q'\|_1 \le \xi\}. \tag{4.50}$$

Note that there exists an integer n_1 that satisfies $t(y) \in \mathcal{Q}_\xi(q_\varepsilon)$ for some $y \in \mathcal{A}^n$ for all integers $n > n_1$. The left hand side of (4.14) is evaluated in the following manner:

$$\begin{aligned}
&Prob[D(p^* \| \hat{p^*}) > \varepsilon] = \sum_{\mathcal{Y} \in \mathcal{A}^{nM}} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \\
&\qquad \times \Pr(D(p^* \| \hat{p^*}) > \varepsilon \mid \mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \\
&= \sum_{\mathcal{Y} \notin [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \Pr(D(p^* \| \hat{p^*}) > \varepsilon \mid \mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \\
&\qquad + \sum_{\mathcal{Y} \in [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \Pr(D(p^* \| \hat{p^*}) > \varepsilon \mid \mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \\
&\le \sum_{\mathcal{Y} \notin [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) + \sum_{\mathcal{Y} \in [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) \\
&\qquad \times \Pr(D(p^* \| \hat{p^*}) > \varepsilon \mid \mathcal{Y} = \{y_1, y_2, \ldots, y_M\}),
\end{aligned} \tag{4.51}$$

where $[\mathcal{A}^{nM}]$ is a set of $\mathcal{Y}$ defined by

$$[\mathcal{A}^{nM}] = \{\mathcal{Y} = \{y_1, y_2, \ldots, y_M\} \mid \text{there exists } y \in \mathcal{Y} \text{ satisfying } t(y) \in \mathcal{Q}_\xi(q_\varepsilon)\}. \tag{4.52}$$

The first term in (4.51) means the probability that $\mathcal{Y}$ does not contain any elements belonging to $\mathcal{Q}_\xi(q_\varepsilon)$. Hence, for all $n > n_1$ it is upper-bounded in the following manner:

$$\begin{aligned}\sum_{\mathcal{Y}\notin[\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}) &= \left[1 - \sum_{q'\in\mathcal{Q}_\xi(q_\varepsilon)\cap\mathcal{T}^n_{\mathcal{A}}} \sum_{y\in\mathcal{A}^n(q')} q^n(y)\right]^M \\ &\overset{1)}{=} \left[1 - \sum_{q'\in\mathcal{Q}_\xi(q_\varepsilon)\cap\mathcal{T}^n_{\mathcal{A}}} 2^{-nD(q'\|q)+o(n)}\right]^M \\ &\overset{2)}{\le} \left[1 - 2^{-n\min_{q'\in\mathcal{Q}_\xi(q_\varepsilon)} D(q'\|q)+o(n)}\right]^M \\ &\overset{3)}{\le} \exp\left[-\exp_2\left[n(R_\mathcal{Y} - \min_{q'\in\mathcal{Q}_\xi(q_\varepsilon)} D(q'\|q)) + o(n)\right]\right], \end{aligned} \tag{4.53}$$

where $\mathcal{T}^n_{\mathcal{A}}$ denotes the set of all types of $\mathcal{A}^n$, $\mathcal{A}^n(q')$ is a subset of $\mathcal{A}^n$ whose all elements have type q', q^n is the probability measure on $\mathcal{A}^n$ induced by q, a property on type is used for obtaining equality 1) in (4.53), $|\mathcal{T}^n_{\mathcal{A}}| < (n+1)^J$ and extension of the domain imply inequality 2) and inequality 3) follows from the definition of $R_\mathcal{Y}$ and inequality $(1-t)^m \le \exp[-mt]$ for $t \in (0,1)$ and $m > 0$. Since (4.49) guarantees

$$R_\mathcal{Y} - \min_{q'\in\mathcal{Q}_\xi(q_\varepsilon)} D(q'\|q)) \ge R_\mathcal{Y} - \max_{q'\in\mathcal{Q}_\xi(q_\varepsilon)} D(q'\|q)) > 0, \tag{4.54}$$

the first term in (4.51) converges to 0 of double-exponential order of n.

Hereafter, the second term in (4.51) is evaluated. Note that there are two cases that $D(p^*\|\hat{p^*}) \le \varepsilon$ is not satisfied as follows:

$$\begin{aligned}\text{Case 1} &: N_{max} < n^\alpha \text{ for } y_{m_0} \text{ satisfying } D(p^*\|t(y_{m_0})) > \varepsilon, \\ \text{Case 2} &: N_{max} \ge n^\alpha \text{ for } y_{m^*} \text{ satisfying } D(p^*\|t(y_{m^*})) > \varepsilon.\end{aligned}$$

Therefore, the second term in (4.51) is upper-bounded in the following manner:

$$\begin{aligned}&\Pr(D(p^*\|\hat{p^*}) > \varepsilon \,|\, \mathcal{Y} = \{y_1, y_2, \ldots y_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &= \Pr(N_{max} < n^\alpha \text{ and } D(p^*\|t(y_{m_0})) > \varepsilon \,|\, \mathcal{Y} = \{y_1, y_2, \ldots y_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\quad + \Pr(N_{max} \ge n^\alpha \text{ and } D(p^*\|t(y_{m^*})) > \varepsilon \,|\, \mathcal{Y} = \{y_1, y_2, \ldots y_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\le \Pr(N_{max} < n^\alpha \,|\, \mathcal{Y} = \{y_1, y_2, \ldots y_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\quad + \Pr(N_{max} \ge n^\alpha \text{ and } D(p^*\|t(y_{m^*})) > \varepsilon \,|\, \mathcal{Y} = \{y_1, y_2, \ldots y_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}])\end{aligned} \tag{4.55}$$

It is shown that two probabilities in (4.55) converges to 0 as n tends to infinity as far as $R_{\mathcal{X}}$ satisfies (4.19).

Since $N_{max} = |\mathcal{N}(\boldsymbol{y}_{m^*}, \Delta)| = \max_{1\le m\le M} |\mathcal{N}(\boldsymbol{y}_m, \Delta)|$ from its definition, for given $\mathcal{Y}$ the probability of $N_{max} < n^\alpha$ is less than the one that $|\mathcal{N}(\boldsymbol{y}_m, \Delta)| < n^\alpha$ for any specified $\boldsymbol{y}_{m'} \in \mathcal{Y}$. This property leads to

$$\begin{aligned}&\Pr(N_{max} < n^\alpha \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \boldsymbol{y}_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}])\\ &\quad\le \Pr(|\mathcal{N}(\boldsymbol{y}_{m'}, \Delta)| < n^\alpha \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \boldsymbol{y}_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}]).\end{aligned} \tag{4.56}$$

Notice that $|\mathcal{N}(\boldsymbol{y}_{m'}, \Delta)|$ is governed by the binomial distribution $B(L, \theta)$, where

$$\theta = \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{U}_\beta(\boldsymbol{p}_e) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_{m'}) \le \Delta\} \tag{4.57}$$

and $\boldsymbol{p}^n$ denotes the probability measure on $\mathcal{A}^n$ induced by $\boldsymbol{p}$. Fix β_2 satisfying $\beta < \beta_2 < \beta_e$ and define a new random variable $|\mathcal{N}_L(\boldsymbol{y}_{m'}, \Delta)|$ by

$$|\mathcal{N}_L(\boldsymbol{y}_{m'}, \Delta)| = |\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_{\beta_2}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_{m'}) \le \Delta\}|, \tag{4.58}$$

where $\mathcal{K}_{\beta_2}(\boldsymbol{p})$ is a set defined by (4.24). Lemma 4.1 guarantees the existence of an integer n_2 satisfying $\mathcal{A}^n \cap \mathcal{K}_{\beta_2}(\boldsymbol{p}) \subset \mathcal{A}^n \cap \mathcal{U}_\beta(\boldsymbol{p}_e)$ for all $n > n_2$. Notice that $|\mathcal{N}_L(\boldsymbol{y}_{m'}, \Delta)|$ is also governed by the binomial distribution $B(L, \theta_L)$, where

$$\theta_L = \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_{\beta_2}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_{m'}) \le \Delta\} \tag{4.59}$$

For $n > \max\{n_1, n_2\}$ Lemma 4.3 shows that θ_L can be expressed in the following form:

$$\theta_L = 2^{-n \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) + o(n)} \tag{4.60}$$

for any $\boldsymbol{y}_{m'} \in \mathcal{A}^n(\tilde{\boldsymbol{q}})$, where $\tilde{\boldsymbol{q}} = \boldsymbol{t}(\boldsymbol{y}_{m'})$. From the assumption that $\boldsymbol{y}_{m'} \in \mathcal{Q}_\xi(\boldsymbol{q}_e)$,

$$\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) \le \max_{\boldsymbol{q}' \in \mathcal{Q}_\xi(\boldsymbol{q}_e)} \min_{V \in \mathcal{V}(\boldsymbol{p}, \boldsymbol{q}', \Delta)} I(\boldsymbol{q}'; V) < R_{\mathcal{X}}, \tag{4.61}$$

where (4.48) is used to establish the last inequality. Consequently, Lemma 4.2 implies that

$$\begin{aligned}&\Pr(|\mathcal{N}(\boldsymbol{y}_{m'}, \Delta)| < n^\alpha \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \boldsymbol{y}_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}])\\ &\quad\le \Pr(|\mathcal{N}_L(\boldsymbol{y}_{m'}, \Delta)| < n^\alpha \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots \boldsymbol{y}_M\}, \mathcal{Y} \in [\mathcal{A}^{nM}])\\ &\quad\le \exp[-\exp_2[n(R_{\mathcal{X}} - R_\theta) + o(n)]]\end{aligned} \tag{4.62}$$

where R_θ denotes the middle term in (4.61) and the first inequality follows from the definition of $|\mathcal{N}_L(\boldsymbol{y}_{m'}, \Delta)|$. Inequality (4.62) shows that the first term in (4.55) converges to 0 of double-exponential order of n by virtue of (4.61).

Now, the second term in (4.55) is evaluated by the similar way that establishes (4.62). First, choose β_1 satisfying $0 < \beta_1 < \beta$ arbitrarily and define a new random variable $|\mathcal{N}_U(\boldsymbol{y}_{m'}, \Delta)|$ as follows:

$$|\mathcal{N}_U(\boldsymbol{y}_{m'}, \Delta)| = |\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_{\beta_1}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_{m'}) \leq \Delta\}| \tag{4.63}$$

Note that $|\mathcal{N}_U(\boldsymbol{y}_{m'}, \Delta)|$ is governed by the binomial distribution $B(L, \theta_U)$, where

$$\begin{aligned} \theta_U &= \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_{\beta_1}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}_{m'}) \leq \Delta\} \\ &= 2^{-n \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) + o(n)}, \end{aligned} \tag{4.64}$$

where $\tilde{\boldsymbol{q}} = \boldsymbol{t}(\boldsymbol{y}_{m'})$ and Lemma 4.3 yields the last equality. For $\boldsymbol{y}_{m'}$ satisfying $D(\boldsymbol{p}^*||\boldsymbol{t}(\boldsymbol{y}_{m'})) > \varepsilon$ (4.64) is evaluated as follows:

$$\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\tilde{\boldsymbol{q}}; V) \geq \min_{\boldsymbol{q}' : D(\boldsymbol{p}^*||\boldsymbol{q}') > \varepsilon} \ \min_{V \in \mathcal{V}(\boldsymbol{p}, \boldsymbol{q}', \Delta)} I(\boldsymbol{q}'; V). \tag{4.65}$$

Notice that the right hand side of (4.65) is greater than the rate-distortion function $R(\boldsymbol{p}, \Delta)$, and hence, it is greater than $R_{\mathcal{X}}$. The second term in (4.55) is evaluated in the following way:

$$\begin{aligned} &\Pr(N_{max} \geq n^\alpha \text{ and } D(\boldsymbol{p}^*||\boldsymbol{t}(\boldsymbol{y}_{m^*})) > \varepsilon \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_M\}, \, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\quad \leq \Pr(|\mathcal{N}(\boldsymbol{y}_m, \Delta)| \geq n^\alpha \text{ for any } \boldsymbol{y}_m \in \mathcal{Y} \text{ satisfying } D(\boldsymbol{p}^*||\boldsymbol{t}(\boldsymbol{y}_m)) > \varepsilon \,| \\ &\qquad\quad \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_M\}, \, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\quad \leq \Pr(|\mathcal{N}_U(\boldsymbol{y}_m, \Delta)| \geq n^\alpha \text{ for any } \boldsymbol{y}_m \in \mathcal{Y} \text{ satisfying } D(\boldsymbol{p}^*||\boldsymbol{t}(\boldsymbol{y}_m)) > \varepsilon \,| \\ &\qquad\quad \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_M\}, \, \mathcal{Y} \in [\mathcal{A}^{nM}]) \\ &\quad \leq \exp[-n^{\alpha+1}(R'_\theta - R_{\mathcal{X}}) \cdot \ln 2 + nR'_\theta \ln 2 + o(n)], \end{aligned} \tag{4.66}$$

where R'_θ denotes the right hand side of (4.65) and Lemma 4.2 is used to deduce the last inequality.

Finally, the second term in (4.51) is evaluated. Notice that both (4.62) and (4.66) do not depend on $\mathcal{Y}$. Therefore,

$$\sum_{\mathcal{Y} \in [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}) \Pr(D(\boldsymbol{p}^*||\hat{\boldsymbol{p}^*}) > \varepsilon \,|\, \mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M\})$$

$$\leq \Big(\exp[-\exp_2[n(R_{\mathcal{X}} - R_\theta) + o(n)]] + \exp[-n^{\alpha+1}(R'_\theta - R_{\mathcal{X}}) \cdot \ln 2 + nR_{\mathcal{X}} \ln 2 + o(n)]\Big)$$
$$\times \sum_{\mathcal{Y} \in [\mathcal{A}^{nM}]} \Pr(\mathcal{Y} = \{y_1, y_2, \ldots, y_M\})$$
$$\leq \exp[-\exp_2[n(R_{\mathcal{X}} - R_\theta) + o(n)]]$$
$$+ \exp[-n^{\alpha+1}(R'_\theta - R_{\mathcal{X}}) \cdot \ln 2 + nR_L \ln 2 + o(n)], \quad (4.67)$$

which becomes less than δ for sufficiently large n. This completes the proof of this theorem. □

Judging from (4.51), (4.53) and (4.67) in the proof of Theorem 4.1), the most dominant term is the second term in (4.67). For making the probability of error caused by the algorithm less than δ, n of $O(\log_2 \frac{1}{\delta})$ should be chosen for sufficiently large n.

4.3 Another Universal Estimation Algorithm

4.3.1 Definition and Properties of the Algorithm

In Section 4.2 a universal estimation algorithm of p^* defined by (4.3) is proposed. The algorithm requires two kinds of training sequences and for arbitrarily chosen $\varepsilon > 0$ and $\delta \in (0,1)$ outputs $\hat{p^*}$ satisfying $D(p^* \| \hat{p^*}) \leq \varepsilon$ with probability at least $1 - \delta$. Sizes of n and the training sequences required by the algorithm is evaluated as functions of ε and δ. In other words, it permits an estimation error up to ε in order to make length of the required training sequences small. In this section another universal estimation algorithm of p^* that outputs a probability distribution arbitrarily close to p^* with high probability is proposed in this section.

The same notations introduced in Section 4.2.1 is also used here. A class of discrete memoryless sources of finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ is of interest. Assume that the probability distribution on $\mathcal{A}$, denoted by $p = (p(a_1), p(a_2), \ldots, p(a_J))$, satisfies $p(a_j) > 0$ for all $j = 1, 2, \ldots, J$. Denote by d a single-letter fidelity criterion satisfying three conditions 1), 2) and 3) given in Section 4.2.1 and define d_n by (4.8). For any fixed distortion level $\Delta \in (0, D_{max})$ define p^* by (4.10). Uniqueness of p^* is also assumed. The uniqueness of p^* implies uniqueness of the stochastic matrix W^* achieving the

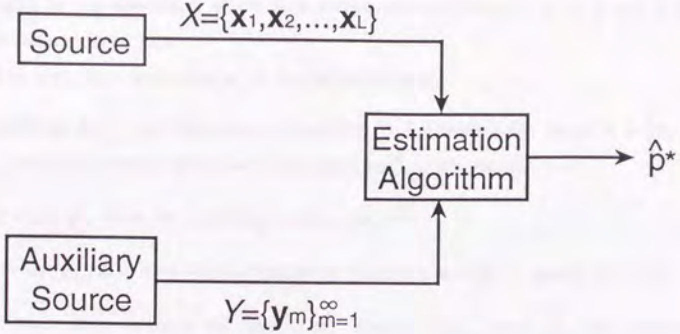


Figure 4.2 Block Diagram of the Universal Estimation System II

minimum in (4.1). It is calculated as

$$W^{*}(a_k|a_j) = \frac{p^{*}(a_k)\exp[sd(a_j,a_k)]}{\sum_{l=1}^{J} p^{*}(a_l)\exp[sd(a_j,a_l)]} \tag{4.68}$$

for all $j, k = 1, 2, \ldots, J$ [2]. *

A block diagram of the universal estimation system considered in this section is given in Fig. 4.2. Suppose that an auxiliary source with probability distribution $\boldsymbol{q} = (q(a_1), q(a_2), \ldots, q(a_J))$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ is available to the estimation algorithm. Denote by $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ L n-tuples drawn from the source and $\mathcal{Y} = \{\boldsymbol{y}_m\}_{m=1}^{\infty}$ an infinite sequence of n-tuples drawn from the auxiliary source. Though the cardinality of $\mathcal{Y}$ is finite in Section 4.2, $\mathcal{Y}$ in a form of $\mathcal{Y} = \{\boldsymbol{y}_m\}_{m=1}^{\infty}$ is required by the algorithm proposed in this section. Define

$$R_{\mathcal{X}} = \frac{1}{n}\log_2 L \tag{4.69}$$

*Instead of assuming the uniqueness of $\boldsymbol{p}^{*}$, $p^{*}(a_j) > 0$ for all $j = 1, 2, \ldots, J$ and regularity of the matrix $A = (e^{-sd(a_j,a_k)})$ for $s = \frac{d}{dD}R(\boldsymbol{p}, D)|_{D=\Delta}$ can be assumed [2]. In each case uniqueness of W^{*} plays an important role in the following sections.

Assume that an estimate of $\boldsymbol{p}$, denoted by $\boldsymbol{p}_e$, satisfying $\|\boldsymbol{p} - \boldsymbol{p}_e\|_1 = O(n^{-\beta_e})$ is also available to the algorithm, where β_e is a constant satisfying $\beta_e \in (0, \frac{1}{2}]$ and $\|\cdot\|_1$ is the norm defined in (4.11).

The algorithm estimates $\boldsymbol{p}^*$ in the following way:

Algorithm 4.2 *1) Choose $\alpha > 0$ and $\beta \in (0, \beta_e)$ arbitrarily. Draw $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$ from the source. Set $m = 1$ and select sufficiently small $\gamma > 0$.*

2) Draw $\boldsymbol{y}_m$ from the auxiliary source. Define

$$\mathcal{N}(\boldsymbol{y}_m, \Delta + \gamma) = \{\boldsymbol{x} \in \mathcal{X} \mid d_n(\boldsymbol{x}, \boldsymbol{y}) \leq \Delta + \gamma \text{ and } \|\boldsymbol{p}_e - \boldsymbol{t}(\boldsymbol{x})\|_1 \leq n^{-\beta}\}, \quad (4.70)$$

where $\boldsymbol{t}(\boldsymbol{x})$ denotes the type of $\boldsymbol{x}$. Count $|\mathcal{N}(\boldsymbol{y}_m, \Delta + \gamma)|$, the cardinality of $\mathcal{N}(\boldsymbol{y}_m, \Delta + \gamma)$.

3) If $|\mathcal{N}(\boldsymbol{y}_m, \Delta + \gamma)| \geq n^\alpha$, then output $\boldsymbol{t}(\boldsymbol{y}_m)$. Otherwise, increment m and go to 2). □

Choices of L and γ are crucial in Algorithm 4.2. The following theorem guarantees that the algorithm outputs $\hat{\boldsymbol{p}^*}$ arbitrarily close to $\boldsymbol{p}^*$ under appropriate choices of $R_\mathcal{X} = \frac{1}{n}\log_2 L$ and γ and stops *with probability one* when n is sufficiently large.

Theorem 4.2 *Let $\Delta \in (0, D_{max})$ be a distortion level arbitrarily fixed and assume that $\boldsymbol{p}^*$ defined by (4.3) is unique. If $R_\mathcal{X}$ is equal to $R(\boldsymbol{p}, \Delta)$, then for any fixed $\eta > 0$ there exists $\gamma_0 = \gamma_0(\eta)$ satisfying the following properties for all $\gamma \in (0, \gamma_0]$:*

(I) *there exist a constant $\xi = \xi(\eta, \gamma) > 0$ and an integer $n_0 = n_0(\eta, \gamma)$ that satisfies*

$$\Pr(|\mathcal{N}(\boldsymbol{y}, \Delta + \gamma)| > n^\alpha) < \exp[-n^{\alpha+1}\xi \ln 2 + nR_\mathcal{X} \ln 2 + o(n)] \quad (4.71)$$

for all integers $n > n_0$ and $\boldsymbol{y} \in \mathcal{A}^n$ satisfying $\|\boldsymbol{p}^* - \boldsymbol{t}(\boldsymbol{y})\|_1 > \eta$,

(II) *there exists an integer n_0' that guarantees the existence of a type $\tilde{\boldsymbol{q}}$ satisfying the following properties:*

(a) *gap between $\boldsymbol{p}^*$ and $\boldsymbol{q}$ in l_1 sense satisfies*

$$\|\boldsymbol{p}^* - \tilde{q}\|_1 < \eta \quad (4.72)$$

(b) *there exists a positive number* $\xi' = \xi'(\eta, \gamma, \tilde{q})$ *that satisfies*

$$\Pr(|\mathcal{N}(y, \Delta + \gamma)| < n^{\alpha}) < \exp[-\exp_2[n\xi' - o(n)]] \qquad (4.73)$$

for all integers $n > n_0'$ *and* y *of type* $\tilde{q}$

where $\Pr(\cdot)$ *means the probability with respect to* $\mathcal{X}$.

Moreover, Algorithm 4.2 *stops with probability one.*

Since arbitrary $\alpha > 0$ is chosen in Algorithm 4.2, property (I) implies that the algorithm outputs types of $\mathcal{A}^n$ not satisfying $\|p^* - \tilde{q}\|_1 < \eta$ with probability tending to 0 of exponential order of n. On the other hand, property (II) guarantees the existence of types that the algorithm puts out with probability going to 1 of double-exponential order of n.

Theorem 4.2 seems to treat only the case that $R_{\mathcal{X}}$ is equal to the rate-distortion bound $R(p, \Delta)$ though any knowledge on $R(p, \Delta)$ is not assumed. Nevertheless, (4.71) holds for any $y \in \mathcal{A}^n$ if $R_{\mathcal{X}} < R(p, \Delta)$ and otherwise (4.73) holds for a type not satisfying (4.72). Consequently, by varying $R_{\mathcal{X}}$ from zero to $H(p)$ Algorithm 4.2 can estimate not only p^* but also $R(p, \Delta)$.

For establishing property (I) an arbitrarily $\gamma \in [0, \gamma_0]$ can be chosen. However, it is hard to guarantee the existence of a type satisfying property (II) when $\gamma = 0$. On the other hand, for establishing only property (II) simple definition of $\mathcal{N}(y, \Delta + \gamma)$ such as

$$\mathcal{N}(y, \Delta + \gamma) = \{x \in \mathcal{X} \mid d_n(x, y) \leq \Delta + \gamma\} \qquad (4.74)$$

can be used, though property (I) is not guaranteed.

4.3.2 Analysis of the Algorithm

This section is devoted to a proof of Theorem 4.2. Three lemmata shown in Section 4.2.3 are also keys to the proof.

Proof of Theorem 4.2: Fix $\Delta \in (0, D_{max})$, $\eta > 0$ and $y \in \mathcal{A}^n$ satisfying $\|p^* - t(y)\|_1 > \eta$ arbitrarily. Define $R_{\mathcal{X}} = R(p, \Delta)$. Firstly, property (I) is established under an appropriate choice of $\gamma > 0$.

Fix any β_1 satisfying $0 < \beta_1 < \beta$ and define

$$\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma) = \{\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K}_{\beta_1}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta + \gamma\} \tag{4.75}$$

where $\mathcal{K}_{\beta_1}(\boldsymbol{p})$ is a set defined by (4.24). Notice that $|\mathcal{N}_U(\boldsymbol{y}, \Delta+\gamma)|$ as well as $|\mathcal{N}(\boldsymbol{y}, \Delta+\gamma)|$ is a binomial random variable. By considering the fact described in Lemma 4.1, there exists an integer n_1 satisfying

$$\Pr(|\mathcal{N}(\boldsymbol{y}, \Delta + \gamma)| > n^\alpha) \le \Pr(|\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma)| > n^\alpha) \tag{4.76}$$

for all $n > n_1$. Therefore, if the existence of $\xi > 0$ that satisfies $E|\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma)| < 2^{-n\xi + o(n)}$ is established, Lemma 4.2 and (4.76) yield (4.71), where E denotes the probability with respect to $\mathcal{X}$.

Since element of $\mathcal{X}$ are independently drawn from the source, $|\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma)|$ is evaluated as follows:

$$\begin{aligned} E|\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma)| &= L \cdot \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{A}^n \cap \mathcal{K}_{\beta_1}(\boldsymbol{p}) \,|\, d_n(\boldsymbol{x}, \boldsymbol{y}) \le \Delta + \gamma\} \\ &< 2^{-n[\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta+\gamma)} I(\tilde{\boldsymbol{q}}; V) - R(\boldsymbol{p}, \Delta)] + o(n)} \end{aligned} \tag{4.77}$$

where $\tilde{\boldsymbol{q}} = \boldsymbol{t}(\boldsymbol{y})$ and the definition of L and Lemma 4.3 deduce the last inequality. From the uniqueness of $\boldsymbol{p}^*$ it is clear that

$$\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta)} I(\boldsymbol{q}; V) > R(\boldsymbol{p}, \Delta) \tag{4.78}$$

for all $\tilde{\boldsymbol{q}}$ satisfying $\|\boldsymbol{p}^* - \tilde{\boldsymbol{q}}\|_1 > \eta$. Since $\boldsymbol{p}^*$ is unique and the mutual information $I(\tilde{\boldsymbol{q}}; V)$ is continuous with respect to V, there exists a real number γ_1 satisfying

$$\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta+\gamma)} I(\tilde{\boldsymbol{q}}; V) > R(\boldsymbol{p}, \Delta) \tag{4.79}$$

for all $\gamma \in [0, \gamma_1]$ and accordingly, there exists $\xi = \xi(\eta, \gamma) > 0$ that satisfies

$$\min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}, \Delta+\gamma)} I(\tilde{\boldsymbol{q}}; V) - R(\boldsymbol{p}, \Delta) \ge \xi \tag{4.80}$$

Combining (4.77) and (4.80) yields

$$E|\mathcal{N}_U(\boldsymbol{y}, \Delta + \gamma)| < 2^{-n\xi + o(n)} \tag{4.81}$$

and hence, Lemma 4.2 implies (4.71).

Secondly, property (II) is proved. Since the rate-distortion function $R(\boldsymbol{p}, D)$ is strictly monotone decreasing function with respect to D, for any fixed $\gamma > 0$

$$I(\boldsymbol{p}; W^*) = R(\boldsymbol{p}, \Delta) > R(\boldsymbol{p}, \Delta + \gamma) = I(\boldsymbol{p}, W^\dagger), \tag{4.82}$$

where $W^\dagger$ is the $J \times J$ stochastic matrix achieving the minimum in (4.1) at $D = \Delta + \gamma$. Define $\boldsymbol{p}^\dagger$ by

$$p^\dagger(a_k) = \sum_{j=1}^{J} p(a_j) W^\dagger(a_k|a_j) \tag{4.83}$$

for all $k = 1, 2, \ldots, J$. Existence of $\gamma_2 = \gamma_2(\eta)$ satisfying $\|\boldsymbol{p}^* - \boldsymbol{p}^\dagger\|_1 < \eta$ for all $\gamma \in (0, \gamma_2]$ is guaranteed by the uniqueness of W^*. Nonexistence of such γ_2 conflicts to the uniqueness of W^* originated from the uniqueness of $\boldsymbol{p}^*$.

Note that an arbitrary approximation of $\boldsymbol{p}^\dagger$ is realized by a rational number of denominator n with increasing n. Hence, there exists a type $\tilde{\boldsymbol{q}}^\dagger$ satisfying $\|\boldsymbol{p}^* - \tilde{\boldsymbol{q}}^\dagger\|_1 < \eta$ if n is sufficiently large. Moreover, continuity of the mutual information leads to $\tilde{\boldsymbol{q}}^\dagger$ that also satisfies

$$R(\boldsymbol{p}, \Delta) - \min_{V \in \mathcal{V}(\boldsymbol{p}, \tilde{\boldsymbol{q}}^\dagger, \Delta+\gamma)} I(\tilde{\boldsymbol{q}}^\dagger; V) \geq \xi' \tag{4.84}$$

for a constant $\xi >' 0$ by choosing smaller γ_2 if necessary. It is shown that any $\boldsymbol{y}$ of type $\tilde{\boldsymbol{q}}^\dagger$ satisfies (4.73).

Fix $\boldsymbol{y} \in \mathcal{A}^n$ of type $\tilde{\boldsymbol{q}}^\dagger$ and a real number β_2 satisfying $\beta < \beta_2 < \beta_e$ arbitrarily, where β_2 is a parameter that determines precise of $\boldsymbol{p}_e$. Define

$$\mathcal{N}_L(\boldsymbol{y}, \Delta + \gamma) = \{\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K}_{\beta_2} \mid d_n(\boldsymbol{x}, \boldsymbol{y}) \leq \Delta + \gamma\} \tag{4.85}$$

and notice again that $|\mathcal{N}(\boldsymbol{y}, \Delta + \gamma)|$ and $|\mathcal{N}_L(\boldsymbol{y}, \Delta + \gamma)|$ are binomial random variables. Since Lemma 4.1 guarantees the existence of an integer n_2 satisfying

$$\Pr(|\mathcal{N}(\boldsymbol{y}, \Delta + \gamma)| < n^\alpha) \leq \Pr(|\mathcal{N}_L(\boldsymbol{y}, \Delta + \gamma)| < n^\alpha) \tag{4.86}$$

for all $n > n_2$. it is sufficient to show that the right hand side of (4.86) converges to zero of double-exponential order of n. If expectation of $|\mathcal{N}_L(\boldsymbol{y}, \Delta + \gamma)|$ with respect to $\mathcal{X}$ turns out to grow of exponential order of n, Lemma 4.2 implies the convergence. It is evaluated in the following manner:

$$\begin{aligned} E|\mathcal{N}_L(\boldsymbol{y}, \Delta+\gamma)| &= L \cdot \boldsymbol{p}^n\{\boldsymbol{x} \in \mathcal{K}_{\beta_2}(\boldsymbol{p}) \mid d_n(\boldsymbol{x}, \boldsymbol{y}) \leq \Delta+\gamma\} \\ &> 2^{n[R(\boldsymbol{p},\Delta)-\min_{V\in\mathcal{V}(\boldsymbol{p},\tilde{\boldsymbol{q}}^\dagger,\Delta+\gamma)} I(\tilde{\boldsymbol{q}}^\dagger;V)]-o(n)}, \end{aligned} \tag{4.87}$$

where $L = 2^{nR(\boldsymbol{p},\Delta)}$ and a property of binomial random valuables yield the equality, Lemma 4.3 implies the first inequality and (4.84) is used for obtaining the inequality. Hence, Lemma 4.2 and (4.86) deduce

$$\Pr(|\mathcal{N}(\boldsymbol{y}, \Delta+\gamma)| < n^\alpha) < \exp[-\exp_2[n\xi' - o(n)]]. \tag{4.88}$$

Choosing $\gamma_0 = \min\{\gamma_1, \gamma_2\}$ completes the proof of the properties (I) and (II) of the theorem.

Finally, it is shown that Algorithm 4.2 stops with probability 1 if $R_\mathcal{X}$ is equal to $R(\boldsymbol{p}, \Delta)$ and an arbitrarily $\gamma \in (0, \gamma_0]$ and an appropriate n are chosen. For all $m = 1, 2, \ldots$ define E_m as the event that the algorithm does not stop by m-th iteration, i.e, by the time when $\boldsymbol{y}_m$ is drawn and tested. Then, for all integers $m > 1$

$$\Pr(E_m) = \Pr(E_{m-1}) \cdot \Pr(|\mathcal{N}(\boldsymbol{y}_m, \Delta+\gamma)| < n^\alpha) \tag{4.89}$$

and hence,

$$\begin{aligned} \frac{\Pr(E_m)}{\Pr(E_{m-1})} &= \Pr(|\mathcal{N}(\boldsymbol{y}_m, \Delta+\gamma)| < n^\alpha) \qquad (4.90) \\ &< \Pr(\boldsymbol{t}(Y_m^n) = \tilde{\boldsymbol{q}}^\dagger) \Pr(|\mathcal{N}(Y_m^n, \Delta+\gamma)| < n^\alpha \mid \boldsymbol{t}(Y_m^n) = \tilde{\boldsymbol{q}}^\dagger) \\ &\quad + \Pr(\boldsymbol{t}(Y_m^n) \neq \tilde{\boldsymbol{q}}^\dagger) \Pr(|\mathcal{N}(Y_m^n, \Delta+\gamma)| < n^\alpha \mid \boldsymbol{t}(Y_m^n) \neq \tilde{\boldsymbol{q}}^\dagger), \qquad (4.91) \end{aligned}$$

where $\tilde{\boldsymbol{q}}^\dagger$ is a type used for establishing property (II) and Y_m^n is a m-th random vector from the auxiliary source. Since $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$ from the assumption of the auxiliary source, $\Pr(\boldsymbol{t}(\boldsymbol{y}) = \tilde{\boldsymbol{q}}^\dagger) \approx 2^{-nD(\tilde{\boldsymbol{q}}^\dagger \| \boldsymbol{q}^*)}$ and $D(\tilde{\boldsymbol{q}}^\dagger \| \boldsymbol{q}) < \infty$. By using property (I) and property (II) $\Pr(E_m)/\Pr(E_{m-1})$ is evaluated in the following manner:

$$\begin{aligned} \frac{\Pr(E_m)}{\Pr(E_{m-1})} &< 2^{-nD(\tilde{\boldsymbol{q}}^\dagger \| \boldsymbol{q})+o(n)} \exp[-n^{\alpha+1}\xi \ln 2 + nR_\mathcal{X} \ln 2 + o(n)] \\ &\quad + (1 - 2^{-nD(\tilde{\boldsymbol{q}}^\dagger \| \boldsymbol{q})+o(n)}) \exp[-\exp_2[n\xi' - o(n)]], \end{aligned} \tag{4.92}$$

where n satisfying $n \geq \max\{n_0, n_0'\}$ is chosen. The right hand side of (4.92) guarantees the existence of an integer $n_0'' = n_0''(\lambda)$ that satisfies $\Pr(E_m)/\Pr(E_{m-1}) < 1 - \lambda$ for all

$n > n_0''$ and $m > 1$, where $\lambda \in (0,1)$ is an arbitrarily fixed constant. If n satisfying $n \geq \max\{n_0, n_0', n_0''\}$ is chosen, $\Pr(E_m) \to 0$ of exponential order of m and $\sum_{m=1}^{\infty} \Pr(E_m) < \infty$. Therefore, the Borel-Cantelli lemma guarantees that Algorithm 4.2 stops *with probability one.* □

4.4 Conclusion

Two universal estimation algorithms of a probability distribution $\boldsymbol{p}^*$ on a source alphabet $\mathcal{A}$ that is the output probability distribution of the test channel for discrete memoryless sources are proposed. The algorithm requires two kinds of training sequences and outputs an estimate of $\boldsymbol{p}^*$. One of the sequence is drawn from the source to be compressed and the other is drawn from an auxiliary memoryless source that puts out all of symbols in the same alphabet.

One of the two algorithm requires $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_L\}$, $\boldsymbol{x}_l \in \mathcal{A}^n$ for all $l = 1, 2, \dots, L$ from the source and $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M\}$ $\boldsymbol{y}_m \in \mathcal{B}^n$ for all $m = 1, 2, \dots, M$ from the auxiliary source. It outputs an estimate of $\boldsymbol{p}^*$, denoted by $\hat{\boldsymbol{p}^*}$, satisfying $D(\boldsymbol{p}^* \| \hat{\boldsymbol{p}^*}) \leq \varepsilon$ with probability at least $1 - \delta$ for any fixed $\varepsilon > 0$ and $\delta \in (0,1)$. Lower bounds of $R_{\mathcal{X}} = \frac{1}{n}\log_2 L$, $R_{\mathcal{Y}} = \frac{1}{n}\log_2 M$ and n required for meeting the criterion are evaluated as a function of ε and δ. A property of binomial random variables and conventional techniques of the Shannon theory such as the type are used in order to justify the algorithm. If the algorithm is applied to the data compression scheme with fidelity criterion proposed in Chapter 3, data compression of rate $R(\boldsymbol{p}, \Delta) + \varepsilon$ and average distortion close to Δ per source symbol becomes possible in an asymptotic sense. The imposed criterion on estimation resembles to the PAC learning models that often appears in the field of the computational learning theory. The algorithm not only is a construction of estimator of $\boldsymbol{p}^*$ but also throws light on the Shannon theoretic field from a viewpoint of the computational complexity.

The other algorithm requires $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_L\}$, $\boldsymbol{x}_l \in \mathcal{A}^n$ for all $l = 1, 2, \dots, L$ from the source and $\mathcal{Y} = \{\boldsymbol{y}_m, \}_{m=1}^{\infty}$, $\boldsymbol{y}_m \in \mathcal{A}^n$ for all $m = 1, 2, \dots,$ from the auxiliary source. It can output an estimate of $\boldsymbol{p}^*$ arbitrarily close to the true distribution with high probability if appropriate parameters are chosen in the algorithm. Though an infinite sequence from the auxiliary source is required, the algorithm is guaranteed to stop *with probability one*.

However, there remains several problems to be solved. Is it possible to obtain a lower bound of $R_{\mathcal{X}}$ that dose not depend on p in the first formulation of estimation problem? This problem is essential for the sake of universal encoding. Development of an algorithm without use the estimate of probability distribution of the sources and establishing required size of training sequence should be further considered.

Chapter 5

Conclusion

In this thesis asymptotic properties of data compression schemes with fidelity criterion are discussed. The rate-distortion function $R(D)$, which is defined as a minimum of the mutual information, describes achievable rate of data compression schemes when average distortion up to D is permitted. Existence of an asymptotically optimal block code that achieves the rate-distortion function is guaranteed by the source coding theorem. For readers' convenience, Chapter 1 is devoted to introduction of such basic properties on data compression schemes with fidelity criterion.

In Chapter 2 encoding of memoryless gaussian sources is considered under the squared-error criterion. For the sources of known mean and variance it is shown that an encoding scheme similar to the shape-gain vector quantization is asymptotically optimal under certain conditions. In the scheme source blocks of blocklength n are encoded by using a scalar quantizer and a set of point on n-dimensional unit hypersphere. If the set of points is appropriately chosen, asymptotic behavior of rate and average distortion is characterized by the scalar quantizer. The encoding method is easily extended to the sources of unknown mean or variance by taking the two-pass encoding strategy. Given a data block of blocklength n, the encoder transmits an estimate of unknown parameters with error $O(\frac{1}{\sqrt{n}})$ to the decoder first and encodes the block as if the parameters are known. The probability that distortion between source blocks and their reproduced form is not greater than an acceptable level is analyzed. Under the assumption that the unknown parameters lie in a known and bounded set, it is proved that the probability tends to one as $n \to \infty$ with a slight loss of rate even if mean or variance are unknown. The extra rate required for transmission of the block is (i)$\frac{1}{2}\log_2 n + C_1$ when only mean is unknown, (ii) $\frac{1}{2}\log_2 n + C_2\sqrt{n} + C_3$ when only variance is unknown and (iii) $\log_2 n + C_4\sqrt{n} + C_5$ when both mean and variance are unknown, where $C_i, i = 1, 2, \ldots, 5$ are constants.

A data compression algorithm based on string matching is treated in Chapter 3. In the algorithm both an encoder and a decoder are supposed to share a data-base sequence of sufficient length in each buffer. Given a source block of blocklength n, the encoder searches for the minimum index of the buffer with property that distortion between the source block and n consecutive symbols beginning from the index is within an acceptable level. It transmits the index in the fixed-length binary form to the decoder. The decoder reproduces the source block as n consecutive symbols from the transmitted index. Rate required by the scheme is analyzed for discrete memoryless sources and memoryless gaussian sources. In case of the discrete memoryless source of alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ and probability distribution $\boldsymbol{p}$ if the data-base sequence is drawn according to probability distribution $\boldsymbol{q}$ satisfying $q(a_j) > 0$ for all $j = 1, 2, \ldots, J$, it is shown that $D(\boldsymbol{q}||\boldsymbol{p}^*)$ expresses redundancy of the scheme, where $\boldsymbol{p}^*$ is the output probability distribution of the test channel and $D(\cdot||\cdot)$ denotes the divergence. The result directly implies a sufficient condition for making the redundancy asymptotically go to zero. In case of memoryless gaussian sources, however, any i.i.d. gaussian sequence of known mean and variance can make the redundancy in rate asymptotically equal to zero by modifying the way of matching. These results unveils an essential difference between compression of discrete memoryless sources and compression of memoryless gaussian sources.

Chapter 4 is devoted to development of algorithms estimating $\boldsymbol{p}^*$ universally, the output probability distribution of the test channel. Two universal estimating algorithm are proposed and analyzed. Only discrete memoryless sources of alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ and probability distribution $\boldsymbol{p}$ is of interest. The first algorithm requires data sequence $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$, $\boldsymbol{x}_l \in \mathcal{A}^n$ from the source and $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$, $\boldsymbol{y}_m \in \mathcal{A}^n$ from another discrete memoryless source of probability distribution $\boldsymbol{q}$ satisfying $q(a_k) > 0$ for all $k = 1, 2, \ldots, J$. It outputs $\hat{\boldsymbol{p}}^*$ as an estimate of p^*. The criterion imposed on the algorithm resembles the PAC (Probably Approximately Correct) learning models that often appear in the field of computational learning theory. For arbitrarily fixed $\varepsilon > 0$ and $\delta \in (0,1)$ the algorithm must output $\hat{p^*}$ satisfying $D(\hat{p^*}||p^*) \leq \varepsilon$ with probability at least $1-\delta$. Lower bounds of $R_{\mathcal{X}} \stackrel{\text{def}}{=} \frac{1}{n}\log_2 L$, $R_{\mathcal{Y}} \stackrel{\text{def}}{=} \frac{1}{n}\log_2 M$, and n that enable the algorithm to meet the criterion of estimation are deduced as a function of ε and δ. As a byproduct of the algorithm, redundancy of the encoding algorithm proposed in Chapter 3 becomes not greater than ε for any discrete memoryless sources when data-base used in the encoding algorithm is drawn according to the probability distribution that the estimation algorithm outputs. The other algorithm requires $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L\}$, $\boldsymbol{x}_l \in \mathcal{A}^n$ from the source and $\mathcal{Y} = \{\boldsymbol{y}_m\}_{m=1}^{\infty}$, $\boldsymbol{y}_m \in \mathcal{A}^n$ from another discrete memoryless source. The algorithm only requires memory for storage of one element of $\mathcal{Y}$, which is different from the previous algorithm. It is shown that the algorithm outputs $\hat{p^*}$ that is arbitrarily close to p^* with high probability and stops *with probability one* when parameters in the algorithm are appropriately chosen,

Appendix A

For Section 2.3

This appendix gives a proof about the convergence of the second term in (2.48). Since the volume of D_{Lm} is infinite for all $m = 1, 2, \ldots, M$, there are blocks $\boldsymbol{x} \in D_{Lm}$ encoded with large distortion. For completing the proof of Theorem 2.1 it should be shown that

$$Q_n \stackrel{\text{def}}{=} E\left[\frac{1}{n}\sum_{m=1}^{M}\int_{D_{Lm}}(\langle \boldsymbol{x}, \boldsymbol{y}_m\rangle - a_L)^2 p(\boldsymbol{x})d\boldsymbol{x}\right]. \tag{A.1}$$

has an upper-bound that decays of exponential order of n under the condition C1).

If $\boldsymbol{x} \in D_{Lm}$ for some $m = 1, 2, \ldots, M$, $\boldsymbol{x}$ satisfies $\langle \boldsymbol{x}, \boldsymbol{y}_m\rangle \geq a_L$ from the definition of φ_2. By applying $(\langle \boldsymbol{x}, \boldsymbol{y}_m\rangle - a_L)^2 \leq (\|\boldsymbol{x}\|_2 - a_L)^2$ for such $\boldsymbol{x}$, Q_n is upper-bounded as follows:

$$\begin{aligned} Q_n &\leq \frac{1}{n}E\left[\sum_{m=1}^{M}\int_{D_{Lm}}(\|\boldsymbol{x}\|_2 - a_L)^2 p(\boldsymbol{x})d\boldsymbol{x}\right] \\ &< \frac{1}{n}\int_{\|\boldsymbol{x}\|_2 \geq a_L}(\|\boldsymbol{x}\|_2 - a_L)^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\int_u^\infty (r-u)^2 r^{n-1}\exp\left[-\frac{r^2}{2}\right]dr, \end{aligned} \tag{A.2}$$

where $r = \|\boldsymbol{x}\|_2$ and $u = a_L$. Since $(r-u)^2 \leq r^2 - u^2$ for any $0 \leq u \leq r$, the formula of partial integral implies that

$$\begin{aligned} Q_n &< \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\int_u^\infty r^{n+1}\exp\left[-\frac{r^2}{2}\right]dr \\ &\quad - \frac{u^2}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\int_u^\infty r^{n-1}\exp\left[-\frac{r^2}{2}\right]dr \end{aligned}$$

$$
\begin{aligned}
&= \frac{u^n}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\exp\left[-\frac{u^2}{2}\right] \\
&\quad - \frac{u^2-n}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\int_u^\infty r^{n-1}\exp\left[-\frac{r^2}{2}\right]dr,
\end{aligned}
\tag{A.3}
$$

The second term in (A.3) is positive since $u^2 = n^{1+2\alpha} > n$ for some fixed $\alpha > 0$ as is found in C1) and the integrand is non-negative. Then, (A.3) is evaluated as follows:

$$
Q_n < \frac{n^{(\frac{1}{2}+\alpha)n}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2}+1)}\exp\left[-\frac{n^{1+2\alpha}}{2}\right]. \tag{A.4}
$$

To show the exponential decay of (A.4), the Stirling formula

$$
\ln\Gamma(s) \approx \frac{1}{2}\ln 2\pi + \left(s-\frac{1}{2}\right)\ln s - s \tag{A.5}
$$

is used. Taking the logarithm of (A.4) yields

$$
\begin{aligned}
\ln Q_n &< \left(\frac{1}{2}+\alpha\right)n\ln n - \frac{n}{2}\ln 2 - \frac{1}{2}\ln 2\pi \\
&\quad - \frac{n+1}{2}\ln\left(\frac{n}{2}+1\right) - \frac{n}{2} - 1 - \frac{1}{2}n^{1+2\alpha} \\
&= -\frac{1}{2}n^{1+2\alpha} + O(n\ln n),
\end{aligned}
\tag{A.6}
$$

where $o(n^{1+2\alpha})$ denotes terms of order lower than $1+2\alpha$. Consequently, Q_n has an upper-bound in a form

$$
Q_n < \exp\left[-\frac{1}{2}n^{1+2\alpha} + O(n\ln n)\right], \tag{A.7}
$$

which decays of exponential order of n.

Appendix B

For Section 2.4

For completing the proof of Theorem 2.4 the existence of the integer n_3 satisfying $\Pr(E_2|E_4^c) < \varepsilon$ for all $n > n_3$ should be established. The probability is expressed as follows:

$$\Pr(E_2|E_4^c) = \int_{\boldsymbol{R}^n} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x}, \tag{B.1}$$

where χ is defined in (2.68). Fix an arbitrary ξ satisfying $\xi \in (0, \sigma^2)$ and define S_ξ by

$$S_\xi = \left\{\boldsymbol{x} \in \boldsymbol{R}^n \mid \sqrt{n(\sigma^2-\xi)} \le \|\boldsymbol{x}-\boldsymbol{\mu}\|_2 \le \sqrt{n(\sigma^2+\xi)}\right\}. \tag{B.2}$$

Then, $\Pr(E_2|E_4^c)$ in (B.1) is written as

$$\begin{aligned}\Pr(E_2|E_4^c) &= \int_{S_\xi} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\ &\quad + \int_{S_\xi^c} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\ &\le \int_{S_\xi} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\ &\quad + \int_{S_\xi^c} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x},\end{aligned} \tag{B.3}$$

where S_ξ^c denotes the complement of S_ξ and $0 \le \chi \le 1$ is used to obtain the last inequality. The weak law of large numbers guarantees the existence of an integer n_5 that satisfies the second term in (B.3) being less than $\frac{\varepsilon}{2}$ for all $n > n_5$. Therefore, it is sufficient to show that the first term in (B.3) becomes less than $\frac{\varepsilon}{2}$ for sufficiently large n.

It is convenient to represent an upper bound of $\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2$ as a function of $\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2$. First, the triangle inequality is used as follows:

$$\|\boldsymbol{x}-\boldsymbol{\mu}\|_2 \geq \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2 - \|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2. \tag{B.4}$$

Notice that the right hand side of (B.4) becomes positive for any $\boldsymbol{x} \in S_\xi$. Under the event E_4^c $\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2$ is evaluated in the following way:

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2 &\geq \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 2\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2 + \|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2^2 \\ &\geq \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 6\sqrt{K}\sigma\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2. \end{aligned} \tag{B.5}$$

Hence, the first term in (B.3) is upper-bounded by

$$\begin{aligned}&\int_{S_\xi} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right] d\boldsymbol{x} \\ &\leq \int_{S_\xi} \chi\left(\frac{\boldsymbol{x}-\hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 6\sqrt{K}\sigma\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}{2\sigma^2}\right] d\boldsymbol{x} \\ &= \int_{T_\xi} \chi\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 6\sqrt{K}\sigma\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}{2\sigma^2}\right] d\boldsymbol{x} \\ &\leq \int_{S_{\xi'}} \chi\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 6\sqrt{K}\sigma\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}{2\sigma^2}\right] d\boldsymbol{x} \end{aligned} \tag{B.6}$$

where T_ξ is a subset of $\boldsymbol{R}^n$ defined as

$$T_\xi = \{\boldsymbol{x} \in \boldsymbol{R}^n \mid \sqrt{n(\sigma^2-\xi)} \leq \|\boldsymbol{x}+\hat{\boldsymbol{\mu}}-2\boldsymbol{\mu}\|_2 \leq \sqrt{n(\sigma^2+\xi)}\,\} \tag{B.7}$$

and ξ' is an arbitrary number satisfying $0 < \xi' < \xi$. Note that there exists an integer n_6 satisfying $T_\xi \subset S_{\xi'}$ for all $n > n_6$ under the event E_4^c.

By transforming the rectangular coordinate system into the polar one, the right hand side of (B.6) is evaluated in the following way:

$$\begin{aligned}&\int_{S_{\xi'}} \chi\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\|\boldsymbol{x}-\boldsymbol{\mu}\|_2}, \mathcal{Y}^*\right)(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2^2 - 6\sqrt{K}\sigma\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|_2}{2\sigma^2}\right] d\boldsymbol{x} \\ &= \int_{\sqrt{n(\sigma^2-\xi')}}^{\sqrt{n(\sigma^2+\xi')}} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{r^2-6\sqrt{K}\sigma r}{2\sigma^2}\right] r^{n-1} dr \cdot \int_{S^{n-1}} \chi(\boldsymbol{z}, \mathcal{Y}^*)\, d\boldsymbol{z} \\ &\leq \int_{\sqrt{n(\sigma^2-\xi')}}^{\sqrt{n(\sigma^2+\xi')}} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{r^2-6\sqrt{K}\sigma r}{2\sigma^2}\right] r^{n-1} dr \cdot \exp[-\exp(n\beta)], \end{aligned} \tag{B.8}$$

where (2.69) yields the last inequality. The integral in (B.8) is upper-bounded as

$$\begin{aligned}
&\int_{\sqrt{n(\sigma^2-\xi')}}^{\sqrt{n(\sigma^2+\xi')}} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{r^2 - 6\sqrt{K}\sigma r}{2\sigma^2}\right] r^{n-1} dr \\
&\leq \sqrt{n(\sigma^2+\xi')}(2\pi\sigma^2)^{-\frac{n}{2}} \\
&\qquad \times \exp\left[-\frac{n(\sigma^2-\xi') - 6\sqrt{K}\sigma\sqrt{n(\sigma^2-\xi')}}{2\sigma^2}\right] (n(\sigma^2+\xi'))^{\frac{n-1}{2}} \\
&= \exp[n \log_2 n + O(n)]. \qquad (B.9)
\end{aligned}$$

Since β in (B.8) is positive, equations (B.8) and (B.9) imply that the right hand side of (B.6) converges to 0 as n tends to infinity. Therefore, there exists an integer n_7 that guarantees the first term in (B.3) being less than $\frac{\varepsilon}{2}$ for all $n > n_7$. Choosing $n_3 = \max\{n_5, n_6, n_7\}$ completes the proof of $\Pr(E_2|E_4^c) < \varepsilon$ for all $n > n_3$.

Appendix C

For Section 2.5

In Section 2.3 a randomly and independently generated points on S^{n-1} lead to the asymptotic optimality for memoryless gaussian sources with known mean and variance. However, this approach does not make sense in case of memoryless Laplacian sources though Fischer [20] establishes the asymptotic optimality by using a random coding argument restricted to an n-dimensional hyperspheres in the sense of $\|\cdot\|_1$. This appendix intends to clarify the reason why such approach fails.

In the case of gaussian, Lemma 2.2 formulates the probability that an arbitrary points on S^{n-1} does not belong to the region that guarantees the point to be encoded within an acceptable distortion Δ. Such approach is written in a form of (2.25) if a codebook is randomly and independently chosen from points on S^{n-1}. For evaluating the right hand side of (2.25), the ratio of $(n-1)$-dimensional volumes must be lower-bounded. If the logarithm of the lower bound divided by n converges to the form of rare-distortion function as $n \to \infty$, the asymptotic optimality of the code follows.

Can the same idea be applied to analysis of an encoding scheme for memoryless Laplacian sources of known mean and first-order absolute moment? The answer is no. Without loss of generality, the memoryless Laplacian source of zero mean unit first order absolute moment is assumed. Define $S_1(n,r)$ by (2.143) and choose an arbitrary $y \in S_1(n,1)$. Notice that $S_1(n,1)$ has $2n$ vertices, $n(n+1)$ edges and 2^n surfaces. Choose $\Delta > 0$ appropriately small, and define $T_n(y,\Delta)$ as follows:

$$T_n(y,\Delta) = \{z \in S_1(n,1) \mid \|z-y\|_1 \leq \Delta\}. \tag{C.1}$$

The $(n-1)$-dimensional volume of $T_n(y,\Delta)$ and $S_1(n,1)$ are written as $|T_n(y,\Delta)|$ and $|S_1(n,1)|$, respectively. Assume that $T_n(y,\Delta)$ does not intersect any edges of $S_1(n,1)$ according to [20].

A lower-bound of the ratio $|T_n(y,\Delta)|$ to $|S_1(n,1)|$ should be obtained to apply the same approach in the case of gaussian sources. If n is even, a complicated calculation yields the following lower-bound:

$$\frac{|T_n(y,\Delta)|}{|S_1(n,1)|} \geq \frac{(2m)!}{(m-1)!}\cdot\frac{\Delta^{2m-1}}{2^{3m-1}}, \tag{C.2}$$

where $n = 2m$. By taking the logarithm of the both hand sides of (C.2), (C.2) is evaluated in the following manner:

$$\begin{aligned}\frac{1}{2m}\log_2\frac{|T_n(y,\Delta)|}{|S_1(n,1)|} &\geq \frac{2m}{\log_2}\frac{(2m)!}{(m-1)!}+\frac{1}{2m}\log_2\frac{\Delta^{2m-1}}{2^{3m-1}}\\ &\geq \frac{1}{2m}\log_2(m+1)^{(m+1)}+\left(1-\frac{1}{2m}\right)\log_2\Delta-\frac{3}{2}+\frac{1}{2m}\\ &\to \infty \qquad (m\to\infty),\end{aligned} \tag{C.3}$$

which shows that the approach used in the case of gaussian breaks down. The failure of this approach results from the assumption that $T_n(y,\Delta)$ has no intersection with $S_1(n,1)$. This approach only clarify the importance to consider effect of edges of $S_1(n,\Delta)$ for encoding of memoryless Laplacian sources.

The remaining of this appendix is devoted to establishment of the lower bound given in (C.2). It is obvious that $|T_n(y,\Delta)|$ is equal to the $(n-1)$-dimensional volume of the intersection of

$$|x_1|+|x_2|+\cdots+|x_n| \leq \Delta \tag{C.4}$$

with

$$x_1+x_2+\cdots+x_n = 0. \tag{C.5}$$

Note that the set satisfying (C.4) is a convex set enclosed by 2^n hyperplanes, and 2^n-2 out of 2^n hyperplanes intersect (C.5). Since hyperplane (C.5) is convex, the intersection is also convex. Let $\beta = \frac{\Delta}{2}$. There are $n(n-1)$ points of intersection between edges of (C.4) and (C.5) whose component consist of one β, one $-\beta$ and $n-2$ zeros. Hence, the intersection between (C.4) and (C.5) is a convex-hull including all of these $n(n-1)$ points.

To evaluate a lower-bound of the $(n-1)$-dimensional volume of the convex-hull, the case $n = 6$ is considered as an example. When $n = 6$, (C.4) is written as

$$|x_1| + |x_2| + |x_3| + |x_4| + |x_5| + |x_6| \leq \Delta, \tag{C.6}$$

and (C.5) is written as

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 0. \tag{C.7}$$

The set given in (C.6) is enclosed by $2^6 = 64$ hyperplanes, and 62 out of 64 hyperplanes intersect with (C.7). Since only lower-bound of the five-dimensional volume of the intersection is needed, hyperplanes with three non-negative components and another three non-positive components such as

$$x_1 + x_2 + x_3 - x_4 - x_5 - x_6 = \Delta, \quad x_1, x_2, x_3 \geq 0,\ x_4, x_5, x_6 \leq 0 \tag{C.8}$$

are considered. There are $\binom{6}{3}$ choices for hyperplanes. Intersection between (C.7) and (C.8) includes nine points represented by the following vectors:

$$\begin{aligned}
\boldsymbol{a}_{1,1} &= (\beta, 0, 0, -\beta, 0, 0) \\
\boldsymbol{a}_{1,2} &= (\beta, 0, 0, 0, -\beta, 0) \\
\boldsymbol{a}_{1,3} &= (\beta, 0, 0, 0, 0, -\beta) \\
\boldsymbol{a}_{2,1} &= (0, \beta, 0, -\beta, 0, 0) \\
\boldsymbol{a}_{2,2} &= (0, \beta, 0, 0, -\beta, 0) \\
\boldsymbol{a}_{2,3} &= (0, \beta, 0, 0, 0, -\beta) \\
\boldsymbol{a}_{3,1} &= (0, 0, \beta, -\beta, 0, 0) \\
\boldsymbol{a}_{3,2} &= (0, 0, \beta, 0, -\beta, 0) \\
\boldsymbol{a}_{3,3} &= (0, 0, \beta, 0, 0, -\beta).
\end{aligned}$$

The five-dimensional volume of the convex-hull including above nine points and the origin should be evaluated. If five linearly independent vector out of nine are picked, a simplex is defined by the five vectors and the origin. Sum of the five-dimensional volume of all such simplexes is equal to the five-dimensional volume of the convex-hull.

The following procedure is used for picking up five linearly independent vectors. First, pick up all of three orthogonal vectors. There are totally 3! choices as follows:

$$\{a_{1,1}, a_{2,2}, a_{3,3}\}, \quad \{a_{1,1}, a_{2,3}, a_{3,2}\},$$
$$\{a_{1,2}, a_{2,1}, a_{3,3}\}, \quad \{a_{1,2}, a_{2,3}, a_{3,1}\},$$
$$\{a_{1,3}, a_{2,2}, a_{3,1}\}, \quad \{a_{1,3}, a_{2,1}, a_{3,2}\}.$$

Next, choose one vector arbitrarily out of three as a standard vector, and find equidistant vectors from the standard vector and another vectors. There are always two such equidistant vectors. In the case of $\{a_{1,1}, a_{2,2}, a_{3,3}\}$, suppose the case that $a_{3,3}$ is chosen as the standard vector. Two vectors $a_{1,3}$ and $a_{3,1}$ are equidistant from $a_{1,1}$ and $a_{3,3}$, and $a_{2,3}$ and $a_{3,2}$ is equidistant from $a_{2,2}$ and $a_{3,3}$. By selecting one vector out of two for each case, the following $2^2 = 4$ sets of five linearly independent vectors are obtained:

$$\begin{aligned} &\{a_{1,1}, a_{2,2}, a_{3,3}, a_{1,3}, a_{2,3}\}, \\ &\{a_{1,1}, a_{2,2}, a_{3,3}, a_{1,3}, a_{3,2}\}, \\ &\{a_{1,1}, a_{2,2}, a_{3,3}, a_{3,1}, a_{2,3}\}, \\ &\{a_{1,1}, a_{2,2}, a_{3,3}, a_{3,1}, a_{3,2}\}. \end{aligned} \tag{C.9}$$

Note that all of them define distinct simplexes.

The five-dimensional volume of the simplex defined by $\{a_{1,1}, a_{2,2}, a_{3,3}, a_{1,3}, a_{2,3}\}$ can be evaluated by calculating Gram matrix as follows:

$$\frac{1}{5!} \cdot \left| \det \begin{bmatrix} 2\beta^2 & 0 & 0 & \beta^2 & 0 \\ 0 & 2\beta^2 & 0 & 0 & \beta^2 \\ 0 & 0 & 2\beta^2 & \beta^2 & \beta^2 \\ \beta^2 & 0 & \beta^2 & 2\beta^2 & \beta^2 \\ 0 & \beta^2 & \beta^2 & \beta^2 & 2\beta^2 \end{bmatrix} \right|^{\frac{1}{2}} = \frac{\sqrt{6}}{5!}\beta^5 = \frac{\sqrt{6}}{5!}\left(\frac{\Delta}{2}\right)^5, \tag{C.10}$$

where definition of β is used to obtain the last equality. Another simplexes have the same volume since their Gram matrices can be made identical to the one shown in (C.10) by appropriate permutations of rows and columns of those matrices. Then, the five-dimensional volume of the intersection (C.6) with (C.7) is lower-bounded by

$$T_6(y, \Delta) \geq \binom{6}{3} \cdot 3! \cdot 3 \cdot 2^2 \cdot \frac{\sqrt{6}}{5!} \cdot \left(\frac{\Delta}{2}\right)^5. \tag{C.11}$$

In the case that n is even. evaluation of the $(n-1)$-dimensional volume of the intersection between (C.4) and (C.5) is the same. Let $n = 2m$. There are $\binom{2m}{m}$ choices of hyperplanes, $m!$ choices for m orthogonal vectors, m choices for a standard vector, and 2^{n-1} choices for equidistant vectors. All of Gram matrices are congruent to the following matrix:

$$\begin{bmatrix} 2I_{m-1} & I_{m-1} \\ I_{m-1} & K_m \end{bmatrix} \cdot \left(\frac{\Delta}{2}\right)^{2m-2}, \tag{C.12}$$

where I_{m-1} is the $(m-1)$-th unit matrix and K_m is the $m \times m$ matrix whose all diagonal components are equal to 2 and all non-diagonal components are equal to 1. It is easy to check that verify

$$\det \begin{bmatrix} 2I_{m-1} & I_{m-1} \\ I_{m-1} & K_m \end{bmatrix} = 2m. \tag{C.13}$$

Hence, the $(2m-1)$-dimensional volume of the intersection of (C.4) with (C.5) is lower-bounded as follows:

$$|T_{2m}(\boldsymbol{y}, \Delta)| \geq \binom{2m}{m} \cdot m! \cdot m \cdot 2^{m-1} \cdot \frac{\sqrt{2m}}{(2m-1)!} \cdot \left(\frac{\Delta}{2}\right)^{2m-1}, \tag{C.14}$$

which implies (C.2).

Bibliography

[1] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, 1968.

[2] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, New Jersey, 1971.

[3] I. Csiszár and J. Körnor, *Information Theory: Coding Theorem for Discrete Memoryless Systems*, Academic Press, 1981.

[4] R. E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, 1987.

[5] J. Thomas and T. Cover, *Elements of Information Theory*, Wiley, New York, 1991.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 379–423, 623–656, 1948.

[7] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. on Information Theory*, vol. IT-23, no. 3, pp. 337–343, 1977.

[8] C. E. Shannon, "Coding theorems for a discrete source with fidelity criterion," *IRE National Convention Record, Part 4*, pp. 142–163, 1959.

[9] D. J. Sakrison, "A geometric treatment of the source encoding of a gaussian random variable," *IEEE Trans. on Information Theory*, vol. IT-14, no. 3, pp. 481–486, 1968.

[10] A. Buzo Y. Linde and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, vol. Com-28, no. 1, pp. 84–95, 1980.

[11] T. Lookabaugh P. A. Chou and R. M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. on Information Theory*, vol. IT-35, no. 1, pp. 1989, 299–315.

[12] T. Lookabaugh P. A. Chou and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-37, no. 1, pp. 31–42, 1989.

[13] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Information Theory*, vol. IT-25, no. 4, pp. 373–380, 1979.

[14] Y. Yamada and S. Tazaki, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Trans. on Information Theory*, vol. IT-26, no. 1, pp. 6–14, 1980.

[15] A. L. Zador, "Asymptotic quantization error of continuous signal and the quantization dimension," *IEEE Trans. on Information Theory*, vol. IT-28, no. 2, pp. 139–149, 1982.

[16] H. Harashima T. Saitoh, T. Hideya and H. Miyakawa, "Gain/shape vector quantization for multidimensional spherically symmetric random source," *IEICEJ Trans.*, vol. J68-B, no. 8, pp. 904–911, 1985.

[17] P. Noll and R. Zelinski, "Bounds on quantizer performance in the Low Bit-Rate Region," *IEEE Trans. on Communication*, vol. COM-26, no. 2, pp. 300–304, 1978.

[18] P. F. Swaszek, "A vector quantizer for the Laplace source," *IEEE Trans. on Information Theory*, vol. IT-37, no. 5, pp. 481–486, 1991.

[19] S. A. Townes and J. B. O'neal Jr., "Permutation codes for Laplacian source," *IEEE Trans. on Information Theory*, vol. IT-30, no. 3, pp. 553–559, 1984.

[20] T. R. Fischer, "A pyramid vector quantizer," *IEEE Trans. on Information Theory*, vol. IT-32, no. 4, pp. 568–583, 1986.

[21] T. Berger, "Permutation entropy quantizers and permutation codes," *IEEE Trans. on Information Theory*, vol. IT-28, no. 2, pp. 149–157, 1982.

[22] J. Ziv and A. Lempel, "Compression of individual sequence via valuable-rate coding," *IEEE Trans. on Information Theory*, vol. IT-24, no. 5, pp. 530–536, 1978.

[23] F. M. Willems, "Universal data compression and repetition times," *IEEE Trans. on Information Theory*, vol. 1, no. IT-35, pp. 54–58, 1989.

[24] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. on Information Theory*, vol. 6, no. IT-35, pp. 1250–1258, 1989.

[25] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. on Information Theory*, vol. IT-39, no. 1, pp. 78–83, 1994.

[26] Y. Steinberg and M. Gutman, "An algorithm for source coding subject to a fidelity criterion, based on string matching," *IEEE Trans. Information Theory*, vol. IT-39, no. 3, pp. 877–886, 1993.

[27] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement," in *Proceedings of IEEE Information Theory Workshop*, 1993, pp. 16–17.

[28] G. Lugosi T. Linder and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. on Information Theory*, , no. to appear Nobember, 1994.

[29] M. Effros P. A. Chou and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," submitted to *IEEE Trans. on Information Theory*.

[30] J. Ziv, "Coding of sources with unknown statistics –Part I: Probability of encoding error," *IEEE Trans. on Information Theory*, vol. IT-18, no. 3, pp. 384–389, 1972.

[31] J. Ziv, "Coding of sources with unknown statistics –Part II: Distortion relative to a fidelity criterion," *IEEE Trans. on Information Theory*, vol. IT-18, no. 3, pp. 389–394, 1972.

[32] R. M. Gray D. L. Neuhoff and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. on Information Theory*, vol. IT-21, no. 5, pp. 511–523, 1975.

[33] F. Kanaya and J. Muramatsu, "An almost sure recurrence theorem with distortion for stationary ergodic sources," in *The Proceedings of The 17th Symposium on Information Theory and Its Applications*, 1994, pp. 517–520.

[34] H. Koga and S. Arimoto, "Asymptotic optimality of modified spherical codes with scalar quantization of gain for memoryless gaussian source," *IEICE Trans. on Fundamentals*, vol. E-76-A, no. 9, pp. 1403–1410, 1993.

[35] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. on Information Theory*, vol. IT-18, no. 4, pp. 460–473, 1972.

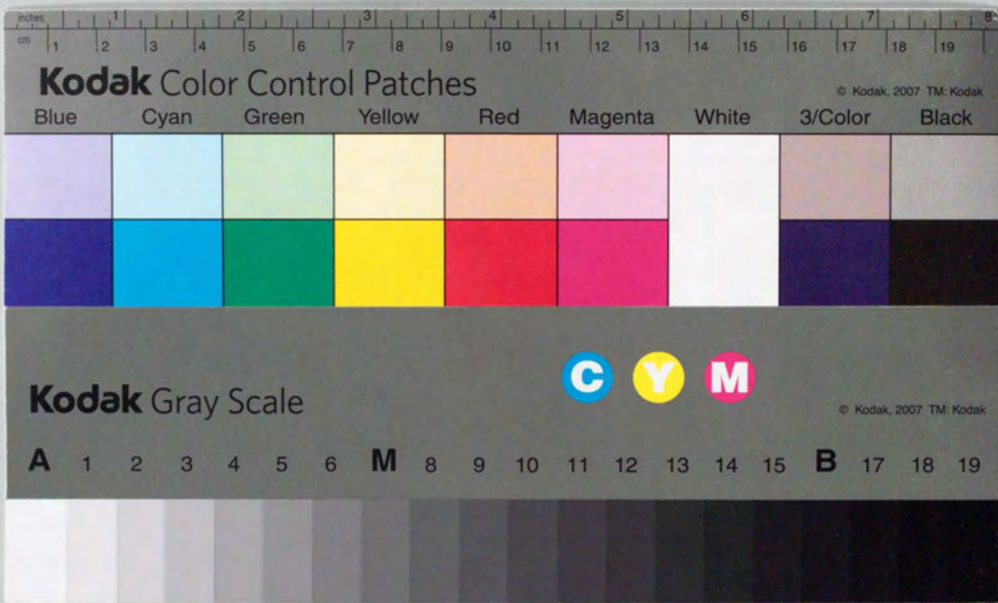
[36] S. Arimoto, "An algorithm for computing the capacity of arbitrarily discrete memoryless channel," *IEEE Trans. on Information Theory*, vol. IT-18, no. 1, pp. 14–20, 1972.

[37] L. G. Valiant, "A theory of the learnable," *Communication of the ACM*, vol. 27, pp. 1134–1142, 1984.

[38] M. Kearns and R. Schapire, "Efficient distribution-free learning of probabilistic concepts," in *Proceedings of IEEE Symposium on Foundations of Computer Science*, 1990, pp. 382–391.

[39] N. Abe and M. K. Warmuth, "On the computational complexity of approximating probability distrubutions by probabilistic automata," in *Proceedings of 3rd Workshop on Computational Learning Theory*, 1990, pp. 52–66.

[40] K. Yamanishi, "A learning criterion for stochastic rule," *Machine Learning*, vol. 9, pp. 165–203, 1992.


inches 1 2 3 4 5 6 7 8
cm 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue Cyan Green Yellow Red Magenta White 3/Color Black
C Y M
Kodak Gray Scale
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19