The Dissertation Committee for Hongbo Si
certifies that this is the approved version of the following dissertation:

Coding Mechanisms for Communication and Compression: Analysis of Wireless Channels and DNA Sequencing

Committee:

Sriram Vishwanath, Supervisor

Alexandros G. Dimakis

O. Ozan Koyluoglu

Haris Vikalo

Felipe Voloch

Coding Mechanisms for Communication and Compression: Analysis of Wireless Channels and DNA Sequencing

by

Hongbo Si, B.S., M.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Dedicated to my parents and my wife.

Acknowledgments

I feel extremely lucky in my years as a graduate student in UT Austin, as I am blessed with help and support from many people.

First and foremost, I express my deepest gratitude to my supervisor. The best thing about my PhD experience is having Professor Sriram Vishwanath as my supervisor, since you cannot ask for a better one. Through the past five years, he has guided me with great patience and enthusiasm, and offered me support and advice on every aspect of my research and career. His broad interests, technical strength and deep insight have immensely affected my development as an academic, and will remain beneficial for my whole life.

Professor Haris Vikalo has been somewhat of a second advisor for me. I am grateful to him for his insightful knowledge and invaluable discussion. Most importantly, he has always encouraged me and pointed me to correct directions. We worked together on many different bio-informatics problems, mainly on haplotyping, and I have learned a great deal from our collaborations, which helped produce a significant portion of this thesis.

My heartfelt thanks also go to Professor O. Ozan Koyluoglu, who has transformed from a student to a professor during my PhD. He is a great collaborator, mentor and friend. He is expert, thorough and careful for every detail of research, and I enjoyed every discussion with him. He was always

able to provide a fresh and critical viewpoint, which helped to force me push forward.

I am also very fortunate and honored to have Professor Alexandros G. Dimakis and Professor Felipe Voloch as my committee members. I still remember how impressed I was when taking their courses, and remember their invaluable comments and suggestions which remarkably improved the quality of my research.

I thank my colleagues of LINC and WNCG. It is hard to imagine how to complete my PhD without the generous help from everyone. Their excellent research work set great examples for me, and their collaborations lead me out of the academic dilemmas.

My life would never have been so exciting without my dear friends in Austin. I wholehearted appreciate their help and kindness in the past years, brightening my life like a rainbow. After many years, I may still take delight in talking about every moment with laughs and tears.

Last, but definitely not least, I would like to thank my parents and my wife, for everything. They have been giving me unconditional support and constant encouragement through these years, for which I will be forever grateful. No words can describe my love for them.

Coding Mechanisms for Communication and Compression: Analysis of Wireless Channels and DNA Sequencing

Hongbo Si, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Sriram Vishwanath

This thesis comprises of two related but distinct components: Coding arguments for communication channels and information-theoretic analysis for haplotype assembly. The common thread for both problems is utilizing information and coding theoretic principles in understanding their underlying mechanisms.

For the first class of problems, I study two practical challenges that prevent optimal discrete codes utilizing in real communication and compression systems, namely, coding over analog alphabet and fading. In particular, I use an expansion coding scheme to convert the original analog channel coding and source coding problems into a set of independent discrete subproblems. By adopting optimal discrete codes over the expanded levels, this low-complexity coding scheme can approach Shannon limit perfectly or in ratio. Meanwhile, I design a polar coding scheme to deal with the unstable state of fading channels. This novel coding mechanism of hierarchically utilizing different types of polar

codes has been proved to be ergodic capacity achievable for several fading systems, without channel state information known at the transmitter.

For the second class of problems, I build an information-theoretic view for haplotype assembly. More precisely, the recovery of the target pair of haplotype sequences using short reads is rephrased as the joint source-channel coding problem. Two binary messages, representing haplotypes and chromosome memberships of reads, are encoded and transmitted over a channel with erasures and errors, where the channel model reflects salient features of high-throughput sequencing. The focus is on determining the required number of reads for reliable haplotype reconstruction.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis deals with two application domains: communication over channels and DNA sequencing. Although at first glance, these two domains appear unrelated, the tools we utilize to analyze them are similar: using information and coding theoretic principles in understanding their underlying mechanisms.

1.1 On the Way to Practical Coding Schemes

The field of information theoretical study on data transmission and compression is started with Shannon's famous theorem proposed in 1948 [1], which says channel capacity is the tightest upper bound on the amount of information that can be reliably transmitted over a noisy communication channel, and rate distortion limit is the tightest lower bound on the amount of information that can be compressed from a source within particular distortion constraint. After that, seeking for practical coding schemes that could approach channel capacity or rate distortion limit became a central objective for researchers. On the way from theory to practice, many coding schemes are proposed. Different types of codes emerge in improving the performance,

giving consideration to the trade-off between coding complexity and rate performance.

The history of channel coding traces back to the era of algebraic coding, including the well-known Hamming codes [2], Golay codes [3], Reed-Muller codes [4][5], Reed-Solomon codes [6], lattice codes [7], and others [8]. However, although making significant achievements, algebraic coding still could not prove to be the way to approach the Shannon limit. The next era of probabilistic coding concerned more with optimizing performance as a function of coding complexity. This line of development included convolutional codes [9], and concatenated codes [10] at earlier times, as well as turbo codes [11] and low-density parity-check (LDPC) codes [12][13] afterwards. Recently, a new class of block codes, polar codes [14] (also see Appendix 1.A for a brief introduction), has been proved to achieve Shannon limit of symmetric binary-input discrete memoryless channels (B-DMC) with low encoding and decoding complexity. In another recent study [15][16], a new type of rateless code, spinal codes, is proposed to achieve the capacity of binary symmetric channels.

For discrete-valued "finite-alphabet" source coding problems, the associated coding theorem [17] and practically-meaningful coding schemes are well known. Trellis based quantizers [18] are the first to achieve the rate distortion trade-off, but with encoding complexity scaling exponentially with the constraint length. Later, Matsunaga and Yamamoto [19] show that a low density parity check (LDPC) ensemble, under suitable conditions on ensemble structure, can achieve the rate distortion bound using an optimal decoder. Af-

ter that, [20] shows that low density generator matrix (LDGM) codes, as the dual of LDPC codes, with suitably irregular degree distributions, empirically perform close to the Shannon rate-distortion bound with message-passing algorithms. More recently, polar codes [14], are the first provably rate distortion limit achievable codes with low encoding and decoding complexity [21][22][23].

However, when utilizing these optimal codes to a practical communication or compression system, new challenges come out. In general, these aforementioned optimal codes are designed on a discrete alphabet, but in practice, the channels and sources we meet are always continuous-valued. Hence, how to design practical coding schemes to bridge the gap from discrete alphabet to analog alphabet is totally nontrivial and meaningful. Besides, in practical communication systems, like wireless systems, channel state is not stable due to the influence of environment. To this end, fading always exists, and the variety of channel states impacts the performance of optimal codes, which forms the second challenge for practical utilization of theoretically optimal codes.

1.2 Information-Theoretical Analysis of DNA Sequencing

A novel branch of information theory is to investigate its application to other research domains, utilizing its inherent ability to reveal the information principle. DNA sequencing and haplotyping are typical instances of those domains.

Current DNA sequencing technology, i.e., next generation sequencing,

aims to enable fast and affordable sequencing tasks. However, various imperfections and uncertainties in the processes employed by the current technologies impose limitations on the achievable read lengths. The read lengths provided by next generation sequencing platforms are much shorter than those provided by the conventional Sanger method [24], while the error rates are higher. In shotgun sequencing [25], a long sequence of DNA is broken up randomly into numerous small reads. Computer programs then use the information of overlapping ends from different reads to assemble them into a continuous sequence [26]. At this point, there is an inherent redundancy due to sequencing a base position multiple times as parts of different reads. This naturally motivates our interest to know how much redundancy is required to obtain a certain level of accuracy, where the accuracy level requirement is often dictated by the downstream application [27].

One of the recently developed genomic techniques based on new generation DNA sequencing is haplotype assembly. A haplotype is the collection of SNPs on a single chromosome within a homologous pair from diploid organisms, and it is believed to contain essential genomic information determining the characteristic and diseases. However, direct analysis and identification of a haplotype is generally challenging, costly, and time and labor intensive. Alternatively, single individual haplotypes can be assembled from short reads provided by high-throughput DNA sequencing systems.

To this end, DNA sequencing and haplotype assembly problems essentially aim to determine the particular order of nucleotides or single nucleotide

polymorphisms, which can be naturally included into the study of information theory and coding theory. Information theoretic view will give a much better explanation of how to find an algorithm free analysis, and reveal the relationship between minimum coverage number and recovery accuracy, while coding scheme will provide a wider horizon for exploring advanced sequencing and assembling schemes.

1.3 Main Contributions of the Thesis

1.3.1 Part I: Expansion Coding for Data Transmission and Compression

A general method of coding over expansion is proposed, which allows one to reduce the highly non-trivial problems of coding over analog channels and compressing analog sources to a set of much simpler subproblems, coding over discrete channels and compressing discrete sources. More specifically, our focus is on coding over additive exponential noise (AEN) channels, and lossy compression of exponential and Laplacian sources. Due to the essential decomposable property of these channels and sources, the proposed expansion method allows for mapping of these problems (either perfectly or approximately) to coding over parallel subproblems, where each level is modeled as an independent coding problem over discrete alphabets. Any feasible solution to the optimization problem after expansion corresponds to an achievable scheme of the original problem. In this mapping, for the cases where finding the optimal solutions is hard to characterize, it is shown that expansion cod-

ing scheme still presents a good performance by specific choices of parameters. More specifically, theoretical analysis and numerical results reveal that expansion coding achieves the capacity of AEN channel in the high SNR regime. It is also shown that for lossy compression, the achievable rate distortion pair by expansion coding approaches the Shannon limit in the low distortion region. Remarkably, by using capacity-achieving codes with low encoding and decoding complexity that are originally designed for discrete alphabets, for instance polar codes, the proposed expansion coding scheme allows for designing low-complexity codes for analog channel coding and source coding.

1.3.2 Part II: Hierarchical Polar Coding Scheme for Fading Channels

The main contribution of this part is to propose polar coding schemes for fading channels. More specifically, the focus is on fading binary symmetric channels, fading additive exponential noise channels, as well as fading wiretap channels.

For fading binary symmetric channels, to overcome the variability of channel states, a coding scheme of hierarchically employing polar codes is proposed. For a two state binary symmetric channel, a polar code designed for the superior fading state is used for each fading block. The receiver, utilizing its channel state information, declares a channel output as erasure whenever its corresponding fading block is at degraded state and its corresponding channel index is a good state of the underlying polar code. Then, an outer polar code,

designed for the corresponding erasure model, is utilized to recover from these erasures by coding over fading blocks. This scheme is generalized to fading scenarios with multiple channel states, and it is shown that the proposed coding scheme, without instantaneous channel state information at the transmitter, achieves the capacity of the fading binary symmetric channel.

For fading additive exponential noise channels, expansion coding is used to convert the problem of coding over these analog fading channels into coding over discrete fading channels. The previously proposed hierarchical polar coding approach is then adopted to resolve these discrete coding problems. Theoretical analysis and numerical results are given, showing that the proposed scheme achieves the ergodic capacity of fading additive exponential noise channel in the high SNR regime.

For fading wiretap channels, a polar coding scheme is proposed to achieve reliability as well as security. Specifically, a block fading model is considered for the wiretap channel that consists of a transmitter, a receiver, and an eavesdropper; and only the information regarding the statistics of the channel state information is assumed at the transmitter. For this model, the aforementioned hierarchical polar coding scheme is combined with the existing polar coding scheme to guarantee security. Message bits are transmitted such that they may be reliably decoded at the receiver, and random bits are introduced to exhaust the leakage seen by the eavesdropper. It is found that this coding scheme is secure capacity achieving for the corresponding fading binary symmetric wiretap channel.

Overall, utilizing polar codes in such a novel (hierarchical) way enables coding without the knowledge of instantaneous channel state information at the transmitter, a practically important scenario in wireless systems.

1.3.3 Part III: Information-Theoretic Analysis for Haplotype Assembly

An information-theoretic analysis is proposed for the haplotype assembly problem. A haplotype is a sequence of nucleotide bases on a chromosome that differ from the bases in the corresponding positions on the other chromosome in a homologous pair. Haplotypes of diploids are typically bi-allelic and hence may conveniently be represented by binary strings. Information about the order of bases in a genome is readily inferred using short reads provided by high-throughput DNA sequencing technologies. Associating reads that cover variant positions with specific chromosomes in a homologous pairs, which enables haplotype assembly, is challenging due to limited lengths of the reads and presence of sequencing errors.

From the view of information theory, the recovery of the target pair of haplotype sequences using short reads is rephrased as the joint source-channel coding problem. Two binary messages, representing haplotypes and chromosome memberships of reads, are encoded and transmitted over a channel with erasures and errors, where the channel model reflects salient features of high-throughput sequencing. The focus here is on determining the required number of reads for reliable haplotype reconstruction. Both the necessary and

sufficient conditions are presented with order-wise optimal bounds.

1.4 Organization of the Thesis

The proposed research is covered from Chapter 2 to Chapter 8. Chapter 2 and 3 present the expansion coding schemes for data transmission and compression respectively. Chapter 4, 5 and 6 present the hierarchical polar coding scheme for fading channels, where the focus ranges from fading binary symmetric channels, to fading additive exponential noise channels, and then to fading wiretap channels. Chapter 7 presents the information-theoretic analysis of haplotype assembly problem. Finally, Chapter 8 concludes the thesis.

1.5 Notations

n and N are both examples of scalars. $a_{1:n}$ or $\boldsymbol{a} = (a_1, \ldots, a_n)$ is a vector with length n. $\boldsymbol{A}$ is a matrix and, a_{ij} is its i-th row and j-th column element. $\mathcal{A}$ is a set, and $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. X is a random variable, and $\mathsf{X}_{1:n}$ is a random vector. $\mathbf{1}_{\{\cdot\}}$ is the indicator function. $\Pr\{\cdot\}$ is the probability measure and $\mathbb{E}[\cdot]$ is the expectation. $H(\cdot)$ is the discrete entropy (in bits), and $h(\cdot)$ is the differential entropy (in bits).

1.A Introduction to Polar Codes

The construction of polar code is based on a phenomenon referred to as *channel polarization*. Consider a binary-input discrete memoryless channel $\mathcal{W}_{\text{B-DMC}} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \{0, 1\}$. Define

$$\boldsymbol{F} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Let $\boldsymbol{B}_N$ be the bit-reversal operator as defined in [14], where $N = 2^n$. By applying the transform $\boldsymbol{G}_N = \boldsymbol{B}_N \boldsymbol{F}^{\otimes n}$ ($\boldsymbol{F}^{\otimes n}$ denotes the n^{th} Kronecker product of $\boldsymbol{F}$) to $u_{1:N}$, the encoding is given by $x_{1:N} = u_{1:N} \cdot \boldsymbol{G}_N$, which is transmitted through N independent copies of $\mathcal{W}_{\text{B-DMC}}$. Now, consider N binary-input coordinate channels $\mathcal{W}_N^{(i)} : \mathcal{X} \to \mathcal{Y}^N \times \mathcal{X}^{i-1}$, where, for $i \in \{1, \ldots, N\}$, the transition probability is given by

$$\mathcal{W}_N^{(i)}(y_{1:N}, u_{1:i-1}|u_i) \triangleq \sum_{u_{i+1:N}} \frac{1}{2^{N-1}} \mathcal{W}_{\text{B-DMC}}^N(y_{1:N}|u_{1:N} \cdot \boldsymbol{G}_N).$$

Remarkably, as $N \to \infty$, the channels $\mathcal{W}_N^{(i)}$ polarize to either noiseless or pure-noisy, and the fraction of noiseless channels is close to $I(\mathcal{W}_{\text{B-DMC}})$, the symmetric capacity of channel $\mathcal{W}_{\text{B-DMC}}$ [14].

Given this polarization phenomenon, polar codes can be considered as $\boldsymbol{G}_N$-coset codes with parameters $(N, K, \mathcal{A}, u_{\mathcal{A}^c})$, where $u_{\mathcal{A}^c} \in \mathcal{X}^{N-K}$ is frozen vector (can be set to all-zeros for binary symmetric channels [14]), and the information set $\mathcal{A}$ is chosen as a K-element subset of $\{1, \ldots, N\}$ such that the Bhattacharyya parameters satisfy $Z(\mathcal{W}_N^{(i)}) \leq Z(\mathcal{W}_N^{(j)})$ for all $i \in \mathcal{A}$ and $j \in \mathcal{A}^c$, i.e., $\mathcal{A}$ denotes *good* channels (that are noiseless in the limit). We use

permutations (namely, π and ϕ in the sequel) to denote the increasing order of Bhattacharyya parameter values for the polarization of underlying channels. (For instance, for block length N, $\pi(1)$ gives the most reliable polarized channel index.)

A decoder for a polar code is the successive cancelation (SC) decoder, which gives an estimate $\hat{u}_{1:N}$ of $u_{1:N}$ given knowledge of $\mathcal{A}$, $u_{\mathcal{A}^c}$, and $y_{1:N}$ by computing

$$\hat{u}_i = \begin{cases} 1, & \text{if } i \in \mathcal{A}, \text{ and } \frac{\mathcal{W}_N^{(i)}(y_{1:N}, \hat{u}_{1:i-1}|1)}{\mathcal{W}_N^{(i)}(y_{1:N}, \hat{u}_{1:i-1}|0)} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

in the order i from 1 to N. It has been shown that, by adopting an SC decoder, polar codes achieve any rate $R < I(\mathcal{W}_{\text{B-DMC}})$ with a decoding error scaling as $O(2^{-N^\beta})$, where $\beta < 1/2$. Moreover, the complexity for both encoding and decoding is $O(N \log N)$.

Chapter 2

Expansion Coding for Data Transmission

2.1 Problem Background and Related Work

The problem of coding over analog noise channels is highly non-trivial in general. To this end, a method of modulation is commonly utilized to map discrete inputs to analog signals for transmission through the physical channel [28]. In this paper, we focus on designing and coding over such mappings. In particular, we propose a new coding scheme for general analog channels with moderate coding complexity based on an expansion technique, where channel noise is perfectly or approximately represented by a set of independent discrete random variables (see Figure 2.1). Via this representation, the problem of coding over an analog noise channel is reduced to that of coding over parallel discrete channels. We focus on additive exponential noise (AEN) channels, and we show that the Shannon limit, i.e., the capacity, is achievable in the high SNR regime. The main advantage of the proposed method lies on its complexity inheritance property, where the encoding and decoding complexities of the proposed schemes follow that of the embedded capacity achieving codes designed for discrete channels, for instance polar codes and spinal codes.

Multilevel coding is a general coding method designed for analog noise

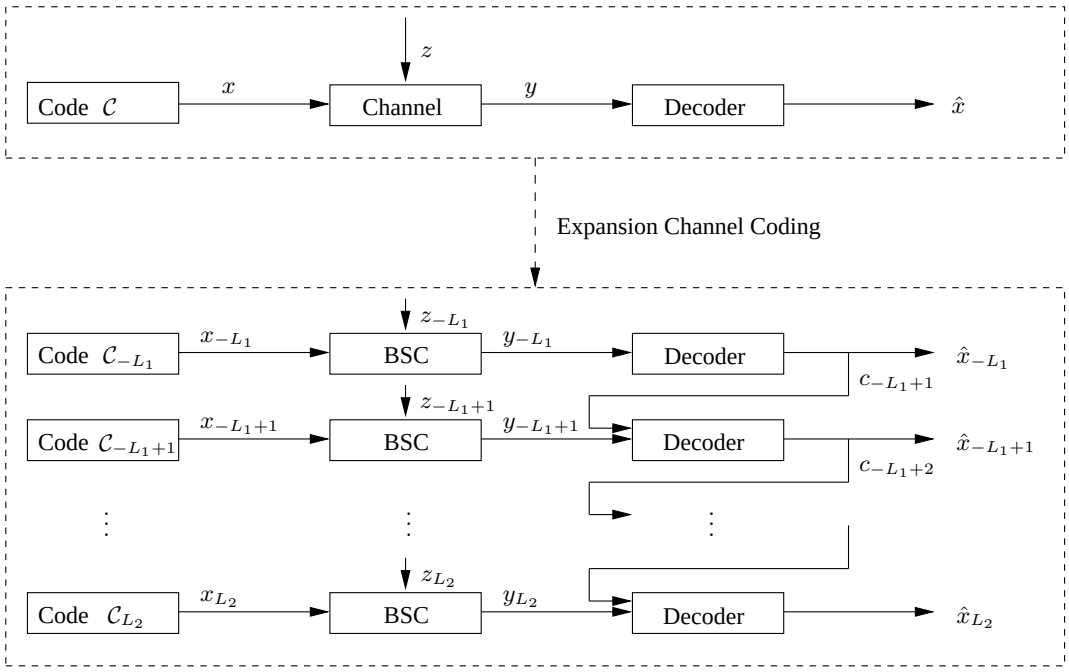

Figure 2.1: **Illustration of expansion channel coding framework.** An analog noise channel is expanded into a set of discrete channels with index from $-L_1$ to L_2. Channel noise is considered as its binary expansion $z = \cdots z_{L_2} \cdots z_1 z_0 . z_{-1} \cdots z_{-L_1} \cdots$, and similar expansions are adopted to channel input and output. Carries exist between neighboring levels.

channel with a flavor of expansion [29]. In particular, a lattice partition chain $\Lambda_1/\cdots/\Lambda_{r-1}/\Lambda_r$ is utilized to represent the channel input, and together with a shaping technique, the reconstructed codeword is transmitted to the channel. It has been shown that optimal latices achieving Shannon limit exist, however, the encoding and decoding complexity is high in general. In the sense of representing the channel input, our scheme is coincident with multilevel coding by choosing $\Lambda_1 = q^{-L_1}\mathbb{Z}, \ldots, \Lambda_r = q^{L_2}\mathbb{Z}$, for some $L_1, L_2 \in \mathbb{Z}^+$, where coding of each level is over q-ary finite field (see Figure 2.2). The difference in the

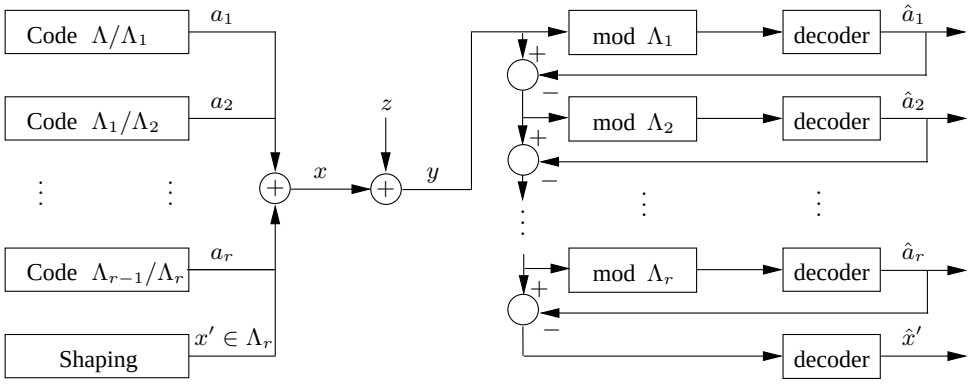

Figure 2.2: **Illustration of multilevel coding framework.** In this example, multilevel coding scheme is illustrated. Comparing to expansion coding in Figure 2.1, only channel input is expressed by multi-levels, but not the channel noise.

proposed method is that besides representing the channel input in this way, we also "expand" the channel noise, such that the coding problem for each level is more suitable to solve by adopting existing discrete coding schemes with moderate coding complexity. Moreover, by adapting the underlying codes to the channel realization dependent variables, such as carries, the Shannon limit is shown to be achievable by expansion coding with moderate number of expanded levels.

Deterministic model, proposed in [30], is another framework to study analog noise channel coding problems, where the basic idea is to construct an approximate channel for which the transmitted signals are assumed to be noiseless above the noise level. This approach is proved to be very effective in analyzing the capacity of networks. In particular, it has been testified that this framework perfectly represents and helps to characterize degrees of freedom of

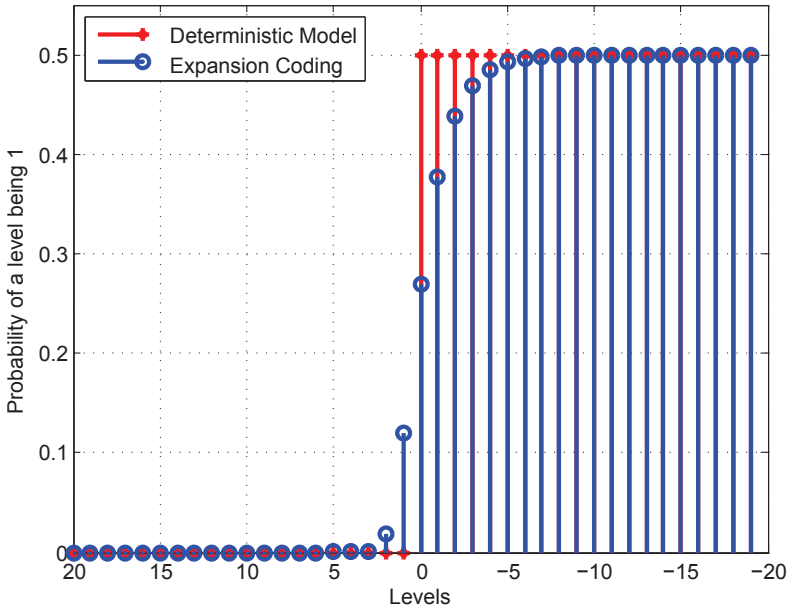

Figure 2.3: **Comparison of noise models between expansion coding and deterministic model.** The noise models of each level for expansion coding and deterministic model are illustrated. Deterministic model cut the noise to a certain level, and expansion coding has a smooth transaction regime.

point-to-point AWGN channels, as well as some multi-user channels of concern. In this sense, our expansion coding scheme can be seen as a generalization of these deterministic approaches. Here, the effective noise in the channel is carefully calculated and the system takes advantage of coding over the noisy levels at any SNR. This generalized channel approximation approach can be useful in reducing the large gaps reported in the previous works, because the noise approximation in our work is much closer to the actual distribution as compared to that of the deterministic model (see Figure 2.3).

There have been many attempts in utilizing discrete codes for analog

channels (beyond simple modulation methods). For example, after the introduction of polar codes, a considerable attention is directed towards utilizing their low complexity property for analog channel coding. A very straightforward way is by central limit theorem, which says certain combination of i.i.d. discrete random variables tends to Gaussian distribution. As reported in [16] and [31], the capacity of AWGN channel can be achieved by coding over large number of BSCs, however, the convergence rate is linear, which limits its application in practice. To this end, [32] proposes a MAC based scheme to improve the rate to exponential, at the expense of having a much larger field size. A newly published result in [33] attempts to combine polar codes with multilevel coding, however many aspects of this optimization of polar-coded modulation still remain open. Along the direction of this research, we also try to utilize capacity achieving discrete codes to approximately achieve the capacity of analog channels.

2.2 Expansion Channel Coding: Theoretical Framework

In general, expansion channel coding is a scheme of reducing the problem of coding over a analog channel to coding over a set of discrete channels. In particular, we consider the additive noise channel given by

$$\mathsf{Y}_i = \mathsf{X}_i + \mathsf{Z}_i,\ i = 1, \cdots, N, \tag{2.1}$$

where X_i are channel inputs with alphabet $\mathcal{X}$ (possibly having channel input requirements, such as certain moment constraints); Y_i are channel outputs; Z_i

are additive noises independently and identically distributed with continuous probability density function; N is blocklength.

When communicating, the transmitter conveys one of the messages, M, which is uniformly distributed in $\mathcal{M} \triangleq \{1, \cdots, 2^{NR}\}$; and it does so by mapping the message to the channel input using encoding function $\psi(\cdot) : \mathcal{M} \to \mathcal{X}^N$ such that $\mathsf{X}_{1:N}(\mathsf{M}) = \psi(\mathsf{M})$. The decoder uses the decoding function $\varphi(\cdot)$ to map its channel observations to an estimate of the message. Specifically, $\varphi(\cdot) : \mathcal{Y}^N \to \mathcal{M}$, where the estimate is denoted by $\hat{\mathsf{M}} \triangleq \varphi(\mathsf{Y}_{1:N})$. A rate R is said to be achievable, if the average probability of error defined by

$$P_{\mathrm{e}} \triangleq \frac{1}{|\mathcal{M}|} \sum_{\mathsf{M} \in \mathcal{M}} \Pr\{\hat{\mathsf{M}} \neq \mathsf{M} | \ \mathsf{M} \text{ is sent.}\}$$

can be made arbitrarily small for large N. The capacity of this channel is denoted by C, which is the maximum achievable rate R, and its corresponding optimal input distribution is denoted as $f_{\mathsf{X}}^*(x)$.

Our proposed coding scheme is based on the idea that by "expanding" the channel noise (i.e., representing it by its q-ary expansion), an approximate channel can be constructed, and proper coding schemes can be adopted to each level in this representation. If the approximation is close enough, then the coding schemes that are optimal for each level can be translated to an effective one for the original channel. More formally, consider the original noise Z and its approximation $\hat{\mathsf{Z}}$, which is defined by the truncated q-ary expansion of Z. For this moment, we simply take $q = 2$ (i.e., considering binary expansion),

and leave the general case for later discussion.

$$\hat{\mathsf{Z}} \triangleq \mathsf{Z}^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \mathsf{Z}_l,$$

where Z^{sign} represents the sign of Z, taking a value from $\{-,+\}$; Z_l's are mutually independent Bernoulli random variables. By similarly expanding the channel input, we convert the problem of coding over analog channels to coding over a set of binary discrete channels. This mapping is highly advantageous, as capacity achieving discrete codes can be adopted for coding over the constructed binary channels. Assume the input distributions for sign channel and discrete channel at l are represented by X^{sign} and X_l correspondingly, then an achievable rate (via random coding) for the approximated channel is given by

$$\hat{R} \triangleq I(\hat{\mathsf{X}}; \hat{\mathsf{X}} + \hat{\mathsf{Z}}),$$

where

$$\hat{\mathsf{X}} \triangleq \mathsf{X}^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \mathsf{X}_l.$$

By adopting the same coding scheme over the original channel, one can achieve a rate given by

$$R \triangleq I(\hat{\mathsf{X}}; \hat{\mathsf{X}} + \mathsf{Z}).$$

The following result provides a theoretical basement for expansion coding. (Here, $\xrightarrow{d.}$ denotes convergence in distribution.)

Theorem 2.1. *If* $\hat{\mathsf{Z}} \xrightarrow{d.} \mathsf{Z}$ *and* $\hat{\mathsf{X}} \xrightarrow{d.} \mathsf{X}^*$, *as* $L_1, L_2 \to \infty$, *where* $\mathsf{X}^* \sim f^*_{\mathsf{X}}(x)$, *i.e., the optimal input distribution for the original channel, then* $R \to C$.

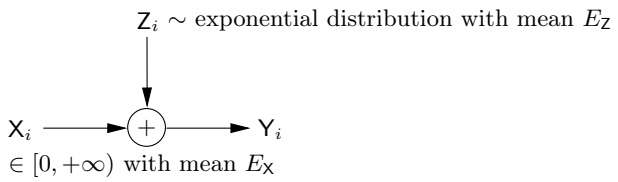

Figure 2.4: **System model for AEN channel.** The noise of AEN channel is independent and distributed as exponential distribution, and channel input is restricted to positive with mean constraint.

The proof of this theorem follows from the continuity property of mutual information. In words, if the approximation channel is close to original channel, and the distribution we adopt is close to optimal input distribution, then expansion coding scheme could achieve the capacity of the channel in concern.

2.3 Expansion Channel Coding: AEN Channel

The particular channel example considered in this section is additive exponential noise (AEN) channel, which models worst-case noise given a mean and a non-negativity constraint on noise [34]. In addition, the AEN model naturally arises in non-coherent communication settings, and in optical communication scenarios. (We refer to [34] and [35] for an extensive discussion on the AEN channel.)

Martinez proposed the pulse energy modulation scheme in [35], which can be seen as a generalization of amplitude modulation for the Gaussian channels. In this scheme, the constellation symbols are chosen as $c(i-1)^l$, for

$i = 1, \cdots, 2^M$ with a constant c, and it is shown that the information rates obtained from this constellation can achieve an energy (SNR) loss of 0.76 dB (with the best choice of $l = \frac{1}{2}(1 + \sqrt{5})$) compared to the capacity in the high SNR regime. Another constellation technique for this coded modulation approach is recently considered in [36], where it is shown that log constellations are designed such that the real line is divided into $(2M - 1)$ equally probable intervals. M of the centroids of these intervals are chosen as constellation points, and, by a numerical computation of the mutual information, it is shown that these constellations can achieve within a 0.12 dB SNR gap in the high SNR regime.

In contrast, our expansion coding approach achieves arbitrarily close to the capacity of the channel, such that it outperforms these previously proposed modulation techniques.

2.3.1 Problem Setup for AEN Channel Coding

More precisely, the model for additive exponential noise (AEN) channel is illustrated in Figure 2.4, where the channel noise Z_i in (2.1) are independently and identically distributed according to an exponential density with mean E_{Z}, i.e., omitting the index i, noise has the following density:

$$f_{\mathsf{Z}}(z) = \frac{1}{E_{\mathsf{Z}}} e^{-\frac{z}{E_{\mathsf{Z}}}} \cdot u(z), \tag{2.2}$$

where $u(z) = 1$ for $z \geq 0$ and $u(z) = 0$ otherwise. Moreover, channel input X_i in (2.1) is restricted to be non-negative and satisfies mean constraint

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathsf{X}_i] \leq E_{\mathsf{X}}. \tag{2.3}$$

The capacity of AEN channel is given by [34],

$$C_{\text{AEN}} = \log(1 + \text{SNR}), \tag{2.4}$$

where $\text{SNR} \triangleq E_{\mathsf{X}}/E_{\mathsf{Z}}$, and the capacity achieving input distribution is given by

$$f^*_{\mathsf{X}}(x) = \frac{E_{\mathsf{X}}}{(E_{\mathsf{X}} + E_{\mathsf{Z}})^2} e^{\frac{-x}{E_{\mathsf{X}}+E_{\mathsf{Z}}}} \cdot u(x) + \frac{E_{\mathsf{Z}}}{E_{\mathsf{X}} + E_{\mathsf{Z}}} \cdot \delta(x), \tag{2.5}$$

where $\delta(x) = 1$ if and only if $x = 0$. Here, the optimal input distribution is not exponentially distributed, but a mixture of an exponential distribution with a delta function. However, we observe that in the high SNR regime, the optimal distribution gets closer to an exponential distribution with mean E_{X}, since the weight of delta function approaches to 0 as SNR tends to infinity.

2.3.2 Binary Expansion of Exponential Distribution

The basis of the proposed coding scheme is the expansion of analog random variables to discrete ones, and the exponential distribution emerges as a first candidate due to its decomposition property. We show the following lemma, which allows us to have independent Bernoulli random variables in the binary expansion of an exponential random variable.

Lemma 2.2. *Let* B_l*'s be independent Bernoulli random variables with parameters given by* b_l*, i.e.,* $Pr\{\mathsf{B}_l = 1\} \triangleq b_l$*, and consider the random variable defined by*

$$\mathsf{B} \triangleq \sum_{l=-\infty}^{\infty} 2^l \mathsf{B}_l.$$

Then, the random variable B *is exponentially distributed with mean* λ^{-1}*, i.e., its pdf is given by*

$$f_{\mathsf{B}}(b) = \lambda e^{-\lambda b}, \quad b \geq 0,$$

if and only if the choice of b_l *is given by*

$$b_l = \frac{1}{1 + e^{\lambda 2^l}}.$$

Proof. See Appendix 2.A. □

This lemma reveals one can reconstruct exponential random variable from a set of independent Bernoulli random variables perfectly. Figure 2.5 illustrates that the distribution of recovered random variable from expanded levels (obtained from the statistics of 100000 independent samples) is a good approximation of original exponential distribution.

A set of typical numerical values of b_ls by fixing $\lambda = 1$ is shown in Figure 2.6. It is evident that b_l approaches 0 for the "higher" levels and approaches 0.5 for what we refer to as "lower" levels. Hence, the primary non-trivial levels within which coding is meaningful are the so-called "middle" ones, which provides the basis for truncating the number of levels to a finite value without a significant loss in performance.

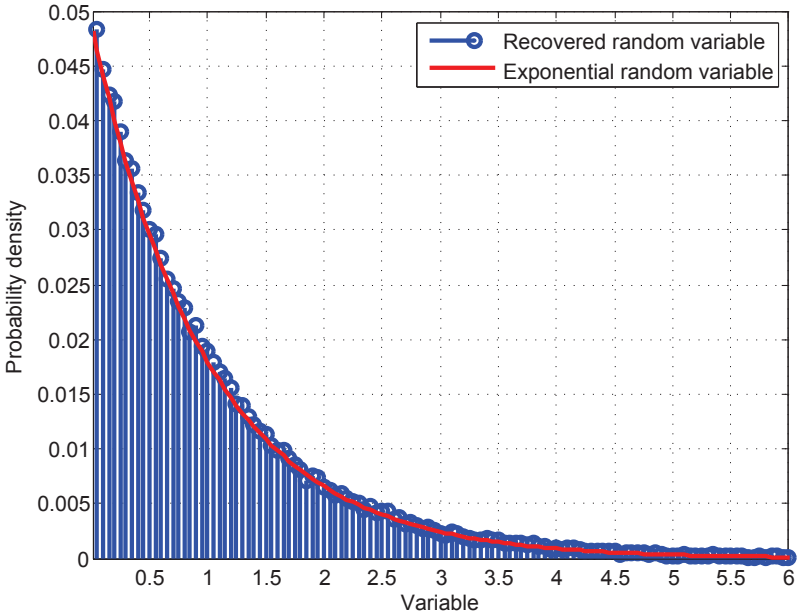

Figure 2.5: **Distribution of recovered random variable from expanded levels, comparing with original exponential distribution ($\lambda = 1$).** 100000 samples are generated from the expansion form of discrete random variables, where expansion levels are truncated from -10 to 10.

2.3.3 Expansion Coding for AEN Channel

We consider the binary expansion of the channel noise

$$\hat{\mathsf{Z}}_i \triangleq \sum_{l=-L_1}^{L_2} 2^l \mathsf{Z}_{i,l}, \tag{2.6}$$

where $\mathsf{Z}_{i,l}$ are i.i.d. Bernoulli random variables with parameters

$$q_l \triangleq \Pr\{\mathsf{Z}_l = 1\} = \frac{1}{1 + e^{2^l/E_{\mathsf{Z}}}}, \quad l = -L_1, \ldots, L_2. \tag{2.7}$$

By Lemma 2.2, $\hat{\mathsf{Z}}_i \xrightarrow{d} \mathsf{Z}_i$ as $L_1, L_2 \to \infty$. In this sense, we approximate the exponentially distributed noise perfectly by a set of discrete Bernoulli

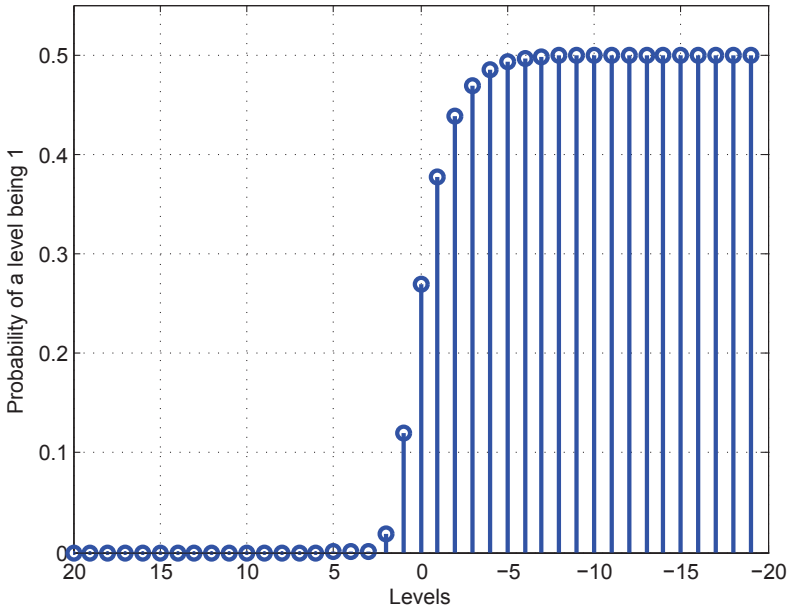

Figure 2.6: **Numerical results for a set of b_l with $\lambda = 1$.** X-axis is the level index for binary expansion (e.g. value -2 means the weight of corresponding level is 2^{-2}), and Y-axis shows the corresponding probability of taking value 1 at each level, i.e., b_l.

distributed noises. Similarly, we also expand channel input and output as in the following,

$$\hat{\mathsf{X}}_i \triangleq \sum_{l=-L_1}^{L_2} 2^l \mathsf{X}_{i,l}, \tag{2.8}$$

$$\hat{\mathsf{Y}}_i \triangleq \sum_{l=-L_1}^{L_2} 2^l \mathsf{Y}_{i,l}, \tag{2.9}$$

where $\mathsf{X}_{i,l}$ and $\mathsf{Y}_{i,l}$ are also Bernoulli random variables with parameters $\Pr\{\mathsf{X}_l = 1\} \triangleq p_l$ and $\Pr\{\mathsf{Y}_l = 1\} \triangleq r_l$ correspondingly. Here, the channel input is chosen as zero for levels $l \notin \{-L_1, \cdots, L_2\}$. Noting that the summation in the original channel is a sum over real numbers, we do not have a binary symme-

try channel (BSC) at each level (from X_ls to Y_ls). If we could replace the real sum by modulo-2 sum such that at each level l we have an independent coding problem, then any capacity achieving BSC code can be utilized over this channel. (Here, instead of directly using the capacity achieving input distribution of each level, we can use its combination with the method of Gallager [37] to achieve a rate corresponding to the one obtained by the mutual information $I(\mathsf{X}_l;\mathsf{Y}_l)$ evaluated with an input distribution Bernoulli with parameter p_l. This helps to approximate the optimal input distribution of the original channel.) However, due to the addition over real numbers, carries exist between neighboring levels, which further implies that levels are not independent (see Figure 2.7). Every level, except for the lowest one, is impacted by carry from lower levels. In order to alleviate this issue, two schemes are proposed in the following to ensure independent operation of the levels. In these models of coding over independent parallel channels, the total achievable rate is the summation of individual achievable rates over all levels.

2.3.3.1 Considering carries as noise

Denoting the carry seen at level l as $\mathsf{C}_{i,l}$, which is also a Bernoulli random variable with parameter $\Pr\{\mathsf{C}_{i,l} = 1\} \triangleq c_l$, the remaining channels can be represented with the following,

$$\mathsf{Y}_{i,l} = \mathsf{X}_{i,l} \oplus \tilde{\mathsf{Z}}_{i,l}, \quad i = 1, \cdots, N,$$

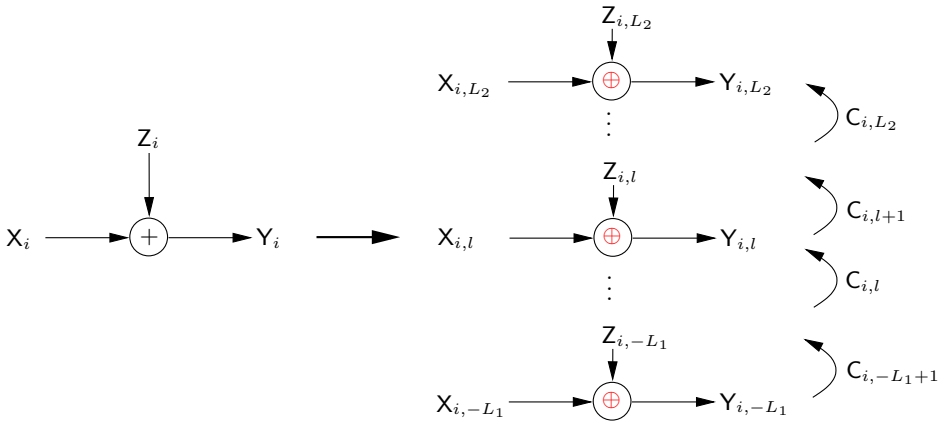

Figure 2.7: **Expansion coding for AEN channel.** Original AEN channel is expanded into a set of BSCs, where carries are considered between neighboring levels.

where the effective noise, $\tilde{\mathsf{Z}}_{i,l}$, is a Bernoulli random variable obtained by the convolution of the actual noise and the carry, i.e.

$$\tilde{q}_l \triangleq \Pr\{\tilde{\mathsf{Z}}_{i,l} = 1\} = q_l \otimes c_l,$$

where the convolution $\otimes$ is defined as $q_l \otimes c_l \triangleq q_l(1-c_l) + c_l(1-q_l)$, and the carry probability is given by the following recursion relationship:

1) For level $l = -L_1$,

$$c_{-L_1} = 0;$$

2) For level $l > -L_1$,

$$c_{l+1} = p_l q_l (1-c_l) + p_l(1-q_l)c_l + (1-p_l)q_l c_l + p_l q_l c_l.$$

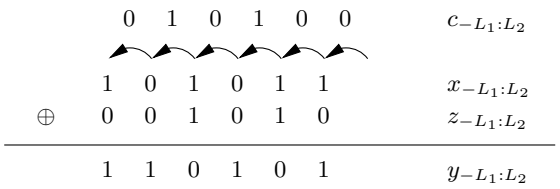

Figure 2.8: **Illustration of decoding carries in expansion coding scheme.** An example to show the mechanism of decoding carries is described, where the decoding process starts from the lowest level and initializes with 0.

Using capacity achieving codes for BSC, e.g., polar codes or spinal codes, combined with the Gallager's method, expansion coding achieves the following rate by considering carries as noise.

Theorem 2.3. *Expansion coding, by considering carries as noise, achieves the rate for AEN channel given by*

$$\hat{R}_1 = \sum_{l=-L_1}^{L_2} \hat{R}_{1,l} = \sum_{l=-L_1}^{L_2} \left[H(p_l \otimes \tilde{q}_l) - H(\tilde{q}_l) \right], \tag{2.10}$$

for any $L_1, L_2 > 0$*, where* $p_l \in [0, 0.5]$ *is chosen to satisfy constraint* (2.3)*, i.e.,*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\hat{\mathsf{X}}_i] = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=-L_1}^{L_2} 2^l \mathbb{E}[\mathsf{X}_{i,l}] = \sum_{l=-L_1}^{L_2} 2^l p_l \leq E_{\mathsf{X}}.$$

2.3.3.2 Decoding carries

In this scheme, let us consider decoding starting from the lowest level $l = -L_1$. The receiver will obtain the correct $\mathsf{X}_{i,-L_1}$ for $i = 1, \cdots, N$ by using powerful discrete coding at this level. As the receiver has the knowledge of $\mathsf{Y}_{i,-L_1}$, it is simple to determine the correct noise sequence $\mathsf{Z}_{i,-L_1}$ for $i =$

$1, \cdots, N$. With this knowledge, the receiver can directly obtain each $\mathsf{C}_{i,-L_1+1}$ for $i = 1, \cdots, N$, which is the carry from level $l = -L_1$ to level $l = -L_1 + 1$. This way (see a particular example in Figure 2.7), by iterating to higher levels, the receiver can recursively subtract off the impact of carry bits. Therefore, when there is no decoding error at each level, the effective channel that the receiver will see is given by

$$\mathsf{Y}_{i,l} = \mathsf{X}_{i,l} \oplus \mathsf{Z}_{i,l}, \quad i = 1, \cdots, N,$$

for $l = -L_1, \cdots, L_2$. We remark that with this decoding strategy, the effective channels will no longer be a set of independent parallel channels, as decoding in one level affects the channels at higher levels. However, if the utilized coding method is strong enough (e.g., if the error probability decays to 0 exponentially with N), then decoding error due to carry bits can be made insignificant by increasing N for a given number of levels. We state the rate resulting from this approach in a theorem.

Theorem 2.4. *Expansion coding, by decoding the carries, achieves the rate for AEN channel given by*

$$\hat{R}_2 = \sum_{l=-L_1}^{L_2} \hat{R}_{2,l} = \sum_{l=-L_1}^{L_2} \left[H(p_l \otimes q_l) - H(q_l) \right], \tag{2.11}$$

for any $L_1, L_2 > 0$, where $p_l \in [0, 0.5]$ is chosen to satisfy constraint (2.3)*, i.e.,*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\hat{\mathsf{X}}_i] = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=-L_1}^{L_2} 2^l \mathbb{E}[\mathsf{X}_{i,l}] = \sum_{l=-L_1}^{L_2} 2^l p_l \leq E_{\mathsf{X}}.$$

Compared to the previous case, the optimization problem is simpler here as the rate expression is simply the sum of the rates obtained from a set of parallel channels. Having these two theoretical achievable rates in hand, it remains to choose proper values for p_l. Note that the optimization problems given by Theorem 2.3 and 2.4 are not easy to solve in general. Here, instead of searching for the optimal solutions directly, we utilize the information from the optimal input distribution of the original channel. Recall that the distribution in (2.5) can be approximated by an exponential distribution with mean E_{X} at high SNR. Hence, one can simply choose p_l from the binary expansion of the exponential distribution with mean E_{X} as an achievable scheme, i.e.,

$$p_l \triangleq \Pr\{\mathsf{X}_l = 1\} = \frac{1}{1 + e^{2^l/E_{\mathsf{X}}}}, \quad l = -L_1, \ldots, L_2. \tag{2.12}$$

We now show that this proposed scheme achieves the capacity of AEN channel in the high SNR regime for a sufficiently high number of levels. For this purpose, we first characterize the asymptotic behavior of entropy at each level for q_l and $\tilde{q}_l$ correspondingly, where the later one is closely related to carries.

Lemma 2.5. *The entropy of noise seen at level l, $H(q_l)$, is bounded by*

$$H(q_l) < 3 \log e \cdot 2^{-l+\eta} \quad \textit{for } \; l > \eta, \tag{2.13}$$

$$H(q_l) > 1 - \log e \cdot 2^{l-\eta} \quad \textit{for } \; l \leq \eta, \tag{2.14}$$

where $\eta \triangleq \log E_{\mathsf{Z}}$.

Proof. See Appendix 2.B. □

Lemma 2.6. *The entropy of equivalent noise at level l, $H(\tilde{q}_l)$, is bounded by*

$$H(\tilde{q}_l) < 6 \log e \cdot (l - \eta) \cdot 2^{-l+\eta} \quad \textit{for } \; l > \eta, \tag{2.15}$$

$$H(\tilde{q}_l) > 1 - \log e \cdot 2^{l-\eta} \quad \textit{for } \; l \leq \eta, \tag{2.16}$$

where $\eta \triangleq \log E_{\mathsf{Z}}$.

Proof. See Appendix 2.C. □

The intuitions behind these lemmas are given by the example scenario in Figure 2.9, which says the tail's bounds of noises are both exponential. Now we are ready to obtain the main result for capacity gap of expansion coding scheme over AEN channel.

Theorem 2.7. *For any positive constant $\epsilon < 1$, if*

- $L_1 \geq -\log \epsilon - \log E_{\mathsf{Z}}$;
- $L_2 \geq -\log \epsilon + \log E_{\mathsf{X}}$;
- $SNR \geq 1/\epsilon$, *where* $SNR = E_{\mathsf{X}}/E_{\mathsf{Z}}$,

then, with the choice of p_l as (2.12),

1. by considering carries as noise, the achievable rate given by (2.10) *satisfies*

$$\hat{R}_1 \geq C_{AEN} - c,$$

where c is a constant not related to SNR or ϵ;

2. by decoding carries, the achievable rate given by (2.11) *satisfies*

$$\hat{R}_2 \geq C_{AEN} - 5 \log e \cdot \epsilon.$$

Proof. The proof of this theorem is based on the observation that the sequence of p_l is a left-shifted version of q_l at high SNR regime. As limited by power constraint, the number of levels shifted is at most $\log(1+\text{SNR})$, which further implies the rate we gain is roughly $\log(1+\text{SNR})$ as well, when carries are decoded. If considering carries as noise, then there is apparent gap between two version of noises, which leads to a constant gap for achievable rate. Figure 2.9 helps to illustrate key steps of the intuition, and a detailed proof with precise calculation is given in Appendix 2.D. □

By Lemma 2.2, $\hat{\mathsf{Z}} \xrightarrow{d} \mathsf{Z}$, and combined with the argument in Theorem 2.1, we have $\hat{R}_2 \to R$ as $L_1, L_2 \to \infty$. Hence, the coding scheme also works well for the original AEN channel. More precisely, expansion coding scheme achieves the capacity of AEN channel at high SNR region using moderate large number of expansion levels.

2.3.4 Numerical results

We calculate the rates obtained from the two schemes above ($\hat{R}_1$ as (2.10) and $\hat{R}_2$ as (2.11)) with input probability distribution given by (2.12).

Numerical results are given in Figure 2.10. It is evident from the figure (and also from the analysis given in Theorem 2.7) that the proposed technique

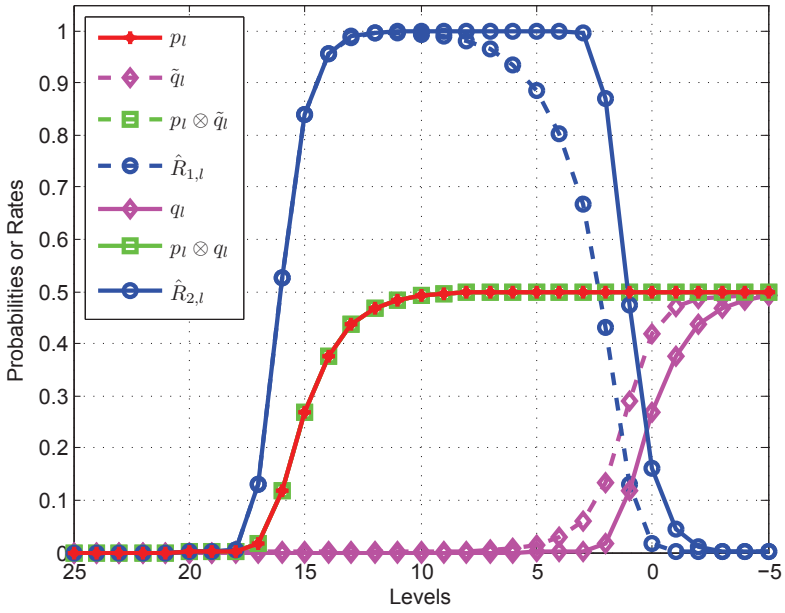

Figure 2.9: **Signal and noise probabilities, and rates per level.** p_l, q_l, $p_l \otimes q_l$, $\tilde{q}_l$, $p_l \otimes \tilde{q}_l$ and rates at each level are shown. In this example, $E_{\mathsf{X}} = 2^{15}$ and $E_{\mathsf{Z}} = 2^0$, which further implies p_l is a left-shifted version of q_l by 15 levels. The coding scheme with $L_1 = 5$ and $L_2 = 20$ covers the significant portion of the rate obtained by using all of the parallel channels.

of decoding carries, when implemented with sufficiently large number of levels, achieves channel capacity at high SNR regime.

Another point is that neither of the two schemes works well in low SNR regime, which mainly results from the fact that input approximation is only perfect for sufficiently high SNR. Nevertheless, the scheme (the rate obtained by decoding carries) performs close to optimal in the moderate SNR regime as well.

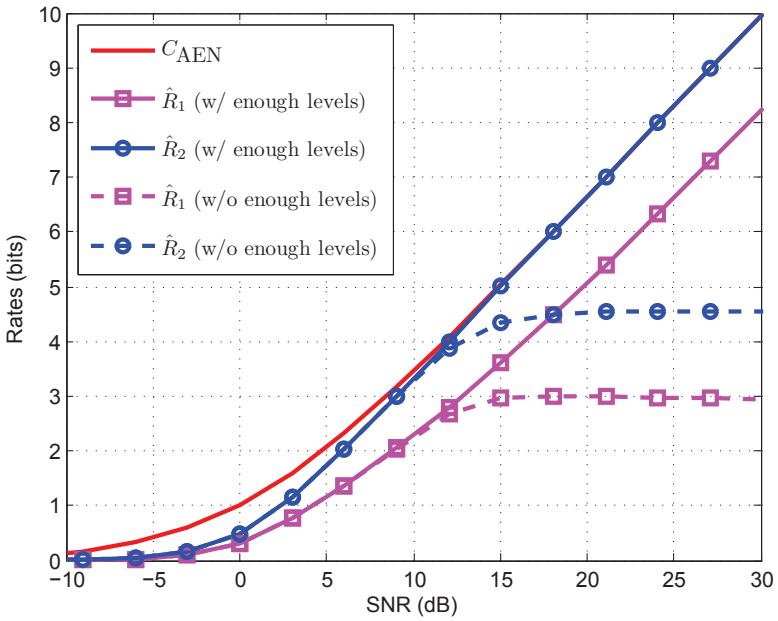

Figure 2.10: **Numerical results of achievable rates for AEN channels using expansion coding.** $\hat{R}_1$: The rate obtained by considering carries as noise. $\hat{R}_2$: The rate obtained by decoding carry at each level. Solid lines represent adopting enough number of levels as indicated in Theorem 2.7, while dashed lines represent only adopting constant number of levels (not scaling with SNR).

2.3.5 Generalization

Up to now, only binary expansion is considered. Generalization to q-ary expansion with $q \geq 2$ is discussed here. Note that this change does not impact the expansion coding framework, whereas the only difference lies that each level after expansion should be modeled as a q-ary discrete memoryless channel. For this, we need to characterize the q-ary expansion of exponential distribution. Mathematically, the parameters of expanded levels for an exponential random

variable B with parameter λ can be calculated as follow:

$$\begin{aligned} b_{l,s} &\triangleq \Pr\{\mathsf{B}_l = s\} \\ &= \sum_{k=0}^{\infty} \Pr\{q^l(qk+s) \le \mathsf{B} < q^l(qk+s+1)\} \\ &= \sum_{k=0}^{\infty} \left[e^{-\lambda q^l (qk+s)} - e^{-\lambda q^l (qk+s+1)} \right] \\ &= \frac{\left(1 - e^{-\lambda q^l}\right) e^{-\lambda q^l s}}{1 - e^{-\lambda q^{l+1}}}, \end{aligned}$$

where $l \in \{-L_1, \ldots, L_2\}$ and $s \in \{0, \ldots, q-1\}$.

Based on this result, consider channel input and noise expansions as

$$p_{l,s} \triangleq \Pr\{\mathsf{X}_l = s\} = \frac{\left(1 - e^{-q^l/E_{\mathsf{X}}}\right) e^{-q^l s/E_{\mathsf{X}}}}{1 - e^{-q^{l+1}/E_{\mathsf{X}}}},$$

and

$$q_{l,s} \triangleq \Pr\{\mathsf{Z}_l = s\} = \frac{\left(1 - e^{-q^l/E_{\mathsf{Z}}}\right) e^{-q^l s/E_{\mathsf{Z}}}}{1 - e^{-q^{l+1}/E_{\mathsf{Z}}}}.$$

Then, the achievable rate by decoding carries (note that in q-ary expansion case, carries are still Bernoulli distributed) can be expressed as

$$\hat{R}_2 = \sum_{l=-L_1}^{L_2} \left[H(p_{l,0:q-1} \otimes q_{l,0:q-1}) - H(q_{l,0:q-1}) \right], \tag{2.17}$$

where $p_{l,0:q-1}$ and $q_{l,0:q-1}$ denote the distribution of expanded random variables at level l for input and noise respectively; $\otimes$ represents for the vector convolution.

When implemented with enough number of levels in coding, the achievable rates given by (2.17) can still achieve the capacity of AEN channel. More

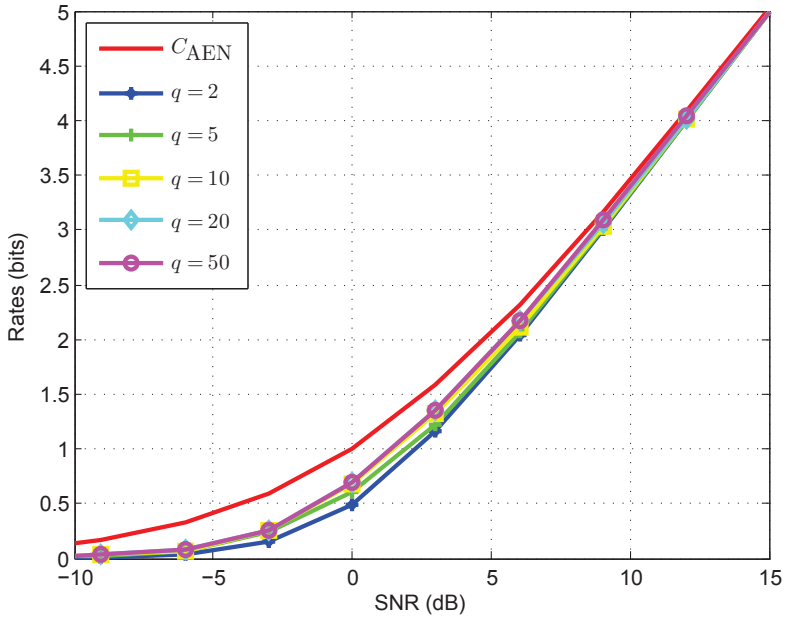

Figure 2.11: **Numerical results for q-ary expansion.** The achievable rates using q-ary expansion coding by decoding carries are illustrated in the figure.

precisely, from the numerical result shown in Figure 2.11, expansion coding with larger q can achieve higher rate (although this enhancement becomes insignificant when q is greater than 10).

2.4 Summary

Here, the method of expansion coding is proposed to construct good codes for analog channel coding. With a perfect or approximate decomposition of channel noises, we consider coding over independent parallel representations, which provides a foundation for reducing the original problems to a set of parallel simpler subproblems. In particular, via expansion channel coding, we

consider coding over q-ary channels for each expanded level. This approximation of the original channel together with capacity achieving codes for each level (to reliably transmit messages over each channel constructed) and Gallager's method (to achieve desired communication rates for each channel) allow for constructing near-capacity achieving codes for the original channel.

One significant benefit from expansion coding is coding complexity. As indicated in theoretical analysis (Theorem 2.7), approximately $-2\log\epsilon + \log\mathrm{SNR}$ number of levels are sufficient for the channel coding schemes. Thus, by choosing "good" low complexity optimal codes within each level (such as polar codes [14], [21]), the overall complexity of the coding scheme can be made small for the original continuous-valued channel coding problems (polynomial of block length N and $\log\mathrm{SNR}$).

Although the discussion in this section focuses on AEN channels, expansion coding scheme is a more general framework and its applications are not limited to such scenarios. Towards this end, any channel noise with decomposable distribution could fit into the range of expansion coding. Moreover, the idea of expansion could also be generalized to network information theory, where it can play a similar role like deterministic models [30]. However, the expanded channels are not totally deterministic in our case, but with different noise levels, which may help to construct a more precise models for network analyses.

2.A Proof of Lemma 2.2

The "if" part follows by extending the one given in [38], which considers the expansion of a truncated exponential random variable. We show the result by calculating the moment generating function of B. Using the assumption that $\{\mathsf{B}_l\}_{l\in\mathbb{Z}}$ are mutually independent, we have

$$M_{\mathsf{B}}(t) = \mathbb{E}\left[e^{t\mathsf{B}}\right] = \prod_{l=-\infty}^{\infty} \mathbb{E}\left[e^{t2^l\mathsf{B}_l}\right].$$

Noting that B_l are Bernoulli random variables, we have

$$\mathbb{E}\left[e^{t2^l\mathsf{B}_l}\right] = \frac{e^{t2^l}}{1+e^{\lambda 2^l}} + \left(1 - \frac{1}{1+e^{\lambda 2^l}}\right) = \frac{1+e^{(t-\lambda)2^l}}{1+e^{-\lambda 2^l}}.$$

Then, using the fact that for any constant $\alpha \in \mathbb{R}$,

$$\prod_{l=0}^{n}(1+e^{\alpha 2^l}) = \frac{1-e^{2^{n+1}\alpha}}{1-e^{\alpha}},$$

we can obtain the following for $t < \lambda$,

$$\prod_{l=0}^{\infty} \mathbb{E}\left[e^{t2^l\mathsf{B}_l}\right] = \lim_{n\to\infty} \prod_{l=0}^{n} \frac{1+e^{(t-\lambda)2^l}}{1+e^{-\lambda 2^l}} = \frac{1-e^{-\lambda}}{1-e^{t-\lambda}}. \tag{2.18}$$

Similarly, for the negative part, we have

$$\prod_{l=-n}^{-1}(1+e^{\alpha 2^l}) = \frac{1-e^{\alpha}}{1-e^{\alpha 2^{-n}}},$$

which further implies that

$$\prod_{l=-\infty}^{-1} \mathbb{E}\left[e^{t2^l\mathsf{B}_l}\right] = \lim_{n\to\infty} \frac{1-e^{t-\lambda}}{1-e^{(t-\lambda)2^{-n}}} \frac{1-e^{-\lambda 2^{-n}}}{1-e^{-\lambda}} = \frac{\lambda(1-e^{t-\lambda})}{(\lambda-t)(1-e^{-\lambda})}. \tag{2.19}$$

Thus, finally for any $t < \lambda$, combining equations (2.18) and (2.19), we get

$$M_{\mathsf{B}}(t) = \frac{\lambda}{\lambda - t}.$$

The observation that this is the moment generation function for an exponentially distributed random variable with parameter λ concludes the proof.

The independence relationships between levels in "only if" part can be simply verified using memoryless property of exponential distribution. Here we just need to show the parameter for Bernoulli random variable at each level. Observe that for any $l \in \mathbb{Z}$,

$$\begin{aligned}\Pr\{\mathsf{B}_l = 1\} &= \Pr\{\mathsf{B} \in \cup_{k\in\mathbb{N}}[2^l(2k-1), 2^l(2k))\} \\ &= \sum_{k\in\mathbb{N}} \Pr\{2^l(2k-1) \leq \mathsf{B} < 2^l(2k)\}. \end{aligned} \tag{2.20}$$

Using cdf of exponential distribution, we obtain

$$\Pr\{2^l(2k-1) \leq \mathsf{B} < 2^l(2k)\} = e^{-\lambda 2^l(2k-1)} - e^{-\lambda 2^l(2k)} = e^{-\lambda 2^l(2k)}\left(e^{\lambda 2^l} - 1\right).$$

Putting this back to (2.20) we have

$$\Pr\{\mathsf{B}_l = 1\} = \sum_{k=1}^{\infty} e^{-\lambda 2^l(2k)}\left(e^{\lambda 2^l} - 1\right) = \frac{1}{e^{\lambda 2^l} + 1}.$$

2.B Proof of Lemma 2.5

From (2.7), and $\eta \triangleq \log E_{\mathsf{Z}}$, we have

$$q_l = \frac{1}{1 + e^{2^l/E_{\mathsf{Z}}}} = \frac{1}{1 + e^{2^{l-\eta}}}.$$

By definition of entropy, we obtain

$$\begin{aligned} H(q_l) &= -q_l \log q_l - (1-q_l)\log(1-q_l) \\ &= -\frac{1}{1+e^{2^{l-\eta}}} \log \frac{1}{1+e^{2^{l-\eta}}} - \frac{e^{2^{l-\eta}}}{1+e^{2^{l-\eta}}} \log \frac{e^{2^{l-\eta}}}{1+e^{2^{l-\eta}}}. \end{aligned}$$

When $l \leq \eta$, we obtain a lower bound as

$$\begin{aligned} H(q_l) &= \frac{1}{1+e^{2^{l-\eta}}} \log \left(1+e^{2^{l-\eta}}\right) + \frac{e^{2^{l-\eta}}}{1+e^{2^{l-\eta}}} \log \left(\frac{1+e^{2^{l-\eta}}}{e^{2^{l-\eta}}}\right) \\ &= \log \left(1+e^{2^{l-\eta}}\right) - \frac{e^{2^{l-\eta}}}{1+e^{2^{l-\eta}}} \log e \cdot 2^{l-\eta} \\ &\overset{(a)}{>} \log(1+1) - \log e \cdot 2^{l-\eta} \\ &= 1 - \log e \cdot 2^{l-\eta}, \end{aligned}$$

where (a) is due to $e^{2^{l-\eta}} > 1$ and $-e^{2^{l-\eta}}/(1+e^{2^{l-\eta}}) > -1$.

On the other hand, when $l > \eta$, we have

$$\begin{aligned} H(q_l) &= \frac{1}{1+e^{2^{l-\eta}}} \log \left(1+e^{2^{l-\eta}}\right) + \frac{e^{2^{l-\eta}}}{1+e^{2^{l-\eta}}} \log \left(\frac{1+e^{2^{l-\eta}}}{e^{2^{l-\eta}}}\right) \\ &\overset{(b)}{<} \frac{1}{1+e^{2^{l-\eta}}} \log \left(2e^{2^{l-\eta}}\right) + \log \left(1+e^{-2^{l-\eta}}\right) \\ &\overset{(c)}{<} \frac{1+2^{l-\eta} \cdot \log e}{1+e^{2^{l-\eta}}} + e^{-2^{l-\eta}} \cdot \log e \\ &\overset{(d)}{<} \frac{1+2^{l-\eta} \cdot \log e}{1+1+2^{l-\eta}+2^{2(l-\eta)}/2} + \frac{\log e}{1+2^{l-\eta}} \\ &\overset{(e)}{<} 2^{\eta-l+1} \cdot \log e + 2^{\eta-l} \cdot \log e \\ &= 3 \log e \cdot 2^{\eta-l}, \end{aligned}$$

where

(b) is from $1 < e^{2^{l-\eta}}$ and $e^{2^{l-\eta}}/(1+e^{2^{l-\eta}}) < 1$;

(c) is from $\log(1+\alpha) < \log e \cdot \alpha$ for any $0 < \alpha < 1$;

(d) is from $e^{\alpha} > 1 + \alpha + \alpha^2/2 > 1 + \alpha$ for any $\alpha > 0$;

(d) is from $1 + 2^{l-\eta} \cdot \log e < (2 + 2^{l-\eta} + 2^{2(l-\eta)}/2)(2^{\eta-l+1} \cdot \log e)$ and $1 < (1 + 2^{l-\eta})2^{\eta-l}$ for any l and η.

2.C Proof of Lemma 2.6

By definition, $\tilde{q}_l = c_l \otimes q_l$, so its behavior is closely related to carries. Note that for any l, we have

$$\tilde{q}_l = c_l(1-q_l) + q_l(1-c_l) = q_l + c_l(1-2q_l) \geq q_l,$$

where the last inequality holds due to $q_l < 1/2$ and $c_l \geq 0$. Then, for $l \leq \eta$, we have

$$H(\tilde{q}_l) \geq H(q_l) > 1 - \log e \cdot 2^{l-\eta},$$

where the first inequality holds due to monotonicity of entropy, and the last inequality is due to (2.14) in Lemma 2.5. For the $l > \eta$ part, we need to characterize carries first. We have the following assertion:

$$c_l < 2^{\eta-l+1} - \frac{2}{1+e^{2^{l-\eta}}}, \quad \text{for } \ l > \eta, \tag{2.21}$$

and the proof is based on the following induction analysis. For $l = \eta + 1$, this is simply true, because $c_l < 1/2$ for any l. Assume (2.21) is true for level $l > \eta$,

then at the $l+1$ level, we have

$$\begin{aligned} c_{l+1} &= p_l q_l(1-c_l) + p_l(1-q_l)c_l + (1-p_l)q_l c_l + p_l q_l c_l \\ &= p_l(c_l + q_l - 2q_l c_l) + q_l c_l \\ &\overset{(a)}{<} \frac{1}{2}(c_l + q_l - 2q_l c_l) + q_l c_l \\ &= \frac{1}{2}(c_l + q_l) \\ &\overset{(b)}{<} \frac{1}{2}\left(2^{\eta-l+1} - \frac{2}{1+e^{2^{l-\eta}}} + \frac{1}{1+e^{2^{l-\eta}}}\right) \\ &\overset{(c)}{<} 2^{-(l-\eta+1)+1} - \frac{2}{1+e^{2^{l-\eta+1}}}, \end{aligned}$$

where

(a) is due to $p_l < 1/2$ and $c_l + q_l - 2q_l c_l = c_l(1-2q_l) + q_l > 0$;

(b) is due to the assumption (2.21) for level l;

(c) is due to the fact that $1/[2(1+e^{2^{l-\eta}})] > 2/(1+e^{2^{l-\eta+1}})$ holds for any $l > \eta$. To this end, the assertion also holds for level $l+1$, and this completes the proof of (2.21).

Using (2.21), we obtain that for any $l > \eta$

$$\begin{aligned} \tilde{q}_l &= q_l + c_l(1-2q_l) \\ &< \frac{1}{1+e^{2^{l-\eta}}} + \left(2^{\eta-l+1} - \frac{2}{1+e^{2^{l-\eta}}}\right)\left(1 - \frac{2}{1+e^{2^{l-\eta}}}\right) \\ &= 2^{\eta-l+1} - \frac{1+2^{\eta-l+2}}{1+e^{2^{l-\eta}}} + \frac{4}{(1+e^{2^{l-\eta}})^2} \\ &< 2^{\eta-l+1}, \end{aligned} \tag{2.22}$$

where the last inequality holds due to $(1+2^{\eta-l+2})(1+e^{2^{l-\eta}}) > 4$ for any $l > \eta$.

Finally, we obtain

$$\begin{aligned} H(\tilde{q}_l) &\overset{(d)}{<} H(2^{\eta-l+1}) \\ &= -2^{\eta-l+1}\log(2^{\eta-l+1}) - (1-2^{\eta-l+1})\log(1-2^{\eta-l+1}) \\ &\overset{(e)}{<} (l-\eta-1)\cdot 2^{\eta-l+1} + (1-2^{\eta-l+1})\cdot 2\log e\cdot 2^{\eta-l+1} \\ &\overset{(f)}{<} (l-\eta)\cdot 2^{\eta-l+1} + (l-\eta)\cdot 2\log e\cdot 2^{\eta-l+1} \\ &< 3\log e\cdot(l-\eta)\cdot 2^{\eta-l+1} \\ &= 6\log e\cdot(l-\eta)\cdot 2^{\eta-l}, \end{aligned}$$

where

(d) is from (2.22) and the monotonicity of entropy;

(e) is from $-\log(l-\eta-\alpha) < 2\log e\cdot\alpha$ for any $\alpha \le 1/2$;

(f) is from $1-2^{\eta-l+1} < l-\eta$ for any $l > \eta$.

From the proof, the information we used for p_l is that $p_l < 1/2$, so this bound is irrelative to SNR, i.e. this upper bound is a uniform one for any SNR.

2.D Proof of Theorem 2.7

We first prove that $\hat{R}_2$ achieves capacity. Denote $\xi = \log E_{\mathsf{X}}$ and $\eta = \log E_{\mathsf{Z}}$. Then, we have an important observation that

$$p_l = \frac{1}{1+e^{2^l/2^\xi}} = q_{l+\eta-\xi}, \tag{2.23}$$

which means channel input is a shifted version of noise with respect to expansion levels (see Figure 2.9 for intuition). Based on this, we have

$$\begin{aligned}
\hat{R}_2 &= \sum_{l=-L_1}^{L_2} [H(p_l \otimes q_l) - H(q_l)] \\
&\overset{(a)}{\geq} \sum_{l=-L_1}^{L_2} [H(p_l) - H(q_l)] \\
&\overset{(b)}{=} \sum_{l=-L_1}^{L_2} [H(q_{l+\eta-\xi}) - H(q_l)] \\
&= \sum_{l=-L_1+\eta-\xi}^{L_2+\eta-\xi} H(q_l) - \sum_{l=-L_1}^{L_2} H(q_l) \\
&\overset{(c)}{=} \sum_{l=-L_1+\eta-\xi}^{-L_1-1} H(q_l) - \sum_{l=L_2+\eta-\xi+1}^{L_2} H(q_l) \\
&\geq \sum_{l=-L_1+\eta-\xi}^{-L_1-1} \left[1 - \log e \cdot 2^{l-\eta}\right] - \sum_{l=L_2+\eta-\xi+1}^{L_2} 3 \log e \cdot 2^{\eta-l} \\
&\overset{(d)}{\geq} (\xi - \eta) - \log e \cdot 2^{-L_1-\eta} - 3 \log e \cdot 2^{-L_2+\xi} \\
&\overset{(e)}{\geq} \log\left(\frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}}\right) - \log e \cdot \epsilon - 3 \log e \cdot \epsilon \\
&\overset{(f)}{\geq} \log\left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}}\right) - \log e \cdot \frac{E_{\mathsf{Z}}}{E_{\mathsf{X}}} - 4 \log e \cdot \epsilon \\
&\overset{(g)}{\geq} \log\left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}}\right) - 5 \log e \cdot \epsilon, \qquad (2.24)
\end{aligned}$$

where

(a) is due to $p_l \otimes q_l = p_l(1-q) + (1-p_l)q_l \geq p_l$, and monotonicity of entropy;

(b) follows from (2.23);

(c) follows from (2.13) and (2.14) in Lemma 2.5;

(d) holds as

$$\sum_{l=-L_1+\eta-\xi}^{-L_1-1} 2^{l-\eta} \leq \sum_{l=-\infty}^{-L_1-1} 2^{l-\eta} = 2^{-L_1-\eta},$$

and

$$\sum_{l=L_2}^{L_2+\eta-\xi+1} 2^{\eta-l} \leq \sum_{l=\infty}^{L_2+\eta-\xi+1} 2^{\eta-l} = 2^{-L_2+\xi};$$

(e) is due to the assumptions that $L_1 \geq -\log\epsilon - \eta$, and $L_2 \geq -\log\epsilon + \xi$;

(f) is due to the fact that

$$\log\left(1+\frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}}\right) - \log\left(\frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}}\right) = \log\left(1+\frac{E_{\mathsf{Z}}}{E_{\mathsf{X}}}\right) \leq \log e \cdot \frac{E_{\mathsf{Z}}}{E_{\mathsf{X}}},$$

as $\log(1+\alpha) \leq \log e \cdot \alpha$ for any $\alpha \geq 0$;

(g) is due to the assumption that $\mathrm{SNR} = E_{\mathsf{X}}/E_{\mathsf{Z}} \geq 1/\epsilon$.

Next, we show the result for $\hat{R}_1$. Observe that

$$\begin{aligned}
\hat{R}_1 &= \sum_{l=-L_1}^{L_2} [H(p_l \otimes \tilde{q}_l) - H(\tilde{q}_l)] \\
&\overset{(h)}{\geq} \sum_{l=-L_1}^{L_2} [H(p_l \otimes q_l) - H(\tilde{q}_l)] \\
&= \sum_{l=-L_1}^{L_2} [H(p_l \otimes q_l) - H(q_l)] + \sum_{l=-L_1}^{L_2} [H(q_l) - H(\tilde{q}_l)] \\
&= \hat{R}_2 - \sum_{l=-L_1}^{L_2} [H(\tilde{q}_l) - H(q_l)] \\
&\overset{(i)}{\geq} \hat{R}_2 - \sum_{-L_1}^{\eta} \left[1 - \left(1 - \log e \cdot 2^{l-\eta}\right)\right] - \sum_{\eta+1}^{L_2} \left[6\log e \cdot (l-\eta)2^{-l+\eta} - 0\right]
\end{aligned}$$

$$
\begin{aligned}
&= \hat{R}_2 - \sum_{-L_1}^{\eta} \log e \cdot 2^{l-\eta} - \sum_{\eta+1}^{L_2} 6 \log e \cdot (l-\eta) 2^{-l+\eta} \\
&\overset{(j)}{\geq} \hat{R}_2 - 2 \log e - 12 \log e \\
&\overset{(k)}{\geq} \log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}}} \right) - 5 \log e \cdot \epsilon - 14 \log e,
\end{aligned}
$$

where

(h) is due to $\tilde{q}_l \geq q_l$, which further implies $p_l \otimes \tilde{q}_l \geq p_l \otimes q_l$;

(i) follows from (2.14) and (2.15), together with the fact that $H(\tilde{q}_l) \leq 1$ and $H(q_l) \geq 0$ for any l;

(j) follows from the observations that

$$
\sum_{-L_1}^{\eta} \log e \cdot 2^{l-\eta} \leq \log e \cdot 2^{-\eta} \cdot \sum_{-\infty}^{\eta} 2^l = 2 \log e,
$$

and

$$
\sum_{\eta+1}^{L_2} 6 \log e \cdot (l-\eta) \cdot 2^{-l+\eta} \leq 6 \log e \cdot \sum_{\eta+1}^{\infty} (l-\eta) \cdot 2^{-l+\eta} = 12 \log e;
$$

(k) is due to (2.24).

Thus, choosing $c = 19 \log e$ completes the proof. Note that, in the course of providing these upper bounds, the actual gap might be enlarged. A precise value of the gap is much smaller, i.e., as shown in Figure 2.10, numerical result for the capacity gap is around 1.72 bits.

Chapter 3

Expansion Coding for Data Compression

3.1 Problem Background and Related Work

Another well-studied (and practically valuable) research direction in information theory is the problem of compression of continuous-valued sources. Given the increased importance of voice, video and other multimedia, all of which are typically "analog" in nature, the value associated with low-complexity algorithms to compress continuous-valued data is likely to remain significant in the years to come.

Although both practical coding schemes as well as theoretical analysis are very heavily studied, a very limited literature exists that connects the theory with low-complexity codes. The most relevant literature in this context is on lattice compression and its low-density constructions [39]. Yet, this literature is also limited in scope and application.

In the domains of image compression and speech coding, Laplacian and exponential distributions are widely adopted as natural models of correlation between pixels and amplitude of voice [37]. Exponential distribution is also fundamental in characterizing continuous-time Markov processes [34]. Although the rate distortion functions for both have been known for decades,

there is still a gap between theory and existing low-complexity coding schemes. Some schemes have been proposed, primarily for the medium to high distortion regime, such as the classical scalar and vector quantization schemes [40], and Markov chain Monte Carlo (MCMC) based approach in [41]. However, the understanding of low-complexity coding schemes, especially for the low-distortion regime, remains limited. To this end, our expansion source coding scheme aims to approach the rate distortion limit with practical encoding and decoding complexity. By expanding the sources into independent levels, and using the decomposition property of exponential distribution, the problem has been remarkably reduced to a set of simpler subproblems, compression for discrete sources.

3.2 Expansion Source Coding: Theoretical Framework

Expansion source coding is a scheme of reducing the problem of compressing analog sources to compressing a set of discrete sources. In particular, consider an i.i.d. source $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_N$. A $(2^{NR}, N)$-rate distortion code consists of an encoding function $\varphi : \mathbb{R}^N \to \mathcal{M}$, where $\mathcal{M} \triangleq \{1, \ldots, 2^{NR}\}$, and a decoding function $\varsigma : \mathcal{M} \to \mathbb{R}^N$, which codes $\mathsf{X}_{1:N}$ to an estimate $\tilde{\mathsf{X}}_{1:N}$. Then, a rate and distortion pair (R, D) is said to be achievable if there exists a sequence of $(2^{NR}, N)$-rate distortion codes with $\lim_{N\to\infty} \mathbb{E}[d(\mathsf{X}_{1:N}, \tilde{\mathsf{X}}_{1:N})] \leq D$ for a given distortion measure of interest $d(\cdot, \cdot)$. The rate distortion function $R(D)$

is the infimum of such rates, and by Shannon's theorem [17], we have:

$$R(D) = \min_{f_{\tilde{\mathsf{X}}|\mathsf{X}}(\tilde{x}|x):\mathbb{E}[d(\mathsf{X}_{1:N},\tilde{\mathsf{X}}_{1:N})]\leq D} \frac{1}{N} I(\mathsf{X}_{1:N};\tilde{\mathsf{X}}_{1:N}),$$

where the optimal conditional distribution is given by $f^*_{\tilde{\mathsf{X}}|\mathsf{X}}(\tilde{x}|x)$.

The expansion source coding scheme proposed here is based on the observation that by expanding the original analog source into a set of independent discrete random variables, proper source coding schemes could be adopted for every expanded level. If this approximation in expansion is close enough, then the overall distortion obtained from expansion coding scheme is also close to the original distortion. More formally, consider the original analog source X and its approximation $\hat{\mathsf{X}}$ given by (omitting index i)

$$\hat{\mathsf{X}} \triangleq \mathsf{X}^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \mathsf{X}_l, \tag{3.1}$$

where X^{sign} represents the sign of $\hat{\mathsf{X}}$, and takes values from $\{-,+\}$; X_l is the expanded Bernoulli random variable at level l. Similarly, if we expand the estimate by

$$\hat{\tilde{\mathsf{X}}} \triangleq \tilde{\mathsf{X}}^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \tilde{\mathsf{X}}_l, \tag{3.2}$$

where $\tilde{\mathsf{X}}^{\text{sign}}$ represents the sign of $\hat{\tilde{\mathsf{X}}}$, random variable taking values from $\{-,+\}$, and $\tilde{\mathsf{X}}_l$ is independent Bernoulli random variable at level l after expansion.

Here, we reduce the original problem to a set of source coding subproblems over levels $-L_1$ to L_2. Similar to the channel coding case analyzed

in Chapter 2, if $\hat{\mathsf{X}} \xrightarrow{d.} \mathsf{X}$, and $\hat{\tilde{\mathsf{X}}} \xrightarrow{d.} \tilde{\mathsf{X}}^*$, as $L_1,\ L_2 \to \infty$, then the achieved rate distortion pair approximates the original one. Note that, in general, the decomposition may not be sufficiently close for most of the sources, and the distribution for the estimate may not be sufficiently approximated, which both add more distortion and result in a gap from the theoretical limit.

3.3 Expansion Source Coding: Exponential Source

In this section, a particular lossy compression example is introduced to illustrate the effectiveness of expansion source coding.

3.3.1 Problem Setup for Exponential Source Coding

Consider an i.i.d. exponential source sequence $\mathsf{X}_1, \ldots, \mathsf{X}_N$, i.e., omitting index i, each variable has a pdf given by

$$f_{\mathsf{X}}(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

where λ^{-1} is the mean of X. Distortion measure of concern is the "one-sided error distortion", i.e.

$$d(x_{1:N}, \tilde{x}_{1:N}) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{x}_i), & \text{if } x_{1:N} \succcurlyeq \tilde{x}_{1:N}, \\ \infty, & \text{otherwise,} \end{cases}$$

where $\succcurlyeq$ means $x_i \geq \tilde{x}_i$ for every $i \in \{1, \ldots, N\}$. This setup is equivalent to the one in [34], where another distortion measure is considered.

Lemma 3.1 ([34]). *The rate distortion function for an exponential source with*

the one-sided error distortion is given by

$$R(D) = \begin{cases} -\log(\lambda D), & 0 \leq D \leq \frac{1}{\lambda}, \\ 0, & D > \frac{1}{\lambda}. \end{cases} \tag{3.3}$$

Moreover, the optimal conditional distribution to achieve the limit is given by

$$f^*_{\mathsf{X}|\tilde{\mathsf{X}}}(x|\tilde{x}) = \frac{1}{D} e^{-(x-\tilde{x})/D}, \quad x \geq \tilde{x} \geq 0. \tag{3.4}$$

Proof. Proof is given in [34], and it is based on the observation that among the ensemble of all probability density functions with positive support set and mean constraint, exponential distribution maximizes the differential entropy. By designing a test channel from $\tilde{\mathsf{X}}$ to X, with additive noise distributed as exponential with parameter $1/D$, both the infimum mutual information and optimal conditional distribution can be characterized. Details can be found in Appendix 3.A. □

3.3.2 Expansion Coding for Exponential Source

Using Lemma 2.2, we reconstruct the exponential distribution by a set of discrete Bernoulli random variables. In particular, the expansion of exponential source over levels ranging from $-L_1$ to L_2 can be expressed as

$$\hat{\mathsf{X}}_i = \sum_{l=-L_1}^{L_2} 2^l \mathsf{X}_{i,l}, \quad i = 1, 2, \ldots, N,$$

where $\mathsf{X}_{i,l}$ are Bernoulli random variables with parameter

$$p_l \triangleq \Pr\{\mathsf{X}_{i,l} = 1\} = \frac{1}{1 + e^{\lambda 2^l}}. \tag{3.5}$$

This expansion perfectly approximates exponential source by letting $L_1, L_2 \to \infty$. Consider a similar expansion of the source estimate, i.e.

$$\hat{\tilde{\mathsf{X}}}_i = \sum_{l=-L_1}^{L_2} 2^l \tilde{\mathsf{X}}_{i,l}, \quad i = 1, 2, \ldots, N,$$

where $\tilde{\mathsf{X}}_{i,l}$ is the resulting Bernoulli random variable with parameter $\tilde{p}_l \triangleq \Pr\{\tilde{\mathsf{X}}_{i,l} = 1\}$. Utilizing the expansion method, the original problem of coding for a continuous source can be translated to a problem of coding for a set of independent binary sources. This transformation, although seemingly obvious, is valuable as one can utilize powerful coding schemes over discrete sources to achieve rate distortion limits with low complexity. In particular, we design two schemes for the binary source coding problem at each level.

3.3.2.1 Coding with one-sided distortion

In order to guarantee $x \geq \tilde{x}$, we formulate each level as a binary source coding problem under the binary one-sided distortion constraint: $d_{\mathrm{O}}(x_l, \tilde{x}_l) = \mathbf{1}_{\{x_l > \tilde{x}_l\}}$. Denoting the distortion at level l as d_l, an asymmetric test channel (Z-channel) from $\tilde{\mathsf{X}}_l$ to X_l can be constructed, where

$$\begin{aligned} \Pr\{\mathsf{X}_l = 1 | \tilde{\mathsf{X}}_l = 0\} &= \frac{d_l}{1 - p_l + d_l}, \\ \Pr\{\mathsf{X}_l = 0 | \tilde{\mathsf{X}}_l = 1\} &= 0. \end{aligned}$$

Based on this, we have $p_l - \tilde{p}_l = d_l$, and the achievable rate at a single level l is given by

$$R_{\mathrm{Z},l} = H(p_l) - (1 - p_l + d_l) H\left(\frac{d_l}{1 - p_l + d_l}\right). \tag{3.6}$$

Due to the decomposability property as stated previously, the coding scheme provided is over a set of parallel discrete levels indexed by $l = -L_1, \ldots, L_2$. Thus, by adopting rate distortion limit achieving codes over each level, expansion coding scheme readily achieves the following result:

Theorem 3.2. *For an exponential source, expansion coding achieves the rate distortion pair given by*

$$R_1 = \sum_{l=-L_1}^{L_2} R_{Z,l}, \tag{3.7}$$

$$D_1 = \sum_{l=-L_1}^{L_2} 2^l d_l + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1}, \tag{3.8}$$

for any $L_1, L_2 > 0$*, and* $d_l \in [0, 0.5]$ *for* $l \in \{-L_1, \cdots, L_2\}$*, where* p_l *is given by* (3.5)*.*

Proof. See Appendix 3.B. Note that, the last two terms in (3.8) are resulting from the truncation and vanish in the limit of large number of levels. In later parts of this section, we characterize the number of levels required in order to bound the resulting distortion within a constant gap. □

3.3.2.2 Successive encoding and decoding

Note that it is not necessary to make sure $x_l \geq \tilde{x}_l$ for every l to guarantee $x \geq \tilde{x}$. To this end, we introduce successive coding scheme, where encoding and decoding start from the highest level L_2 to the lowest. At a certain level, if all higher levels are encoded as equal to the source, then we must model this level as binary source coding with the one-sided distortion. Otherwise,

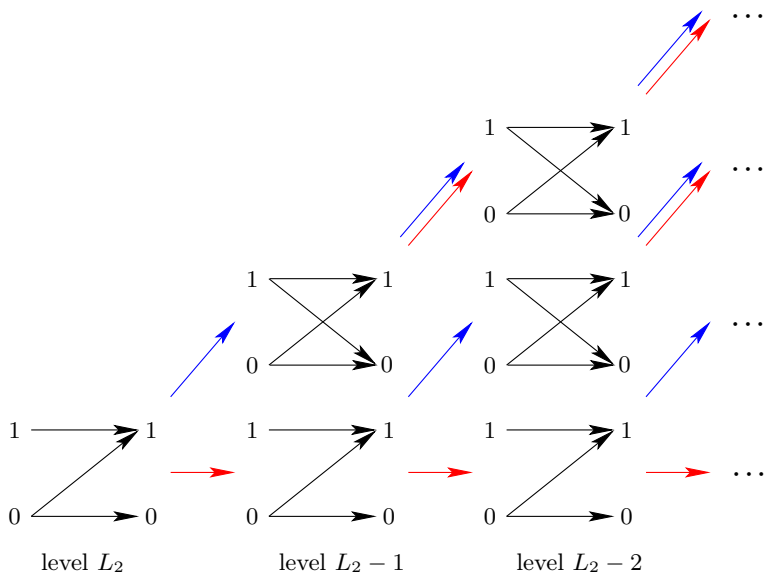

Figure 3.1: **Illustration of successive encoding and decoding.** Encoding and decoding start from the highest level. A lower level is modeled as one-side distortion (test channel is Z-channel) if and only if estimates in all higher levels are decoded as equal to the source. In this illustration, red arrows represent for decoded as equal, while blue ones represent for decoded as unequal.

we formulate this level as binary source coding with the symmetric distortion (see Figure 3.1 for an illustration of this successive coding scheme). In particular for the later case, the distortion of concern is Hamming distortion, i.e. $d_{\mathrm{H}}(x_l, \tilde{x}_l) = \mathbf{1}_{\{x_l \neq \tilde{x}_l\}}$. Denoting the equivalent distortion at level l as d_l, i.e. $\mathbb{E}[\mathsf{X}_l - \tilde{X}_l] = d_l$, then the symmetric test channel from $\hat{\mathsf{X}}_l$ to X_l is modeled as

$$\Pr\{\mathsf{X}_l = 1 | \hat{\mathsf{X}}_l = 0\} = \Pr\{\mathsf{X}_l = 0 | \tilde{\mathsf{X}}_l = 1\} = \frac{d_l}{1 - 2p_l + 2d_l}.$$

Hence, the achievable rate at level l is given by

$$R_{\mathrm{X},l} = H(p_l) - H\left(\frac{d_l}{1 - 2p_l + 2d_l}\right). \tag{3.9}$$

Based on these, we have the following achievable result:

Theorem 3.3. *For an exponential source, applying successive coding, expansion coding achieves the rate distortion pairs*

$$R_2 = \sum_{l=-L_1}^{L_2} \left[\rho_l R_{Z,l} + (1-\rho_l) R_{X,l}\right], \tag{3.10}$$

$$D_2 = \sum_{l=-L_1}^{L_2} 2^l d_l + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1}, \tag{3.11}$$

for any $L_1, L_2 > 0$, *and* $d_l \in [0, 0.5]$ *for* $l \in \{-L_1, \cdots, L_2\}$. *Here,* p_l *is given by* (3.5), *and the values of* ρ_l *are determined by:*

1) For $l = L_2$,

$$\rho_{L_2} = 1;$$

2) For $l < L_2$,

$$\rho_l = \prod_{k=l+1}^{L_2} (1 - d_k).$$

Proof. See Appendix 3.C. □

In this sense, the achievable pairs in both theorems are given by optimization problems over a set of parameters $\{d_{-L_1}, \ldots, d_{L_2}\}$. However, the problems are not convex, so an effective theoretical analysis may not be performed here for the optimal solution. But, by a heuristic choice of d_l, we can still get a good performance. Inspired by the fact that the optimal scheme

models noise as exponential with parameter $1/D$ in the test channel, we design d_l as the expansion parameter from this distribution, i.e.

$$d_l = \frac{1}{1 + e^{2^l/D}}. \tag{3.12}$$

We note that higher levels get higher priority and lower distortion with this choice, which is consistent with the intuition. This choice of d_l may not guarantee any optimality, although simulation results imply that it can be a local optimum. In the following, we show that the proposed expansion coding scheme achieves within a constant gap to the rate distortion function (at each distortion value).

Theorem 3.4. *For any $D \in [0, 1/\lambda]$, there exists a constant $c > 0$ (not related to λ or D), such that if*

- $L_1 \geq -\log D$,
- $L_2 \geq -\log \lambda^2 D$,

then, with a choice of d_l as in (3.12)*, the achievable rate pairs obtained from expansion coding schemes are both within c bit gap to Shannon rate distortion function, i.e.*

$$R_1 - R(D_1) \leq c,$$
$$R_2 - R(D_2) \leq c.$$

Proof. See Appendix 3.D. □

Remark 3.5. *We remark that the requirement for highest level is much more restricted than channel coding case. The reason lies that number of levels should be large enough to contain the main part of both rate and distortion. From the proof of Appendix 3.D, it is evident that* $L_2 \geq -\log \lambda$ *is enough to bound the rate, however, another* $-\log \lambda D$ *is required to approximate the distortion closely (if only concerning relative distortion, these extra levels may not be essential).*

3.3.3 Numerical Results

Numerical results showing achievable rates along with the rate distortion limit are given in Figure 3.2. It is evident that both forms of expansion coding perform within a constant gap of the limit over the whole distortion region, which outperforms existing linear and non-linear scalar quantization technique especially in the low distortion regime (since samples are independent, the simulations for vector quantization are expected to be close to scalar quantization and omitted in this result).

Theorem 3.4 shows that this gap is bounded by a constant. Here, numerical results show that the gap is not necessarily as wide as predicted by the theoretical analysis. Especially for the low distortion region, the gap is numerically found to correspond to 0.24 bits and 0.43 bits for each coding scheme respectively.

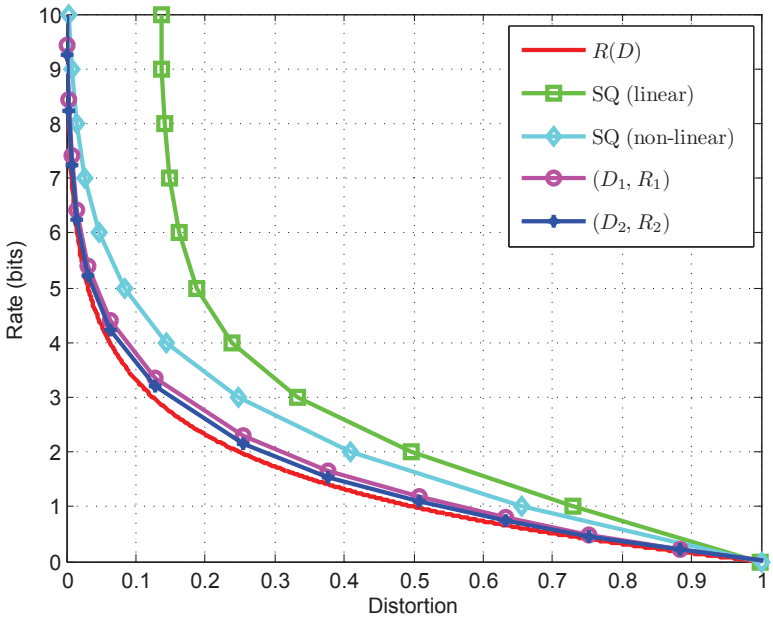

Figure 3.2: **Achievable rate distortion pairs using expansion coding for exponential distribution with one-sided error distortion.** In this numerical result, we set $\lambda = 1$. $R(D)$ (red) is rate distortion limit; (D_1, R_1) (purple) is given by Theorem 3.2; (D_2, R_2) (blue) is given by Theorem 3.3. Linear and non-linear scalar quantization methods are simulated for comparison.

3.4 Expansion Source Coding: Laplacian Source

In this section, we move on to introduce another example, where expansion source coding can be effectively adopted.

3.4.1 Problem Setup for Laplacian Source Coding

Consider an i.i.d. Laplacian source sequence $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_N$, i.e., omitting index i, the probability density function is given by

$$f_{\mathsf{X}}(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R},$$

where λ^{-1} is the magnitude's mean of Laplace distribution, i.e., $\mathbb{E}[|\mathsf{X}|] = 1/\lambda$.

Distortion measure here is the absolute value error distortion, i.e.

$$d(x_{1:N}, \tilde{x}_{1:N}) = \frac{1}{N}\sum_{i=1}^{N} |x_i - \tilde{x}_i|.$$

Lemma 3.6 ([42])**.** *The rate distortion function for Laplacian source with absolute error distortion is given by*

$$R(D) = \begin{cases} -\log(\lambda D), & 0 \le D \le \frac{1}{\lambda}, \\ 0, & D > \frac{1}{\lambda}. \end{cases} \tag{3.13}$$

Moreover, the optimal conditional distribution is

$$f^*_{\mathsf{X}|\tilde{\mathsf{X}}}(x|\tilde{x}) = \frac{1}{2D} e^{-|x-\tilde{x}|/D}, \quad \forall\, x, \tilde{x} \in \mathbb{R}. \tag{3.14}$$

Proof. The proof is given by [42], where the noise in test channel is given by Laplacian with parameter $1/D$. See also Appendix 3.E. □

3.4.2 Expansion Coding for Laplacian Source

By noting that Laplacian is symmetric and two-sided exponential, the expansion of source and its estimate over levels ranging from $-L_1$ to L_2 can be expressed as

$$\hat{\mathsf{X}}_i = \mathsf{X}_i^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \mathsf{X}_{i,l}, \quad i = 1, 2, \ldots, N, \tag{3.15}$$

$$\hat{\tilde{\mathsf{X}}}_i = \tilde{\mathsf{X}}_i^{\text{sign}} \sum_{l=-L_1}^{L_2} 2^l \tilde{\mathsf{X}}_{i,l}, \quad i = 1, 2, \ldots, N, \tag{3.16}$$

where $\mathsf{X}_i^{\text{sign}}$ and $\tilde{\mathsf{X}}_i^{\text{sign}}$ represent the sign of $\hat{\mathsf{X}}_i$ and $\hat{\tilde{\mathsf{X}}}_i$ correspondingly, both random variables uniformly distributed from $\{-1, +1\}$.

In a manner similar to exponential source coding case, expansion reduces the original problem to coding for a set of independent binary sources. However, particularly for Laplacian case, we let $\mathsf{X}^{\text{sign}} = \tilde{\mathsf{X}}^{\text{sign}}$, i.e. using 1 bit to perfectly recover the sign bit. This scheme is more significant and efficient in the low distortion region, since a decoding error of sign bit leads to huge distortion. Then, for the other levels, we formulate each as a binary source coding with Hamming distortion, i.e., $d_{\text{H}}(x_l, \tilde{x}_l) = \mathbf{1}_{\{x_l \neq \tilde{x}_l\}}$. In particular, for level l, we design a symmetric test channel from $\tilde{\mathsf{X}}_l$ to X_l, where the cross probability is given by

$$d_l = \frac{p_l - \tilde{p}_l}{1 - 2\tilde{p}_l},$$

which also implies $\mathbb{E}[|\mathsf{X} - \tilde{\mathsf{X}}|] = d_l$.

Hence, the achievable rate at level l is given by

$$R_l = H(p_l) - H(d_l). \tag{3.17}$$

Due to the decomposability of exponential distribution, the levels after expansion are independent. Based on this property, we have the following result.

Theorem 3.7. *For Laplacian source X, expansion source coding, where the estimate $\tilde{\mathsf{X}}$ is constructed as in the form of* (3.16)*, achieves the rate distortion pair (R, D) with*

$$R = 1 + \sum_{l=-L_1}^{L_2} \left[H(p_l) - H(d_l) \right],$$

for any $L_1, L_2 > 0$ and d_l such that $\mathbb{E}[|\mathsf{X} - \tilde{\mathsf{X}}|] \leq D$.

Note that the absolute value error distortion $\mathbb{E}[|\mathsf{X}-\tilde{\mathsf{X}}|]$ cannot be written as a form of simple weighted sum from Hamming distortions of expanded levels. In fact, we have to use an induction method to characterize the complicated relation. Denote

$$\Delta_k \triangleq \mathbb{E}\left[\left|\sum_{l=-L_1}^{k} 2^l(\mathsf{X}_l - \tilde{\mathsf{X}}_l)\right|\right], \tag{3.18}$$

for any $-L_1 \leq k \leq L_2$, which represents the accumulative distortion up to level k.

1) Initialization: at level $-L_1$,

$$\Delta_{-L_1} = 2^{-L_1} d_{-L_1}.$$

2) Induction: for levels $-L_1 + 1 \leq k \leq L_2$,

$$\begin{aligned}\Delta_k =& \Delta_{k-1}(1-d_k) + 2^k d_k \\ &+ \frac{2^k d_k(1-2p_k)}{1-2d_k} \sum_{l=-L_1}^{k-1} \frac{2^l d_l(1-2p_l)}{1-2d_l}.\end{aligned}$$

To this end, the expansion based coding scheme can be clearly expressed as an optimization problem with variables $\{d_{-L_1}, \ldots, d_{L_2}\}$, but not convex. We have to step back to heuristically choose the value of d_ls in order to get a suboptimal result. More precisely, targeting a distortion D, we construct a set of distortions d_l at each level as

$$d_l = \frac{1}{1+e^{2^l/D}}. \tag{3.19}$$

Then, by Theorem 3.7 and via the iterative algorithm above, we claim the rate distortion pair (R_1, D_1) is achievable, where

$$R_1 = 1 + \sum_{l=-L_1}^{L_2} \left[H(p_l) - H(d_l) \right], \tag{3.20}$$

$$D_1 = \Delta_{L_2}. \tag{3.21}$$

Evidently, this coding scheme may not behave well at high distortion region, since R_1 is at least 1 (which is utilized to represent the sign bit). In the high-distortion regime, precisely compressing the sign bit seems inefficient. To this end, a time sharing scheme is utilized to reduce the gap in high distortion region. More precisely, for any $\rho \in [0, 1]$, we compress ρ fraction of source sequences into codeword 0 (whose corresponding distortion is $1/\lambda$), then the following rate distortion pair is found to be achievable:

$$R_2 = (1 - \rho)R_1, \tag{3.22}$$

$$D_2 = (1 - \rho)D_1 + \rho/\lambda. \tag{3.23}$$

The following theorem provides an upper bound on rate distortion gap of expansion coding scheme.

Theorem 3.8. *For any $D \in [0, 1/\lambda]$, there exists a constant $c' > 0$ (not related to λ or D), such that if*

- $L_1 \geq -\log D$,
- $L_2 \geq -\log \lambda^2 D$,

then, with a choice of d_l as in (3.19)*, the achievable rate pairs obtained from expansion coding schemes are both within c' bit gap to Shannon rate distortion function, i.e.*

$$R_1 - R(D_1) \leq c',$$
$$R_2 - R(D_2) \leq c'.$$

Proof. See Appendix 3.F. □

3.4.3 Numerical Results

We find that the expansion coding scheme is provably within 1 bit constant gap of the rate distortion function. At this point, the calculation of R_1 is fairly tight, however, the upper bound on D_1 can be loose, especially in the low distortion regime. As the calculation of D_1 from d_ls is non-trivial, it is hard to characterize the extent to which the overall distortion is overestimated by the bound. Here, we numerically analyze this gap, and found it to be 0.52 bits in the low distortion regime. (See Figure 3.3.)

3.5 Summary

Similar to the case of channel coding, we utilize expansion source coding to adopt discrete source codes achieving rate distortion limit on each level after expansion, and design codes achieving near-optimal performance for the original source. Theoretical analysis and numerical results are provided to detail performance guarantees of the proposed expansion coding scheme.

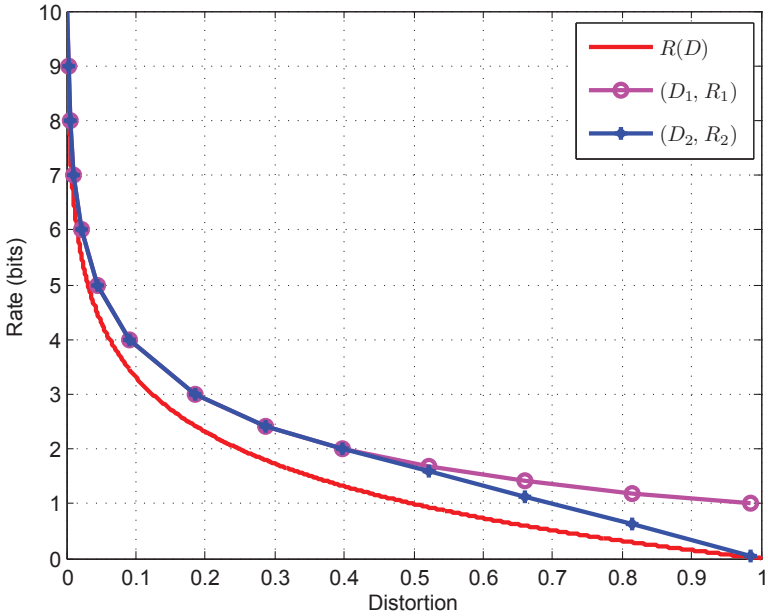

Figure 3.3: **Achievable rate distortion pairs using expansion coding.** In this numerical result, we set $\lambda = 1$. $R(D)$ (red) is rate distortion limit; (D_1, R_1) (purple) is achievable rate using expansion coding: and (D_2, R_2) (blue) is achievable rate using expansion coding and time sharing.

From the analyses in Chapter 2 and Chapter 3, expansion coding is proved to be an effective coding scheme for both data transmission over analog noise channel, and data compression of analog sources. The advantages of expansion coding scheme lie in its ability to shoot for near optimal rate, and to guarantee coding complexities tractable at the same time. Although only two examples are illustrated to show the effectiveness of expansion coding, it is believed to be a uniform coding framework for general coding problems.

3.A Proof of Lemma 3.1

Note that the maximum entropy theorem implies that the distribution maximizing differential entropy over all probability densities f_{X} on support set $\mathbb{R}^+$ satisfying

$$\int_0^\infty f\mathsf{X}(x)xdx = 1, \qquad \int_0^\infty f\mathsf{X}(x)xdx = 1/\lambda,$$

is exponential distribution with parameter λ. Based on this result, in order to satisfy $\mathbb{E}[d(\mathsf{X},\tilde{\mathsf{X}})] \leq D$, where $d(\mathsf{X},\tilde{\mathsf{X}}) = \infty$ for $\mathsf{X} < \tilde{\mathsf{X}}$, we have to restrict $\mathsf{X} \geq \hat{\mathsf{X}}$ with probability 1. To this end, we have

$$\begin{aligned} I(\mathsf{X};\tilde{\mathsf{X}}) &= h(\mathsf{X}) - h(\mathsf{X}|\tilde{\mathsf{X}}) \\ &= \log(e/\lambda) - h(\mathsf{X}-\tilde{\mathsf{X}}|\tilde{\mathsf{X}}) \\ &\geq \log(e/\lambda) - h(\mathsf{X}-\tilde{\mathsf{X}}) \\ &\geq \log(e/\lambda) - \log(e\mathbb{E}[\mathsf{X}-\tilde{\mathsf{X}}]) \\ &\geq \log(e/\lambda) - \log(eD) \\ &= -\log(\lambda D). \end{aligned}$$

Here, we need $\mathsf{X}-\tilde{\mathsf{X}}$ to be exponentially distributed and independent with $\tilde{\mathsf{X}}$ as well. More specifically, we can consider a test channel from $\tilde{\mathsf{X}}$ to X with additive noise $\mathsf{Z} = \mathsf{X}-\tilde{\mathsf{X}}$ distributed as exponential with parameter $1/D$, which gives the conditional distribution given by (3.4).

3.B Proof of Theorem 3.2

Due to decomposability of exponential distribution, the levels after expansion are independent, hence, the achievable rate in this theorem is obtained by additions of individual rates. On the other hand, for the calculation of distortion, we have

$$\begin{aligned} D_1 &= \mathbb{E}[\mathsf{X} - \hat{\tilde{\mathsf{X}}}] \\ &= \mathbb{E}\left[\sum_{l=-\infty}^{\infty} 2^l \mathsf{X}_l - \sum_{l=-L_1}^{L_2} 2^l \tilde{\mathsf{X}}_l\right] \\ &\overset{(a)}{=} \sum_{l=-L_1}^{L_2} 2^l d_l + \sum_{l=L_2+1}^{\infty} 2^l p_l + \sum_{l=-\infty}^{-L_1-1} 2^l p_l \\ &\overset{(b)}{\leq} \sum_{l=-L_1}^{L_2} 2^l d_l + \sum_{l=L_2+1}^{\infty} 2^{-l+1}/\lambda^2 + \sum_{l=-\infty}^{-L_1-1} 2^{l-1} \\ &\overset{(c)}{=} \sum_{l=-L_1}^{L_2} 2^l d_l + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1}, \end{aligned} \tag{3.24}$$

where

(a) follows from $p_l - \tilde{p}_l = d_l$;

(b) follows from

$$p_l = \frac{1}{1+e^{\lambda 2^l}} \leq \frac{1}{1+(1+\lambda 2^l + \lambda^2 2^{2l}/2)} \leq \frac{1}{\lambda^2 2^{2l}/2} = 2^{-2l+1}/\lambda^2,$$

and $p_l < 1/2$ for any l;

(c) follows from

$$\sum_{l=L_2+1}^{\infty} 2^{-l} = 2^{-L_2}, \text{ and } \sum_{l=-\infty}^{-L_1-1} 2^l = 2^{-L_1}.$$

3.C Proof of Theorem 3.3

By the design of coding scheme, if all higher levels are decoded as equivalence, then they must be encoded with one-sided distortion. Recall that for Z-channel, we have

$$\Pr\{\mathsf{X}_l \neq \tilde{\mathsf{X}}_l\} = \Pr\{\mathsf{X}_l = 1, \tilde{\mathsf{X}}_l = 0\} = d_l.$$

Hence, due to independence of expanded levels,

$$\rho_l = \prod_{k=l+1}^{L_2} (1 - d_k).$$

Then, at each level, the achievable rate is $R_{\mathrm{Z},l}$ with probability ρ_l and is $R_{\mathrm{X},l}$ otherwise. From this, we obtain the expression of R_2 given by the theorem. On the other hand, since in both cases we have $p_l - \tilde{p}_l = d_l$, the form of distortion remains the same.

3.D Proof of Theorem 3.4

Denote $\gamma = -\log\lambda$, and $\xi = -\log D$. Then, from $D \leq 1/\lambda$, we have

$$\gamma + \xi \geq 0. \tag{3.25}$$

By noting that p_l and d_l are both expanded parameters from exponential distribution, we have

$$p_l = \frac{1}{1 + e^{\lambda 2^l}} = \frac{1}{1 + e^{2^{l-\gamma}}},$$
$$d_l = \frac{1}{1 + e^{2^l/D}} = \frac{1}{1 + e^{2^{l+\xi}}}.$$

Hence, p_l is shifted version of d_l (analog to the channel coding case), i.e.,

$$d_l = p_{l+\gamma+\xi}. \tag{3.26}$$

Using this relationship, we obtain

$$\begin{aligned}
\sum_{l=-L_1}^{L_2} [H(p_l) - H(d_l)] &\overset{(a)}{=} \sum_{l=-L_1}^{L_2} H(p_l) - \sum_{l=-L_1}^{L_2} H(p_{l+\gamma+\xi}) \\
&= \sum_{l=-L_1}^{L_2} H(p_l) - \sum_{l=-L_1+\gamma+\xi}^{L_2+\gamma+\xi} H(p_l) \\
&\overset{(b)}{=} \sum_{l=-L_1}^{-L_1+\gamma+\xi-1} H(p_l) - \sum_{l=L_2+1}^{L_2+\gamma+\xi} H(p_l) \\
&\overset{(c)}{\leq} \gamma + \xi,
\end{aligned} \tag{3.27}$$

where

(a) follows from (3.26);

(b) follows from (3.25) and theorem assumptions;

(c) follows from $0 \leq H(p_l) \leq 1$ for any l.

From the expression of R_1, we have

$$\begin{aligned}
R_1 &= \sum_{l=-L_1}^{L_2} \left[H(p_l) - (1 - p_l + d_l) H\left(\frac{d_l}{1 - p_l + d_l} \right) \right] \\
&= \sum_{l=-L_1}^{L_2} [H(p_l) - H(d_l)] + \sum_{l=-L_1}^{L_2} \left[H(d_l) - (1 - p_l + d_l) H\left(\frac{d_l}{1 - p_l + d_l} \right) \right] \\
&\leq \gamma + \xi + \sum_{l=-L_1}^{L_2} \left[H(d_l) - (1 - p_l + d_l) H\left(\frac{d_l}{1 - p_l + d_l} \right) \right],
\end{aligned} \tag{3.28}$$

where (3.27) has been utilized in the last inequality. To this end, we need to bound

$$\begin{aligned}\Delta_l &\triangleq H(d_l) - (1-p_l+d_l)H\left(\frac{d_l}{1-p_l+d_l}\right)\\ &= (1-p_l)\log(1-p_l) - (1-d_l)\log(1-d_l) - (1-p_l+d_l)\log(1-p_l+d_l).\end{aligned}$$

For this, two cases are considered:

1) For $l \leq -\xi$, d_l and p_l are close and both tend to 0.5. More precisely, we have

$$\begin{aligned}\Delta_l &\overset{(d)}{\leq} -(1-p_l+d_l)\log(1-p_l+d_l)\\ &\overset{(e)}{\leq} 2\log e\cdot(p_l-d_l)\\ &\overset{(f)}{\leq} 2\log e\cdot\left[\frac{1}{2}-\left(\frac{1}{2}-2^{l+\xi-1}\right)\right]\\ &= \log e\cdot 2^{l+\xi},\end{aligned} \tag{3.29}$$

where

(d) follows from the fact that $(1-\alpha)\log(1-\alpha)$ is a decreasing function over $[0, 0.5]$, hence, $(1-p_l)\log(1-p_l) \leq (1-d_l)\log(1-d_l)$;

(e) follows from the observation that $-(1-\alpha)\log(1-\alpha) \leq 2\log e\cdot\alpha$ for any $\alpha \in [0, 0.5]$;

(f) follows from the fact that $p_l \leq 0.5$ and

$$d_l = \frac{1}{1+e^{2^{l+\xi}}} \geq \frac{1}{1+(1+2\cdot 2^{l+\xi})} \geq \frac{1}{2} - 2^{l+\xi-1},$$

where the first inequality is due to $e^{\alpha} \leq 1+2\alpha$ for any $\alpha \in [0,1]$ ($2^{l+\xi} \leq 1$ due to $l \leq -\xi$), and the last inequality holds for any l.

2) On the other hand, for $l > -\xi$, d_l tends to 0, so as $1 - p_l$ and $1 - p_l + d_l$ get close. More precisely, we have

$$\begin{aligned}\Delta_l &\overset{(g)}{\leq} -(1-d_l)\log(1-d_l) \\ &\overset{(h)}{\leq} 2\log e \cdot d_l \\ &\overset{(i)}{\leq} \log e \cdot 2^{-l-\xi}, \end{aligned} \tag{3.30}$$

where

(g) follows from the fact $(1-p_l)\log(1-p_l) \leq (1-p_l+d_l)\log(1-p_l+d_l)$;

(h) follows from the observation that $-(1-\alpha)\log(1-\alpha) \leq 2\log e \cdot \alpha$ for any $\alpha \in [0, 0.5]$;

(i) follows from the fact that

$$d_l = \frac{1}{1+e^{2^{l+\xi}}} \leq \frac{1}{e^{2^{l+\xi}}} \leq \frac{1}{2 \cdot 2^{l+\xi}} = 2^{-l-\xi-1},$$

where the second inequality holds from $e^\alpha \geq 2\alpha$ for any $\alpha > 1$ ($2^{l+\xi} > 1$ due to $l > -\xi$).

Putting (3.29) and (3.30) back to (3.28), we have

$$\begin{aligned} R_1 &\leq \gamma + \xi + \log e \cdot \sum_{l=-L_1}^{-\xi} 2^{l+\xi} + \log e \cdot \sum_{l=-\xi+1}^{L_2} 2^{-l-\xi} \\ &\leq \gamma + \xi + 2\log e + \log e \\ &= R(D) + 3\log e, \end{aligned} \tag{3.31}$$

where we use the definitions of γ and ξ, such that $\gamma + \xi = R(D)$.

Finally, using the result from Theorem 3.2 that

$$D \leq D_1 \leq \sum_{l=-L_1}^{L_2} 2^l d_l + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1} \leq D + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1}, \quad (3.32)$$

we obtain

$$\begin{aligned} R(D) &\overset{(j)}{\leq} R(D_1) + \frac{\log e}{D}(D_1 - D) \\ &\overset{(k)}{\leq} R(D_1) + \frac{\log e}{D}(2^{-L_2+1}/\lambda^2 + 2^{-L_1-1}) \\ &\overset{(l)}{\leq} R(D_1) + 2.5 \log e, \end{aligned} \quad (3.33)$$

where

(j) follows from $R(D)$ is convex such that for any α_1 and α_2,

$$R(\alpha_1) \geq R(\alpha_2) + R'(\alpha_2)(\alpha_1 - \alpha_2),$$

where $R'(\alpha_2) = -\log e/\alpha_2$ is the derivative of $R(\cdot)$, and setting $\alpha_1 = D_1$, $\alpha_2 = D$ completes the proof of this step;

(k) follows from (3.32);

(l) follows from theorem assumptions that $L_1 \geq -\log D$ and $L_2 \geq -\log \lambda^2 D$.

Combining (3.33) with (3.31), we have

$$R_1 \leq R(D_1) + 5.5 \log e,$$

which completes the proof for R_1 and D_1 by taking $c = 5.5 \log e$.

For the other part of the theorem, observe that

$$H\left(\frac{d_l}{1-2p_l+2d_l}\right) \geq (1-p_l+d_l)H\left(\frac{d_l}{1-p_l+d_l}\right).$$

Hence, for any $-L_1 \leq l \leq L_1$, we have $R_{\mathrm{X},l} \leq R_{\mathrm{Z},l}$. Thus, we have $R_2 \leq R_1$. Combing with the observation that $D_1 = D_2$, we have $R_2 \leq R_1 + 5.5 \log e$.

Note that in the process of providing bounds, the actual gap may be enlarged. A precise value of the gap can be estimated from numerical results, which is 0.24 bit and 0.43 bit for each coding scheme respectively.

3.E Proof of Lemma 3.6

Maximum entropy theorem implies that Laplace distribution with parameter λ has the maximum differential entropy $h(f_{\mathsf{X}})$ over all probability densities f_{X} on support set $\mathbb{R}$ satisfying

$$\int_{-\infty}^{\infty} f_{\mathsf{X}}(x)x dx = 1,$$
$$\int_{-\infty}^{\infty} f_{\mathsf{X}}(x)|x| dx = 1/\lambda.$$

Based on this result, it is evident to note that

$$\begin{aligned} I(\mathsf{X};\tilde{\mathsf{X}}) &= h(\mathsf{X}) - h(\mathsf{X}|\tilde{\mathsf{X}}) \\ &= \log\left(\frac{2e}{\lambda}\right) - h(\mathsf{X}-\hat{\mathsf{X}}|\hat{\mathsf{X}}) \\ &\geq \log\left(\frac{2e}{\lambda}\right) - h(\mathsf{X}-\tilde{\mathsf{X}}) \\ &\geq \log\left(\frac{2e}{\lambda}\right) - \log(2e \cdot \mathbb{E}[|\mathsf{X}-\hat{\mathsf{X}}|]) \end{aligned}$$

$$
\begin{aligned}
&\geq \log\left(\frac{2e}{\lambda}\right) - \log(2eD) \\
&= -\log(\lambda D),
\end{aligned}
$$

where we have used the fact that $\mathbb{E}[|\mathsf{X} - \tilde{\mathsf{X}}|] \leq D$. Here, we need $\mathsf{X} - \tilde{\mathsf{X}}$ to be Laplace distributed and independent with $\tilde{\mathsf{X}}$ as well. More specifically, we can design a test channel from $\tilde{\mathsf{X}}$ to X with additive noise $\mathsf{Z} = \mathsf{X} - \tilde{\mathsf{X}}$ distributed as Laplace with parameter $1/D$, as shown in (3.14).

3.F Proof to Theorem 3.7

Note that the expressions of p_l and d_l are the same as the exponential case (although the potential test channel models are different), and the theorem assumptions are identical. Then, from (3.27), we already have

$$
\sum_{l=-L_1}^{L_2} [H(p_l) - H(d_l)] \leq -\log(\lambda D). \tag{3.34}
$$

Moreover, noting that $\mathbb{E}[|\mathsf{X}_l - \tilde{\mathsf{X}}_l|] = d_l$, we have

$$
\begin{aligned}
D_1 &= \mathbb{E}\left[\left|\sum_{l=-\infty}^{\infty} 2^l \mathsf{X}_l - \sum_{l=-L_1}^{L_2} 2^l \tilde{\mathsf{X}}_l\right|\right] \\
&\leq \sum_{l=-L_1}^{L_2} 2^l \mathbb{E}[|\mathsf{X}_l - \tilde{\mathsf{X}}_l|] + \sum_{l=-\infty}^{-L_1-1} 2^l \mathbb{E}[|\mathsf{X}_l|] + \sum_{l=L_2+1}^{\infty} 2^l \mathbb{E}[|\mathsf{X}_l|] \\
&= \sum_{l=-L_1}^{L_2} 2^l d_l + \sum_{l=-\infty}^{-L_1-1} 2^l p_l + \sum_{l=L_2}^{\infty} 2^l p_l \\
&\leq \sum_{l=-L_1}^{L_2} 2^l d_l + 2^{-L_2+1}/\lambda^2 + 2^{-L_1-1},
\end{aligned} \tag{3.35}
$$

where the last inequality holds analog to (3.24). Hence, following the same steps in (3.33), we obtain

$$R(D) \leq R(D_1) + 2.5 \log e. \tag{3.36}$$

Combining the pieces together, we obtain

$$\begin{aligned} R_1 &= 1 + \sum_{l=-L_1}^{L_2} [H(p_l) - H(d_l)] \\ &\overset{(a)}{\leq} 1 - \log(\lambda D) \\ &= 1 + R(D) \\ &\overset{(b)}{\leq} 1 + R(D_1) + 2.5 \log e, \end{aligned}$$

where

(a) is due to (3.34);

(b) is due to (3.36).

Hence, by choosing $c' = 1 + 2.5 \log e$ complete the proof for R_1 and D_1.

On the other hand, (R_2, D_2) is obtained by convex combination of (R_1, D_1) and $(0, 1/\lambda)$, which means $R_2 \leq R_1$ and $D_2 \geq D_1$. Thus, utilizing the result for R_1 and D_1, we have

$$R_2 \leq R_1 \leq R(D_1) + 1 + 2.5 \log e \leq R(D_2) + 1 + 2.5 \log e.$$

Chapter 4

Polar Coding for Fading BSCs

4.1 Background of Polar Coding for Fading Channels

Polar codes are the first family of provably capacity achieving codes for symmetric binary-input discrete memoryless channels (B-DMC) with low encoding and decoding complexity [14] [43]. These codes polarize the underlying channel in the sense that, via channel combining and channel splitting stages, multiple uses of the given channel are transformed into equivalent polarized ones: either purely noisy (referred to as "bad" channel instances) or noiseless (referred to as "good" channel instances). Then, information symbols are mapped to the good instances of polarized channels, whereas channel inputs corresponding to the bad instances are fixed and shared between the transmitter and receiver. It is shown in [14] that the fraction of the good channel instances approaches the symmetric capacity of the channel, which is equal to the capacity of the underlying channel if the channel is symmetric. That is, polar codes achieve the capacity of symmetric B-DMCs. This phenomenon of channel polarization has then been generalized to arbitrary discrete memoryless channels with a construction complexity to the same order and a similar error probability behavior [44].

However, the analysis of polar coding for fading channels, with either discrete-valued or continuous-valued noises, is still limited. Recent work [45] investigates binary input real number output AWGN fading channel, where the fading coefficient is assumed to be one of the two states with equal probabilities. These fading coefficients are assumed to follow arbitrary distributions with the requirement of satisfying some tail probability constraints. For this setup, the authors proposed polar coding schemes where symbols are multiplexed in a specific fashion at the encoder. In particular, the paper analyzes diagonal, horizontal, and uniform multiplexers; and, the corresponding diversity and outage analysis have been performed. Another recent work [46] focuses on polar coding schemes for Rayleigh fading channel under two scenarios: block fading with known channel state information (CSI) at the transmitter and fast fading with fading distribution known at the transmitter. For the latter case, the channel is shown to be symmetric, and through quantization of the channel output, the polar coding scheme is shown to achieve a constant gap to the capacity.

In my thesis, we focus on a block fading model without the CSI at the transmitter, and propose a hierarchical polar coding scheme for such channels. More precisely, in this chapter, we focus on fading binary symmetric channel (BSC), which is an important model as it is closely related to an AWGN block fading channel with BPSK modulation and demodulation. Such binary input AWGN models are previously analyzed in [47][48] to evaluate the performance of polar codes over AWGN channels. Here, we focus on communication chan-

nel models that involve fading, where the channel coefficients vary according to a block fading model. This scenario of fading AWGN with BPSK modulation resembles a fading binary symmetric channel model, where each fading block has a cross-over probability depending on the corresponding channel state realization. Specifically, AWGN channel states with higher SNRs map to binary symmetric channels with lower crossover probabilities. For this binary symmetric fading model, we propose a novel polar coding approach that utilizes polarization in a *hierarchical* manner *without* channel state information (CSI) at the transmitter (with channel state statistics assumed to be known at the transmitter). The key factor enabling our coding scheme is the hierarchical utilization of polar coding. More precisely, polar codes are not only designed over channel uses for each fading state, but also utilized over fading blocks. By taking advantage of the degradedness property of channel polarization between different BSCs, an erasure model (over fading blocks) is constructed for every channel instance that polarizes differently depending on the channel states. It is shown that this proposed coding scheme, without instantaneous CSI at the transmitter, achieves the ergodic capacity of the fading binary symmetric channel.

4.2 System Model of Fading BSCs

Fading channels characterize the wireless communication channels, where the channel states vary over channel uses. Fading coefficients typically vary much slower than transmission symbol duration in practice. For such cases,

a block fading model is considered, wherein the channel state is assumed to be a constant over each coherence time interval, and follows a stationary ergodic process across fading blocks. For such a block fading model, we consider the practical scenario where the channel state information is available only at the decoder (CSI-D) [49], while the transmitter is assumed to know only the statistics of the channel states.

Binary symmetric channel (BSC) is a channel with binary input X, binary noise Z, and a binary output $\mathsf{Y} = \mathsf{X} \oplus \mathsf{Z}$. Here, for the fading BSC, the channel noise is a Bernoulli distributed random variable, where its statistics depend on the channel states. For the block fading BSC considered in this work, the channel is modeled as follows.

$$\mathsf{Y}_{b,i} = \mathsf{X}_{b,i} \oplus \mathsf{Z}_{b,i}, \quad b = 1, \ldots, B, \quad i = 1, \ldots, N,$$

where N is the block length, and B is the number of fading blocks. Here, $\mathsf{Z}_{b,i}$ are assumed to be identically distributed within a block and follow an i.i.d. fading process over blocks. That is, if we consider fading BSC with S states, with probability ϱ_s the parameter p_s is chosen for the fading block b, where the channel noise $\mathsf{Z}_{b,i}$ is sampled from a Bernoulli random variable with parameter p_s for all $i \in \{1, \ldots, N\}$. Here, $1 \leq s \leq S$ and $\sum_{s=1}^{S} \varrho_s = 1$. See Figure 4.1 for an illustration of defined fading BSCs with two states.

In wireless communications, the fading binary symmetric channel is utilized to model a fading AWGN channel with BPSK modulation and demodulation. In particular, for a fading AWGN channel with input power constraint

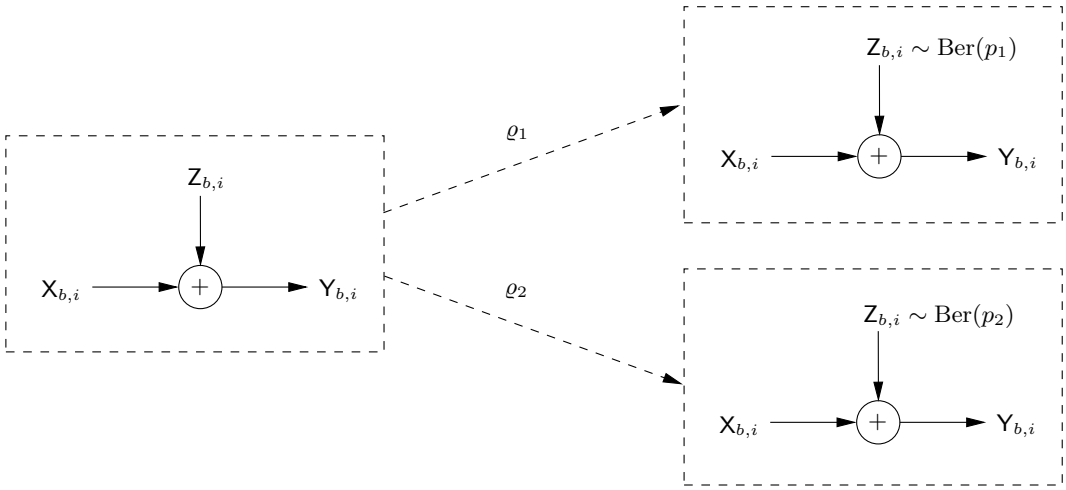

Figure 4.1: **Illustration of fading binary symmetric channels with two states.** Within a particular block b, the noise random variables $\mathsf{Z}_{b,i}$ are identically distributed. Moreover, with probability ϱ_1, they are identically distributed as $\mathrm{Ber}(p_1)$, while with the rest probability ϱ_2, they are identically distributed as $\mathrm{Ber}(p_2)$.

P_{X}, the channel noise is distributed as i.i.d. Gaussian with variance P_{Z}, and the channel gain (the factor H in the AWGN channel $\hat{\mathsf{Y}} = \mathsf{H}\hat{\mathsf{X}} + \hat{\mathsf{Z}}$) remains the same statistic within a fading block, and follows an ergodic process over different blocks. After utilizing the BPSK modulation and demodulation at the encoder and decoder, respectively, the equivalent channel is a binary input and binary output channel, with transition probability relating to AWGN channel state. More precisely, if the channel gain $\mathsf{H}_{b,i}$ within a particular fading block b is equal to h_s with probability ϱ_s for some $s \in \{1, 2, \ldots, S\}$, then the corresponding binary noise in the equivalent fading BSC has the statistics of

$$p_s \triangleq \Pr\{\mathsf{Z}_{b,i} = 1\} = 1 - \Phi(h_s\sqrt{\mathrm{SNR}}),$$

where $\Phi(\cdot)$ is the CDF of normal distribution and SNR is the signal-to-noise ratio, i.e. $\text{SNR} = P_{\mathsf{X}}/P_{\mathsf{Z}}$. In other words, the channel at each fading block can be modeled as $\mathcal{W}_s \triangleq \text{BSC}(p_s)$ with probability ϱ_s.

The ergodic capacity of a fading binary symmetric channel is given by [50]

$$C_{\text{CSI-D}} = \sum_{s=1}^{S} [1 - H(p_s)], \tag{4.1}$$

where $H(\cdot)$ is the binary entropy function, and CSI-D refers to channel state information at the decoder. Note that, the ergodic capacity of fading BSC is an average over the capacities of all possible channels corresponding to different channel states. In this section, we propose a polar coding scheme that achieves the capacity of this fading BSC with low encoding and decoding complexity, without having instantaneous channel state information at the transmitter. Towards this end, we first focus on a fading BSC with two channel states, and then generalize our results to arbitrary finite number of channel states.

4.3 Intuition

In polar coding for a BSC, we see that the channel can be polarized by transforming a set of independent copies of given channels into a new set of channels whose symmetric capacities tend to 0 or 1 (for all but a vanishing fraction of indices). Towards applying such a polarization phenomenon to fading BSC, we first focus on how two binary symmetric channels polarize at the same time. We summarize a result given in [51] regarding the polarization

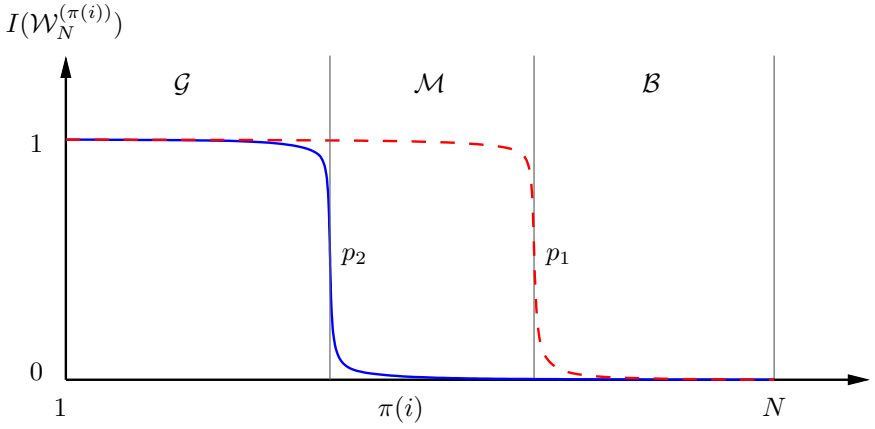

Figure 4.2: **Illustration of polarizations for two binary symmetric channels.** The blue-solid line represents the degraded channel with transition probability p_2, and the red-dashed one represents the superior channel with p_1 ($p_1 \leq p_2$). Values of $I(\mathcal{W}_N^{(\pi(i))})$, the reordered symmetric mutual information, are shown for both channels.

of degraded channels.

Lemma 4.1 ([51]). *For two binary symmetric channels $\mathcal{W}_1 \triangleq BSC(p_1)$ and $\mathcal{W}_2 \triangleq BSC(p_2)$, if $\mathcal{W}_2$ is degraded with respect to $\mathcal{W}_1$, i.e. $p_1 \leq p_2$, then for any channel index $i \in \{1, \ldots, N\}$, the reconstructed channels after polarization have the relationship that $\mathcal{W}_{2,N}^{(i)}$ is degraded with respect to $\mathcal{W}_{1,N}^{(i)}$, and hence $I(\mathcal{W}_{2,N}^{(i)}) \leq I(\mathcal{W}_{1,N}^{(i)})$.*

That is, when polarizing two binary symmetric channels, the reconstructed channels of the degraded channel have lower symmetric rate compared to that of the other channel. This statement also implies that

$$\mathcal{A}_2 \subseteq \mathcal{A}_1,$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ denote the information sets of the superior and degraded channels, respectively. This relationship is illustrated in Figure 4.2. Based on this observation, when polarizing two BSCs, the channel indices after reordering permutation π can be divided into three categories (we assume that channel 2 is degraded, i.e., $p_1 \leq p_2$):

1) Set $\mathcal{G}$: both channels are good, i.e.,

$$I(\mathcal{W}_{1,N}^{(\pi(i))}) \to 1, \quad I(\mathcal{W}_{2,N}^{(\pi(i))}) \to 1.$$

2) Set $\mathcal{M}$: only channel 1 is good, while channel 2 is bad, i.e.,

$$I(\mathcal{W}_{1,N}^{(\pi(i))}) \to 1, \quad I(\mathcal{W}_{2,N}^{(\pi(i))}) \to 0.$$

3) Set $\mathcal{B}$: both channels are bad, i.e.,

$$I(\mathcal{W}_{1,N}^{(\pi(i))}) \to 0, \quad I(\mathcal{W}_{2,N}^{(\pi(i))}) \to 0.$$

We have the following relationships between these sets. First, information sets for two channels are given by $\mathcal{A}_2 = \mathcal{G}$, and $\mathcal{A}_1 = \mathcal{G} \cup \mathcal{M}$. Moreover, considering the sizes of these sets, we have:

$$|\mathcal{G}| = |\mathcal{A}_2| = N \cdot [1 - H(p_2) - \epsilon], \tag{4.2}$$

$$|\mathcal{M}| = |\mathcal{A}_1| - |\mathcal{A}_2| = N \cdot [H(p_2) - H(p_1)], \tag{4.3}$$

$$|\mathcal{B}| = N - |\mathcal{A}_1| = N \cdot [H(p_1) + \epsilon], \tag{4.4}$$

where ϵ is an arbitrarily small positive number (that vanishes as $N \to \infty$).

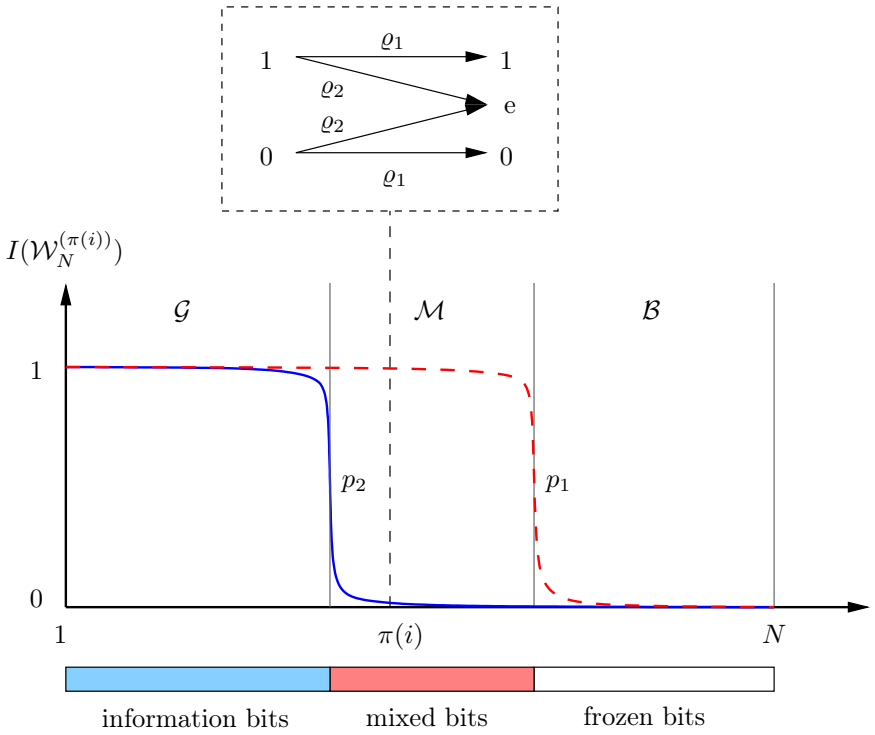

Figure 4.3: **Illustration of polarizations for fading binary symmetric channels.** The blue-solid line represents the degraded state with transition probability p_2, and the red-dashed one represents the superior state with p_1 ($p_1 \leq p_2$). For those channel indices after polarization in mixed set $\mathcal{M}$, an erasure channel is constructed to model its either noiseless or purely noisy behavior.

For a fading binary symmetric channel, we utilize Figure 4.3 to illustrate our coding scheme. Here, consider a fading BSC with only two fading states, the superior state and the degraded one (denoted as state 1 and 2, respectively). If channel is in state 2, which happens with probability ϱ_2, the fading channel polarizes to the blue-solid curve, and otherwise the channel is in state 1, which happens with the probability $\varrho_1 = 1 - \varrho_2$, and the fading

channel polarizes to the red-dashed curve. Hence, the reconstructed channel with index in set $\mathcal{G}$ always polarizes to a good one, i.e., its symmetric mutual information is close to 1 no matter what the fading state is. And, the reconstructed channel with index in set $\mathcal{B}$ always polarizes to a bad one, i.e., its symmetric mutual information is close to 0 no matter what the fading state is. Therefore, one can reliably transmit information for channel instances belonging to $\mathcal{G}$, whereas one may not transmit any information for channel instances belonging to $\mathcal{B}$. The novel part of the proposed coding scheme is for the middle region, i.e., coding over the set $\mathcal{M}$, where reconstructed channels polarize differently depending on the channel states. Since we consider the transmitter has no prior knowledge of channel states before transmitting, coding over channels with indices in $\mathcal{M}$ is challenging. At this point, we observe that for these channels, with probability ϱ_1 they are nearly noiseless, and with probability ϱ_2 they are purely noisy. Thus, each channel can be modeled as a binary erasure channel (BEC) from the viewpoint of blocks, where the erasure probability is equal to ϱ_2. Here, we denote this channel as

$$\mathcal{W}_{\mathrm{e}} \triangleq \mathrm{BEC}(\varrho_2).$$

This observation motivates our design of hierarchical encoder and decoder for fading BSCs.

4.4 Hierarchical Polar Encoder

The encoding process has two phases, hierarchically using polar codes to generate NB-length codewords, where N is blocklength and B is the number

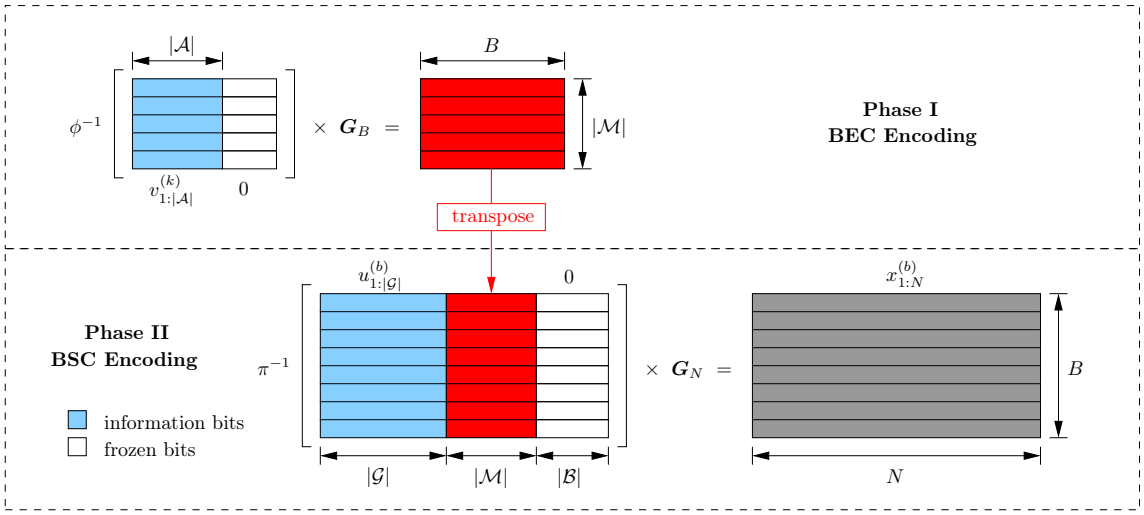

Figure 4.4: **Illustration of proposed polar encoder for a fading binary symmetric channel with two states.** Bits in blue are information bits, and those in white are frozen as zeros. The codewords generated from Phase I are transposed and embedded into the messages of Phase II to generate the final codeword of length NB. ϕ and π are column reordering permutations with respect to BEC and BSC, correspondingly.

of blocks.

4.4.1 Phase I: BEC Encoding

In this phase, we generate $|\mathcal{M}|$ number of BEC polar codes, each with length B. Consider a set of blockwise messages $v^{(k)}_{1:|\mathcal{A}|}$ with $k \in \{1, \ldots, |\mathcal{M}|\}$. For every $v^{(k)}_{1:|\mathcal{A}|}$, we construct polar codeword $\tilde{v}^{(k)}_{1:B}$, which is formed by the $\boldsymbol{G}_B$-coset code with parameter $(B, |\mathcal{A}|, \mathcal{A}, 0)$, where $\mathcal{A}$ is the information set for $\mathcal{W}_{\text{BEC}} = \text{BEC}(\varrho_1)$, and we choose the rate to be optimal, i.e.

$$|\mathcal{A}| = B \cdot (\varrho_1 - \epsilon), \tag{4.5}$$

$$|\mathcal{A}^c| = B \cdot (\varrho_2 + \epsilon). \tag{4.6}$$

In other words, we construct a set of polar codes, where each code corresponds to an index in set $\mathcal{M}$, with the same rate ϱ_1, the same information set $\mathcal{A}$, and the same frozen values, 0. More precisely, if we denote the reordering permutation for $\mathcal{W}_{\mathrm{e}} = \mathrm{BEC}(\varrho_2)$ as ϕ, then

$$\tilde{v}_{1:B}^{(k)} = \nu_{1:B}^{(k)} \times \boldsymbol{G}_B,$$
$$\phi(\nu_{1:B}^{(k)}) = \left[\begin{array}{c|c|c} v_{1:|\mathcal{A}|}^{(k)} & | & 0 \end{array} \right],$$

for every $k \in \{1, \ldots, |\mathcal{M}|\}$. The collection of all $\tilde{v}_{1:B}^{(k)}$ is denoted as a $|\mathcal{M}| \times B$ matrix $\tilde{\boldsymbol{V}}$. We denote $\tilde{\boldsymbol{V}}_b^T$ as the b-th row of the transpose of $\tilde{\boldsymbol{V}}$.

4.4.2 Phase II: BSC Encoding

In this phase, we generate B number of BSC polar codes, each with length N. Consider a set of messages $u_{1:|\mathcal{G}|}^{(b)}$ with $b \in \{1, \ldots, B\}$. For every $u_{1:|\mathcal{G}|}^{(b)}$, construct polar codeword $x_{1:N}^{(b)}$, which is $\boldsymbol{G}_N$-coset code with parameter $(N, |\mathcal{G}|, \mathcal{G}, [\tilde{\boldsymbol{V}}_b^T | 0])$, where $\mathcal{G}$ is BSC information set with size given by (4.2). Remarkably, we do not set all non-information bits to be 0, but transpose the blockwise codewords generated from Phase I and embed them into the messages of this phase. More precisely, if denote the reordering permutation operator of BSC as π, then

$$x_{1:N}^{(b)} = \mu_{1:N}^{(b)} \times \boldsymbol{G}_N.$$
$$\pi\left(\mu_{1:N}^{(b)}\right) = \left[\begin{array}{c|c|c|c} u_{1:|\mathcal{G}|}^{(b)} & | & \tilde{\boldsymbol{V}}_b^T & | \quad 0 \end{array} \right],$$

for every $b \in \{1, \ldots, B\}$. By collecting all $\{x^{(b)}\}_{1:B}$ together, the encoder outputs a codeword with length NB. We equivalently express these codewords

by a $B \times N$ matrix $\boldsymbol{X}$. The proposed encoder for fading binary symmetric channel is illustrated in Figure 4.4.

4.5 Decoder

After receiving the sequence $y_{1:NB}$ from the channel, the decoder's task is to make estimates $\hat{v}_{1:|\mathcal{A}|}^{(k)}$ and $\hat{u}_{1:|\mathcal{G}|}^{(b)}$, such that the information bits in both sets of messages match the ones at the transmitter with high probability. Rewrite channel output $y_{1:NB}$ as a $B \times N$ matrix, with row vectors $y_{1:N}^{(b)}$. As that of the encoding process, the decoding process also works in phases.

4.5.1 Phase I: BSC Decoding for the Superior Channel State

In this phase, we decode part of the output blocks using the BSC SC decoder with respect to the superior channel state. More precisely, since the receiver knows channel states, it can adopt the correct SC decoder (BSC(p_1) SC decoder in this case) to obtain $\hat{\mu}_{1:N}^{(b)}$ from $y_{1:N}^{(b)}$ for every b corresponding to the superior channel state. To this end, the decoder for block b with the superior fading state in this phase is the classical BSC SC polar decoder with parameter p_1, i.e.,

$$\hat{\mu}_i^{(b)} = \begin{cases} 1, & \text{if } i \notin \mathcal{B}, \text{ and } \frac{\mathcal{W}_{1,N}^{(i)}(y_{1:N}^{(b)}, \hat{\mu}_{1:i-1}^{(b)}|1)}{\mathcal{W}_{1,N}^{(i)}(y_{1:N}^{(b)}, \hat{\mu}_{1:i-1}^{(b)}|0)} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

in the order i from 1 to N, and $\mathcal{W}_{1,N}^{(i)}$ is the i-th polarized channel from BSC(p_1). In this phase, one can reliably decode the information bits in blocks with respective to the superior channel states (with the knowledge of frozen

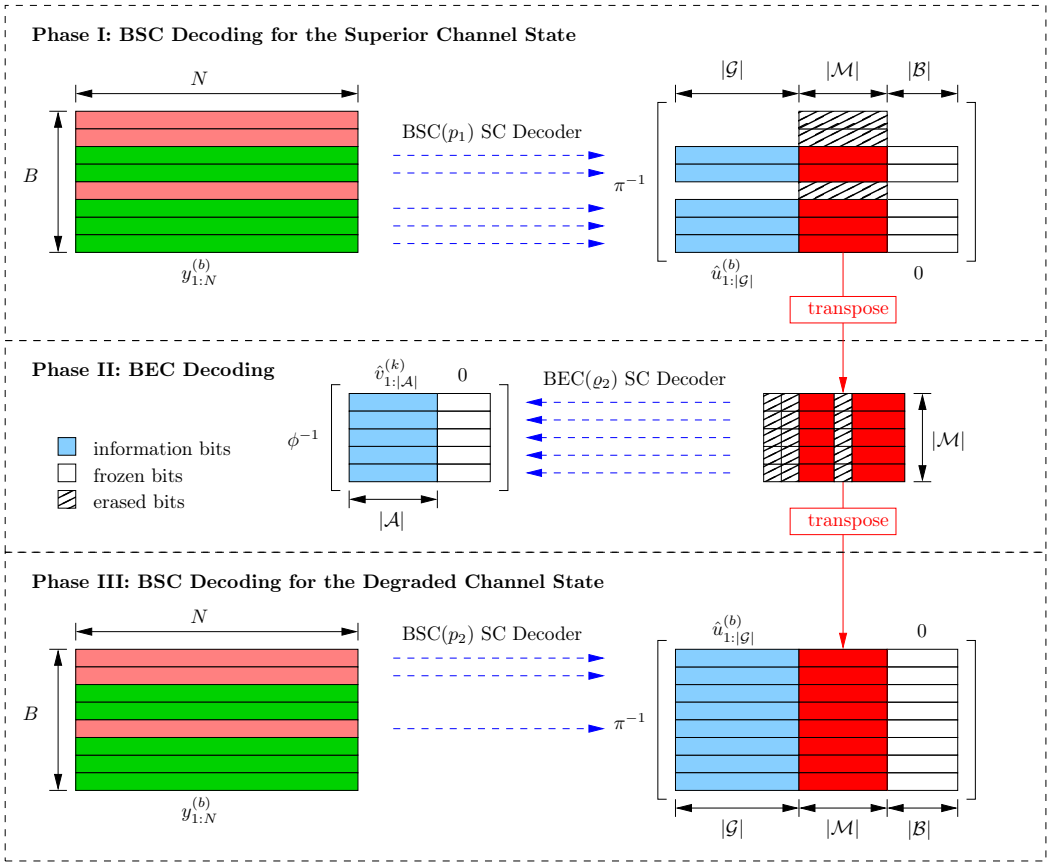

Figure 4.5: **Illustration of proposed polar decoder for a fading binary symmetric channel with two states.** In Phase I, decoder outputs all estimates using BSC SC decoders corresponding to the superior channel state. Selected columns are transposed and delivered as inputs to next phase, by adding all-erasures rows for blocks with the degraded channel state. In Phase II, the decoder continues to use BEC SC decoders to decode all the blockwise information bits, and to recover all erased bits in shade. In Phase III, the BSC SC decoders corresponding to the degraded channel state are utilized to decode the remaining information bits, by taking values of frozen bits in set $\mathcal{M}$ as the decoded results from the previous phase. ϕ and π are column reordering permutations with respect to BEC and BSC, correspondingly.

symbols corresponding to $\mathcal{B}$ indices). Formally, the decoder can declare

$$\pi\left(\hat{\mu}_{1:N}^{(b)}\right) = \left[\begin{array}{c|c|c} \hat{u}_{1:|\mathcal{G}|}^{(b)} & \hat{\tilde{\boldsymbol{V}}}_b^T & 0 \end{array}\right],$$

for every b corresponding to the superior channel state.

However, for the blocks with degraded channel state, one cannot decode reliably because the frozen bits corresponding to set $\mathcal{M}$ are unknown at the decoder (for the degraded channel state, frozen set include $\mathcal{M}$ and $\mathcal{B}$). At this point, we use the next phase to decode these frozen bits using a BEC SC decoder. To proceed, we construct a $B \times |\mathcal{M}|$ matrix $\hat{\tilde{\boldsymbol{V}}}^T$ such that its rows corresponding to the superior state are determined in previous decoding process, while the ones corresponding to the degraded states are all set to erasures. See Figure 4.4 for an intuitive illustration.

4.5.2 Phase II: BEC Decoding

In this phase, we decode the frozen bits with respect to the degraded channel state. More precisely, each row of matrix $\hat{\tilde{\boldsymbol{V}}}$, denoted by $\hat{\tilde{\boldsymbol{V}}}_k$ for $k \in \{1, \ldots, |\mathcal{M}|\}$, is considered as the input to the decoder, and the receiver aims to obtain an estimate of the information bits from it using BEC SC decoder. To this end, the decoder adopted in this phase is the classical BEC SC decoder with parameter ϱ_2, i.e.,

$$\hat{\nu}_b^{(k)} = \begin{cases} 1, & \text{if } b \in \mathcal{A}, \text{ and } \frac{\mathcal{W}_{\mathrm{e},B}^{(b)}(\hat{\tilde{\boldsymbol{V}}}_k, \hat{\nu}_{1:b-1}^{(k)}|1)}{\mathcal{W}_{\mathrm{e},B}^{(b)}(\hat{\tilde{\boldsymbol{V}}}_k, \hat{\nu}_{1:b-1}^{(k)}|1)} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

in the order b from 1 to B, and $\mathcal{W}_{\mathrm{e},B}^{(b)}$ is the b-th polarized channel from BEC(ϱ_2). Then, for every $k \in \{1, \ldots, |\mathcal{M}|\}$, the decoder can declare

$$\phi\left(\hat{\nu}_{1:B}^{(k)}\right) = \left[\begin{array}{c|c|c} \hat{v}_{1:|\mathcal{A}|}^{(k)} & | & 0 \end{array} \right].$$

At this point, the decoder can reconstruct all erased bits as well. More precisely, the erased rows in $\hat{\tilde{\boldsymbol{V}}}^T$ can be recovered, and they can be further utilized to decode the information bits in blocks with the degraded channel state in the next phase.

4.5.3 Phase III: BSC Decoding for the Degraded Channel State

In this phase, the remaining blocks from Phase I are decoded by using BSC SC decoders with respect to degraded channel states. In particular, bits in the frozen set for the degraded channel state (set $\mathcal{B}$ and set $\mathcal{M}$) are known due to the previous phases. Hence, the receiver can decode from $y_{1:N}^{(b)}$ using BSC SC decoder with parameter p_2, i.e.,

$$\hat{\mu}_i^{(b)} = \begin{cases} 1, & \text{if } i \in \mathcal{G}, \text{ and } \frac{\mathcal{W}_{2,N}^{(i)}(y_{1:N}^{(b)}, \hat{\mu}_{1:i-1}^{(b)}|1)}{\mathcal{W}_{2,N}^{(i)}(y_{1:N}^{(b)}, \hat{\mu}_{1:i-1}^{(b)}|0)} \geq 1, \\ \hat{\tilde{v}}_{bi}^T, & \text{if } i \in \mathcal{M}, \\ 0, & \text{otherwise}, \end{cases}$$

in the order i from 1 to N, where $\hat{\tilde{v}}_{bi}^T$ denotes the b-th row and i-th column element of matrix $\hat{\tilde{\boldsymbol{V}}}^T$, and $\mathcal{W}_{2,N}^{(i)}$ is the i-th polarized channel from BSC(p_2). Then, for every b corresponding to the degraded channel state, the decoder declares

$$\pi\left(\hat{\mu}_{1:N}^{(b)}\right) = \left[\begin{array}{c|c|c|c|c} \hat{u}_{1:|\mathcal{G}|}^{(b)} & | & \hat{\tilde{\boldsymbol{V}}}_b^T & | & 0 \end{array} \right].$$

The whole decoding process for fading binary symmetric channel is illustrated in Figure 4.5.

4.6 Performance Evaluation

Here, we summarize the performance of the proposed polar coding scheme. Intuitively, by using BSC SC decoders corresponding to the superior channel state, the output from Phase I successfully recovers all information bits, because the size of information set is equal to the size of $\mathcal{G}$. Then, for decoding at Phase II, the input vector $\hat{\tilde{\boldsymbol{V}}}_k$ can be considered as a ϱ_2-fraction erased polar codeword, hence, BEC SC decoder can decode all information bits in $v_{1:|\mathcal{A}|}^{(k)}$ correctly for all $k \in \{1, \ldots, |\mathcal{M}|\}$, and recover the erased entries correctly as well. Finally, in Phase III of decoding, the bits in $\mathcal{M}$ have the correct frozen values, and by adopting BSC SC decoders corresponding to the degraded channel state, all the remaining information bits can be decoded correctly.

Therefore, as long as the designed rates of polar codes do not exceed the corresponding channel capacities, all information bits in our proposed polar coding scheme are reliably decodable. Hence, more formally, we have the following theorem.

Theorem 4.2. *The proposed polar coding scheme achieves any rate* $R < C_{\text{CSI-D}}$*, for sufficiently large* N *and* B*, and the decoding error probability scales as* $\max\{O(B2^{-N^{\beta}}), O(N2^{-B^{\beta}})\}$ *with* $\beta < 1/2$*. Moreover, the complexity of the*

encoding and decoding processes are both given by $O(NB\log(NB))$, where N is the block length and B is the number of blocks.

Proof. The achievable rate (corresponding to the transmission of information bits in $v_{1:|\mathcal{A}|}^{(k)}$ and $u_{1:|\mathcal{G}|}^{(b)}$) is given by

$$\begin{aligned} R &= \frac{1}{NB}\Big\{|\mathcal{M}|\cdot|\tilde{\mathcal{A}}| + B\cdot|\mathcal{G}|\Big\} \\ &= [H(p_2) - H(p_1)][\varrho_1 - \epsilon] + [1 - H(p_2) - \epsilon] \\ &= \varrho_1[1 - H(p_1)] + \varrho_2[1 - H(p_2)] - \delta(\epsilon), \\ &= C_{\text{CSI-D}} - \delta(\epsilon), \end{aligned}$$

where we have used (4.2)), (4.3), (4.5), and

$$\delta(\epsilon) \triangleq \epsilon[1 + H(p_2) - H(p_1)] \to 0, \text{ as } \epsilon \to 0.$$

The proof for error exponent is obtained by utilizing error bound from polar coding. In Phase I and III of decoding, the error probability of recovering $u_{1:|\mathcal{G}|}^{(b)}$ correctly for each $b \in \{1,\ldots,B\}$ is given by $P_{\text{e},1}^{(b)} = O(2^{-N^\beta})$. Similarly, in decoding Phase II, the error probability of recovering $v_{1:|\mathcal{A}|}^{(k)}$ correctly for each $k \in \{1,\ldots,|\mathcal{M}|\}$ is given by $P_{\text{e},2}^{(k)} = O(2^{-B^\beta})$. Hence, by union bound, the total decoding error probability is upper bounded by

$$P_{\text{e}} \le \sum_{b=1}^{B} P_{\text{e},1}^{(b)} + \sum_{k=1}^{|\mathcal{M}|} P_{\text{e},2}^{(k)} = O(B2^{-N^\beta}) + O(N2^{-B^\beta}),$$

as N and B tend to infinity. Therefore, P_{e} vanishes if $B = o(2^{N^\beta})$ and $N = o(2^{B^\beta})$.

Finally, since we have $|\mathcal{M}|$ number of B-length polar codes as well as B number of N-length polar codes utilized, the overall complexity of the coding scheme for both encoding and decoding is given by

$$|\mathcal{M}| \cdot O(B \log B) + B \cdot O(N \log N) = O(NB \log(NB)).$$

□

This theorem shows that our proposed polar coding scheme achieves the capacity of fading BSC with low encoding and decoding complexity. In addition, the error scaling performance, which is inherited from polar codes, implies that long coherence intervals as well as large number of blocks are required for this coding scheme to make the error probability arbitrarily small.

4.7 Generalization to Arbitrary Finite Number of States

Here, we generalize the polar coding scheme to fading binary symmetric channel with arbitrary finite number of states. Consider S number of BSCs, each with a different transition probability. Without loss of generality, consider $\mathcal{W}_1 \triangleq \text{BSC}(p_1), \ldots, \mathcal{W}_S \triangleq \text{BSC}(p_S)$, with $p_1 \leq p_2 \leq \cdots \leq p_S$. Then, a fading BSC with S fading states is modeled as the channel being $\mathcal{W}_s$ with probability ϱ_s for a given fading block, where $\sum_{s=1}^{S} \varrho_s = 1$. The polarization of a fading BSC with S fading states is illustrated in Figure 4.6, where the reconstructed channel indices are divided into $S+1$ sets after permutation π. In addition to $\mathcal{G}$ and $\mathcal{B}$, there exist $S-1$ middle sets $\mathcal{M}_1, \ldots, \mathcal{M}_{S-1}$ in this case. For each

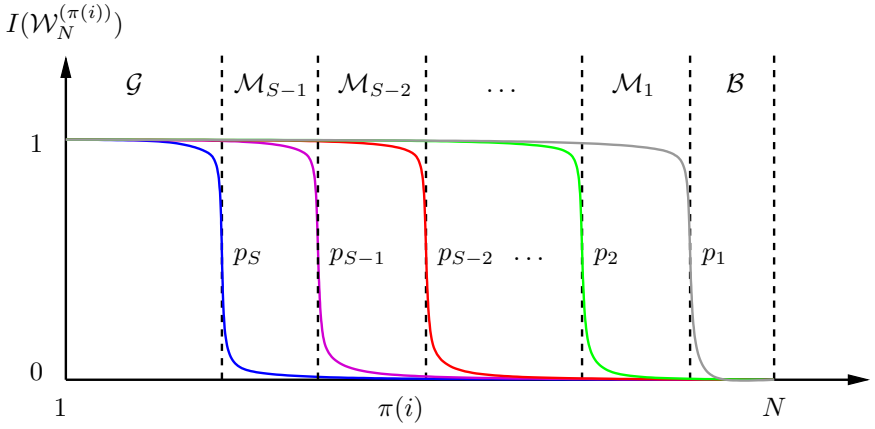

Figure 4.6: **Illustration of polarization for a fading binary symmetric channel with S channel states.** Besides $\mathcal{G}$ and $\mathcal{B}$, there are $S-1$ middle sets, denoted as $\mathcal{M}_1, \ldots, \mathcal{M}_{S-1}$.

channel index in set $\mathcal{M}_s$, channels having statistics being one of $\mathcal{W}_1, \ldots, \mathcal{W}_s$ are polarized to be noiseless and the remaining ones are purely noisy. Therefore, for channel indices belonging to $\mathcal{M}_s$, we consider modeling them as BECs with erasure probability given by

$$e_s \triangleq \sum_{t=s+1}^{S} \varrho_t, \quad 1 \leq s \leq S-1.$$

Based on this, we have

$$\begin{aligned}
&|\mathcal{G}| = |\mathcal{A}_S| = N \cdot [1 - H(p_S) - \epsilon], \\
&|\mathcal{M}_s| = N \cdot [H(p_{s+1}) - H(p_s)], \quad 1 \leq s \leq S-1, \\
&|\mathcal{B}| = N - |\mathcal{A}_1| = N \cdot [H(p_1) + \epsilon].
\end{aligned}$$

Here, the polarization result is similar to the case of two fading states, and we utilize a similar hierarchical coding scheme. In Phase I of encoding,

transmitter generates $S-1$ sets of polar codes, where each one is a G_B-coset code with parameter $(B, |\mathcal{A}_{\mathrm{e},s}|, \mathcal{A}_{\mathrm{e},s}, 0)$ with respect to BEC(e_s) (where $\mathcal{A}_{\mathrm{e},s}$ is the information set for channel BEC(e_s)), and all the encoded codewords are embedded into messages of Phase II in order. Then, in Phase II of encoding, we use BSC polar encoders with information set $\mathcal{G}$ to generate the final codeword with length NB. At the receiver end, we need $2S-1$ number of phases. Phase I utilizes the BSC(p_1) SC decoders to decode blocks with respective to the best channel state (state 1 in this case). Consider all decoded bits in $\mathcal{M}_1$, as well as adding erasures to undecoded blocks, we could decode all erased bits by using BEC(e_1) SC decoders in Phase II. Then, using the decoded information as frozen values for blocks with respective to state 2, BSC(p_2) SC decoders are adopted in Phase III to decode information bits in the blocks corresponding to channel state 2. Recursively, all information bits for both BSC encoding and BEC encoding could be reliably decoded, as long as the designed rates of polar codes do not exceed corresponding channel capacities. Hence, by adopting this hierarchical polar coding scheme, the achievable rate is given by

$$\begin{aligned} R &= \frac{1}{NB}\left\{B\cdot|\mathcal{G}| + \sum_{s=1}^{S-1}|\mathcal{M}_s|\cdot|\mathcal{A}_s|\right\} \\ &= [1-H(p_S)-\epsilon] + \sum_{s=1}^{S-1}[H(p_{s+1})-H(p_s)]\cdot(1-e_s-\epsilon) \\ &= \sum_{s=1}^{S}\varrho_s[1-H(p_s)] - \delta'(\epsilon), \end{aligned}$$

where $\delta'(\epsilon) \triangleq \epsilon[1+H(p_S)-H(p_1)]$ tends to 0 as $\epsilon \to 0$. Thus, to this end, the proposed polar coding scheme achieves the capacity of channel, and the

encoding and decoding complexities are both given by

$$\sum_{s=1}^{S-1} |\mathcal{M}_s| \cdot O(B \log B) + B \cdot O(N \log N) = O(NB \log(NB)),$$

which is independent to the value of S as $\sum_{s=1}^{S-1} |\mathcal{M}_s| \leq N$. For the same reason, the decoding error bound also remains the same as the case of only two fading states. Thus, our proposed polar coding scheme achieves the capacity of fading binary symmetric channel with arbitrary finite number of fading states, and the encoding and decoding complexity are both guaranteed to be tractable in practice.

4.8 Summary

In this section, a hierarchical polar coding scheme is proposed for the fading BSC. This novel scheme, by exploiting an erasure decoding approach at the receiver, utilizes the polarization results of different BSCs. (These BSCs are defined over channel uses at a given fading block and over fading blocks at a given channel use index.) This novel polar coding technique is shown to be capacity achieving for fading BSC. Remarkably, the proposed scheme does not assume channel state information at the transmitter and fading BSC models the fading additive white Gaussian noise (AWGN) channel with a BPSK modulation. Therefore, our results are quite relevant to the practical channel models considered in wireless communications.

We remark that the advantages of polar codes in rate and complexity are both inherited in the proposed coding schemes. More precisely, as polar

codes achieve channel capacity of BSC and BEC, our hierarchical utilization of polar codes also achieves the capacity of fading BSC. Meanwhile, the property of low complexity for polar codes is also inherited to the coding scheme for fading channels.

Finally, we note that the proposed coding scheme requires long codeword lengths to make the error probability arbitrarily small. This requirement translates to requiring long coherence intervals and large number of fading blocks as our approach utilizes coding over both channel uses and fading blocks. (This is somewhat similar to the analyses in Shannon theory, where the guarantee of the coding is that the error probability vanishes as the block length gets large.) Therefore, our coding scheme fits to the fading channels with moderate/long coherence time and large number of fading blocks. Here, we comment on applicability of the proposed coding scheme in typical wireless systems. As reported in [52][53], LTE systems operating at 1.8GHz frequency with 20MHz bandwidth typically have fading durations of 2.8×10^5 to 1.0×10^7 channel uses. In addition, WiFi systems operating at 5GHz frequency with 20MHz bandwidth typically have fading durations of 7.7×10^5 to 1.8×10^7 channel uses [54]. (Here, a mobile speed of 1m/s is assumed for both systems.) Polar codes, on the other hand, typically have error rates around 10^{-6} when the blocklength is around 2^{10}, and a smaller error probability is even possible, when the decoding is implemented with a better decoder. For instance, instead of the classical SC decoder, a list decoder [48] can be utilized. Finally, besides long coherence intervals, another requirement for the proposed coding scheme

is to have large number of fading blocks. This requirement can be satisfied in many practical scenarios at the expense of having large decoding delays.

Chapter 5

Polar Coding for Fading AEN Channels

5.1 System Model of Fading AEN Channels

In this section, we proceed to consider another fading channel model with analog noise statistics known. In particular, the hierarchical polar coding scheme is combined with the aforementioned expansion coding technique (Chapter 2) to achieve ergodic capacity for fading channels with analog noises. More precisely, we consider the fading additive exponential noise (AEN) channel given by

$$\mathsf{Y}_{b,i} = \mathsf{X}_{b,i} + \mathsf{Z}_{b,i}, \quad b = 1, \ldots, B, \quad i = 1, \ldots, N,$$

where $\mathsf{X}_{b,i}$ is channel input and restricted to be positive and with mean E_{X}; N is block length; and B is the number of blocks. In this model, $\mathsf{Z}_{b,i}$ are assumed to be identically distributed within a block and follow an ergodic i.i.d. fading process over blocks. That is, if we consider a fading AEN channel with S states, then, with probability ϱ_s channel noise $\mathsf{Z}_{b,i}$ is distributed as an exponential random variable with parameter E_{Z_s} for a given b and all $i \in \{1, \ldots, N\}$, i.e.,

$$f_{\mathsf{Z}_{b,i}}(z) = \frac{1}{E_{\mathsf{Z}_s}} e^{-\frac{z}{E_{\mathsf{Z}_s}}}, \quad z \geq 0, \tag{5.1}$$

where $1 \leq s \leq S$ and $\sum_{s=1}^{S} \varrho_s = 1$.

We first state the following upper bound on the ergodic channel capacity in the high SNR regime.

Lemma 5.1. *The ergodic capacity of a fading AEN channel, with channel state information known at the decoder, is upper bounded as follows.*

$$\lim_{E_{\mathsf{X}} \to \infty} C_{\textit{CSI-D}} \leq \sum_{s=1}^{S} \varrho_s \left[\log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) \right] \tag{5.2}$$

Proof. See Appendix 5.A. □

In the following, we show that our proposed polar coding scheme achieves the upper bound above in the high SNR regime.

Remark 5.2. *Note that the capacity of the fading AEN channel with CSI-D approaches to the bound above in the high SNR regime. (For example, our coding scheme, as shown below, provides one such achievable rate.) This observation is similar to the Gaussian counterpart [49], where in the high SNR regime, the performance obtained from a waterfilling strategy - the optimal solution for the case where encoder can adapt its power based on the channel state, i.e., CSI-ED - approaches to the performance of utilizing the same power allocation for each fading channel.*

Remark 5.3. *The model above assumes a mean constraint on the channel input where the average is over channel blocks and channel states. If the mean constraint is per block (abbreviated as MPB - Mean Per fading Block - in the*

following), i.e., $E[\mathsf{X}_{b,i}] \leq E_{\mathsf{X}}$ for each fading block b, then by following steps similar to the ones above, we have

$$C_{\textit{CSI-D, MPB}} \leq C_{\textit{CSI-ED, MPB}} = \sum_{s=1}^{S} \varrho_s \left[\log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) \right].$$

5.2 Expansion Coding with Hierarchical Polar Coding

Similar to the expansion coding technique utilized to achieve the capacity of static AEN channel in Chapter 2, the binary expansion of channel noise is considered as

$$\hat{\mathsf{Z}}_{b,i} \triangleq \sum_{l=-L_1}^{L_2} 2^l \mathsf{Z}_{b,i,l}, \tag{5.3}$$

where $\mathsf{Z}_{b,i,l}$ is a discrete random variable taking value in $\{0, 1\}$. However, the distribution of $\mathsf{Z}_{b,i,l}$ depends on the fading state. More precisely, if the noise for a fading block b is exponential with parameter E_{Z_s}, then $\mathsf{Z}_{b,i,l}$ is a Bernoulli random variable with parameter

$$q_{l,s} \triangleq \Pr\{\mathsf{Z}_{b,i,l} = 1\} = \frac{1}{1 + e^{2^l / E_{\mathsf{Z}_s}}}. \tag{5.4}$$

Then, by the decomposability of exponential random variable, $\hat{\mathsf{Z}}_{b,i} \xrightarrow{d.} \mathsf{Z}_{b,i}$ as L_1 and L_2 tend to infinity. In this sense, we approximate the original exponential noise perfectly by a set of discrete noises.

Similarly, we also expand channel input and output as follows,

$$\hat{\mathsf{Y}}_{b,i} \triangleq \sum_{l=-L_1}^{L_2} 2^l \mathsf{Y}_{b,i,l} = \sum_{l=-L_1}^{L_2} 2^l (\mathsf{X}_{b,i,l} + \mathsf{Z}_{b,i,l}),$$

where $\mathsf{X}_{b,i,l}$ is a Bernoulli random variable with parameter $p_l \triangleq \Pr\{X_{b,i,l} = 1\}$. At this point, we model the expanded channels as

$$\mathsf{Y}_{b,i,l} = \mathsf{X}_{b,i,l} + \mathsf{Z}_{b,i,l}, \quad l = -L_1, \ldots, L_2.$$

Note that the summation is a real sum here, and hence, the channel is not a fading BSC for a given block. If we replace the real sum by modulo-2 sum, then, at level l, any capacity achieving code for fading BSC, for example the one constructed in Chapter 4, can be utilized in combination with the method of Gallager [37] [51] to achieve a rate corresponding to the one obtained by the mutual information $I(\mathsf{X}_{b,l}; \mathsf{Y}_{b,l})$ evaluated with a desired input distribution on $\mathsf{X}_{b,l}$.

Then, using the technique to essentially remove carries as discussed in Chapter 2, each level could be modeled as a fading BSC. Thus, expansion coding reduces the problem of coding over a fading exponential noise channel into a set of simpler subproblems, coding over fading BSCs. By adopting capacity achieving polar coding scheme proposed in Chapter 4 for each expanded fading BSC, we have the following achievable rate result for these channels.

Theorem 5.4. *By decoding carries in expansion coding, and adopting hierarchical polar coding scheme for fading BSC in each expanded level, the proposed scheme achieves the rate given by*

$$R = \sum_{l=-L_1}^{L_2} \sum_{s=1}^{S} \varrho_s [H(p_l \otimes q_{l,s}) - H(q_{l,s})], \tag{5.5}$$

for any $L_1, L_2 > 0$*, where* $p_l \in [0, 0.5]$ *is chosen to satisfy*

$$\sum_{l=-L_1}^{L_2} 2^l p_l \leq E_{\mathsf{X}}.$$

We note the followings. First, the achievable scheme we utilize satisfies the mean constraint on the channel input for each block, i.e., averaged over channel uses, $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathsf{X}_{i,b} \leq E_{\mathsf{X}}$ for each block b. (This implies satisfying power constraint averaged over the blocks as well.) Secondly, the maximum rate from our coding scheme could be considered as an optimization problem over finite number of parameters p_l, $-L_1 \leq l \leq L_2$. However, it is not clear how to solve this non-convex problem. Here, instead of searching for an optimal solution, we shift our focus to finding a sub-optimal choice of p_l such that the achievable rate is close the optimal one in the high SNR regime. From the proof of Lemma 5.1, we observe that the optimal input distribution for the case with the CSI at the transmitter could be approximated with an exponential with parameter E_{X_s} as SNR $= E_{\mathsf{X}_s}/E_{\mathsf{Z}_s}$ gets large. As we do not have CSI at the transmitter in our model, we consider choosing the same energy level, E_{X}, for each fading block. Noting again that the optimal input distribution is unknown for our fading model, the high SNR observation inspires us to choose p_l as

$$p_l = \frac{1}{1 + e^{2^l/E_{\mathsf{X}}}}. \tag{5.6}$$

The following theorem gives the main result of our polar coding scheme over fading AEN channel.

Theorem 5.5. *For any positive constant $\epsilon < 1$, if*

- $L_1 \geq -\log \epsilon - \min_s \log E_{\mathsf{Z}_s}$;
- $L_2 \geq -\log \epsilon + \log E_{\mathsf{X}}$;
- $\min_s SNR_s \geq 1/\epsilon$, *where* $SNR_s \triangleq E_{\mathsf{X}}/E_{\mathsf{Z}_s}$,

then by decoding carries and adopting hierarchical polar codes at each fading BSC after expansion, the achievable rate R given by (5.5)*, with a choice of p_l as* (5.6)*, satisfies*

$$R \geq \sum_{s=1}^{S} \varrho_s \left[\log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) \right] - 5 \log e \cdot \epsilon.$$

Proof. The spirit of the proof is analogous to the one of static AEN channel case, except for taking into the impact of fading. More precisely, in order to achieve the capacity of fading AEN channel, first, SNR should be large enough, and secondly, the number of expanded levels should also be large enough such that the highest level exceeds all the left shifted levels of expanded signal, and the lowest level exceeds the right shifted levels of expanded noises. Hence, in total, basically we need $\log \mathrm{SNR}_{\max}$ ($\mathrm{SNR}_{\max} \triangleq \max_s E_{\mathsf{X}}/E_{\mathsf{Z}_s}$) number of levels to cover all "non-trivial" levels for coding, as well as extra $-2\log\epsilon$ number of levels to shoot for accuracy. The details of the proof are illustrated in Appendix 5.B. □

Remark 5.6. *We note that the proposed scheme achieves a rate*

$$\sum_{s=1}^{S} \varrho_s \left[\log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) \right] = C_{\textit{CSI-ED, MPB}},$$

which is an upper bound on the capacity for the CSI-D case in the high SNR regime. (See Lemma 5.1.) Therefore, the proposed scheme achieves the capacity in the high SNR regime.

5.3 Numerical Results

Numerical results for achievable rate given by (5.5) with p_l chosen as (5.6) are illustrated in Figure 5.1, where we consider the case of two fading states. It is evident from the figure, and also from the theoretical analysis given in Theorem 5.5, that our proposed polar coding scheme together with expansion coding achieves the upper bound on the channel capacity (Lemma 5.1) in the high SNR regime. Therefore, the proposed coding scheme achieves the channel capacity for sufficiently large SNR.

We also note that, similar to static AEN channel case, the coding scheme does not perform well in the low SNR regime, which mainly results from two reasons. First, the upper bound we derived in Lemma 5.1, which is the target rate in our coding scheme, is not tight in the low SNR regime. Secondly, our choice of p_l only behaves as a good approximation for sufficiently high SNR, which limits the proposed scheme to be effective at the corresponding regime. However, as evident from the numerical results, for a fairly large set of SNR values the proposed scheme is quite effective. In addition, the upper bound curve is equal to $C_{\text{CSI-ED, MPB}}$, the capacity when the input mean constraint is imposed per block (instead of averaging over the blocks). Therefore, for the scenario of having input constraint per each fading block, the upper

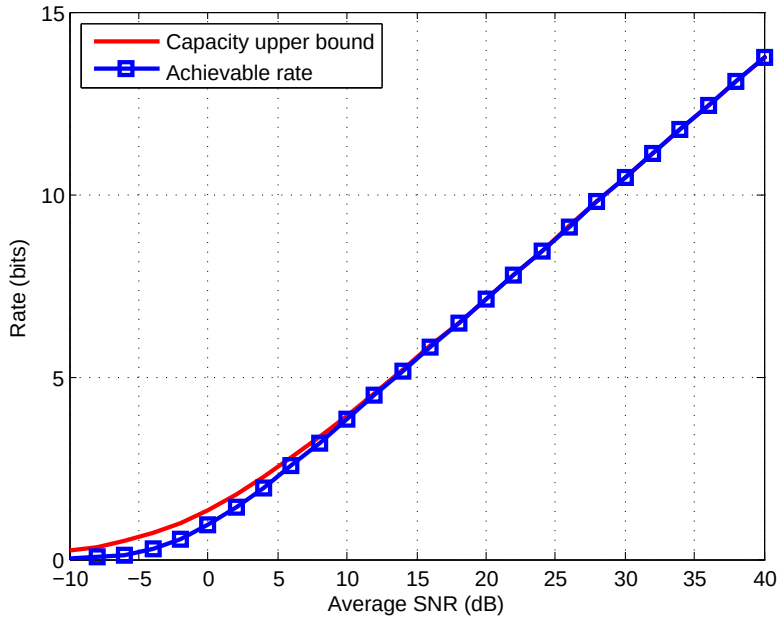

Figure 5.1: **Numerical results.** The upper bound of ergodic capacity, $C_{\text{CSI-ED, MPB}}$, which is equal to $C_{\text{CSI-ED}}$ for sufficiently large SNR, is given by the red curve. The achievable rate is given by the blue curve. In this analysis, only two fading states are concerned, and the parameters are chosen as $E_{\mathsf{Z}_1} = 0.5$, $E_{\mathsf{Z}_2} = 3$, $\varrho_1 = 0.8$, and $\varrho_2 = 0.2$. Average SNR is defined as $E_{\mathsf{X}}/(\sum_{s=1}^{S} \varrho_s E_{\mathsf{Z}_s})$.

bound $C_{\text{CSI-D, MPB}} \leq C_{\text{CSI-ED, MPB}}$ holds at any SNR, and the only degradation in our coding scheme is due to the second point discussed above.

5.4 Summary

In this section, we illustrate the utilization of hierarchical polar coding scheme for encoding over another fading channel model. For the fading AEN channel model, expansion coding is adopted to convert the problem of cod-

ing over an analog fading channel into coding over discrete fading channels. By performing this expansion approach and making the resulting channels independent (via decoding the underlying carries), a fading AEN channel is decomposed into multiple independent fading BSCs (with a reliable decoding of the carries). By utilizing the hierarchical polar coding scheme for fading BSC, both theoretical proof and numerical results show that the proposed approach achieves the capacity of this fading channel in the high SNR regime.

5.A Proof of Lemma 5.1

Denote the channel state as a random variable S, which is discrete on set $\{1, 2, \ldots, S\}$. If the channel state information is known at the decoder, then we have

$$\begin{aligned}
\lim_{E_{\mathsf{X}}\to\infty} C_{\text{CSI-D}} &\overset{(a)}{\leq} \lim_{E_{\mathsf{X}}\to\infty} C_{\text{CSI-ED}} \\
&\overset{(b)}{=} \lim_{E_{\mathsf{X}}\to\infty} \max_{\mathbb{E}[\mathsf{X}]\leq E_{\mathsf{X}}} I(\mathsf{X};\mathsf{Y}|\mathsf{S}) \\
&= \lim_{E_{\mathsf{X}}\to\infty} \max_{\mathbb{E}[\mathsf{X}]\leq E_{\mathsf{X}}} h(\mathsf{Y}|\mathsf{S}) - h(\mathsf{Y}|\mathsf{S},\mathsf{X}) \\
&= \lim_{E_{\mathsf{X}}\to\infty} \max_{\mathsf{X}_s:\sum\limits_s q_s\mathbb{E}[\mathsf{X}_s]\leq E_{\mathsf{X}}} \sum_{s=1}^{S} q_s[h(\mathsf{X}_s+\mathsf{Z}_s) - h(\mathsf{Z}_s)] \\
&\overset{(c)}{=} \lim_{E_{\mathsf{X}}\to\infty} \max_{E_{\mathsf{X}_s}:\sum\limits_s q_s E_{\mathsf{X}_s}\leq E_{\mathsf{X}}} \sum_{s=1}^{S} q_s \left[\log\left(1+\frac{E_{\mathsf{X}_s}}{E_{\mathsf{Z}_s}}\right)\right] \\
&\overset{(d)}{=} \sum_{s=1}^{S} q_s \left[\log\left(1+\frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}}\right)\right],
\end{aligned}$$

where

(a) is due to upper bounding the channel capacity with the case where encoder also has CSI and adapts its coding according to the channel states;

(b) is the ergodic capacity of the channel where both encoder and decoder has CSI (see, e.g., [49, pages 203-209]);

(c) holds as exponential distribution maximizes the differential entropy on positive support with a mean constraint [34] [17, page 412]. Here, we choose X_s to be a weighted sum of an exponential distribution with mean $E_{\mathsf{X}_s} + E_{\mathsf{Z}_s}$

and a delta function in order to make the output $\mathsf{X}_s + \mathsf{Z}_s$ to be exponentially distributed random variable. That is, the pdf of X_s is given by

$$f_{\mathsf{X}_s}(x) = \frac{E_{\mathsf{X}_s}}{E_{\mathsf{X}_s} + E_{\mathsf{Z}_s}} \frac{e^{-x/(E_{\mathsf{X}_s}+E_{\mathsf{Z}_s})}}{E_{\mathsf{X}_s} + E_{\mathsf{Z}_s}} u(x) + \frac{E_{\mathsf{Z}_s}}{E_{\mathsf{X}_s} + E_{\mathsf{Z}_s}} \delta(x),$$

where $\delta(x) = 1$ if $x = 0$, and $\delta(x) = 0$ if $x \neq 0$; $u(x) = 1$ if $x \geq 0$, and $u(x) = 0$ if $x < 0$.

(d) follows by taking the limit.

5.B Proof of Theorem 5.5

We first state bounds for the entropy of channel noise with mean E_{Z_s} at level l, which are obtained from the Lemma 2.6:

$$H(q_{l,s}) \leq 3 \log e \cdot 2^{-l+\eta_s} \quad \text{for } l > \eta_s, \tag{5.7}$$

$$H(q_{l,s}) \geq 1 - \log e \cdot 2^{l-\eta_s} \quad \text{for } l \leq \eta_s, \tag{5.8}$$

where $\eta_s \triangleq \log E_{\mathsf{Z}_s}$.

Now, if we denote $\xi \triangleq \log E_{\mathsf{X}}$, then comparing the definitions of p_l and $q_{l,s}$, we get

$$p_l = \frac{1}{1 + e^{2^l/E_{\mathsf{X}}}} = q_{l+\eta_s-\xi,s}. \tag{5.9}$$

Based on these observations, we have

$$\begin{aligned} &\sum_{l=-L_1}^{L_2} [H(p_l \otimes q_{l,s}) - H(q_{l,s})] \\ &\quad \overset{(a)}{\geq} \sum_{l=-L_1}^{L_2} [H(p_l) - H(q_{l,s})] \end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} \sum_{l=-L_1}^{L_2} [H(q_{l+\eta_s-\xi,s}) - H(q_{l,s})] \\
&= \sum_{l=-L_1+\eta_s-\xi}^{L_2+\eta_s-\xi} H(q_{l,s}) - \sum_{l=-L_1}^{L_2} H(q_{l,s}) \\
&= \sum_{l=-L_1+\eta_s-\xi}^{-L_1-1} H(q_{l,s}) - \sum_{l=L_2+\eta_s-\xi+1}^{L_2} H(q_{l,s}) \\
&\overset{(c)}{\geq} \sum_{l=-L_1+\eta_s-\xi}^{-L_1-1} \left[1 - \log e \cdot 2^{l-\eta_s}\right] - \sum_{l=L_2+\eta_s-\xi+1}^{L_2} 3 \log e \cdot 2^{-l+\eta_s} \\
&\overset{(d)}{\geq} \xi - \eta_s - \log e \cdot 2^{-L_1-\eta_s} - 3 \log e \cdot 2^{-L_2+\xi} \\
&\overset{(e)}{\geq} \log \left(\frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) - \log e \cdot \epsilon - 3 \log e \cdot \epsilon \\
&\overset{(f)}{\geq} \log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) - \log e \cdot \frac{E_{\mathsf{Z}_s}}{E_{\mathsf{X}}} - \log e \cdot \epsilon - 3 \log e \cdot \epsilon \\
&\overset{(g)}{\geq} \log \left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}} \right) - 5 \log e \cdot \epsilon, \qquad (5.10)
\end{aligned}
$$

where

(a) is due to $p_l \otimes q_{l,s} \triangleq p_l(1-q_{l,s}) + q_{l,s}(1-p_l) \geq p_l$, and then due to the fact that entropy function is increasing on $[0, 0.5]$ (and, we have $p_l \otimes q_{l,s} \leq 0.5$);

(b) follows from equation (5.9);

(c) follows from bounds (5.7) and (5.8);

(d) follows as

$$\sum_{l=-L_1+\eta_s-\xi}^{-L_1-1} 2^{l-\eta_s} \leq 2^{-L_1-\eta_s},$$

and

$$\sum_{l=L_2+\eta_s-\xi+1}^{L_2} 2^{-l+\eta_s} = \sum_{l=-L_2+\eta_s}^{-L_2+\xi-1} 2^{l} \leq 2^{-L_2+\xi};$$

(e) follows from theorem assumptions that $L_1 \geq -\log\epsilon - \min_s \eta_s$, and $L_2 \geq -\log\epsilon + \xi$;

(f) is due to the fact that $\log(1+E_{\mathsf{X}}/E_{\mathsf{Z}_s}) - \log(E_{\mathsf{X}}/E_{\mathsf{Z}_s}) = \log(1+E_{\mathsf{Z}_s}/E_{\mathsf{X}}) \leq \log e \cdot E_{\mathsf{Z}_s}/E_{\mathsf{X}}$ (as $\log(1+\alpha) \leq \log e \cdot \alpha$ for any $\alpha \geq 0$);

(g) is due to the assumption in theorem that $\min_s \mathrm{SNR}_s \geq 1/\epsilon$.

Then, using (5.10) in (5.5) of Theorem 5.4, we have

$$\begin{aligned} R &= \sum_{s=1}^{S} q_s \left\{ \sum_{l=-L_1}^{L_2} [H(p_l \otimes q_{l,s}) - H(q_{l,s})] \right\} \\ &\geq \sum_{s=1}^{S} q_s \left\{ \log\left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}}\right) - 5\log e \cdot \epsilon \right\} \\ &= \sum_{s=1}^{S} q_s \left[\log\left(1 + \frac{E_{\mathsf{X}}}{E_{\mathsf{Z}_s}}\right) \right] - 5\log e \cdot \epsilon. \end{aligned}$$

Chapter 6

Polar Coding for Fading Wiretap BSCs

6.1 Background of Polar Coding for Wiretap Channels

Wiretap channels, introduced in the seminal paper of Wyner [55], model the communication between a transmitter and a receiver in the presence of an eavesdropper that overhears the transmitted signals via the channel between transmitter and eavesdropper (e.g., by tapping the wire between the legitimate nodes). The task of transmitter is to hide information from the eavesdropper while communicating reliably to the receiver. Wyner studied this problem and characterized the capacity region for certain channel models including the case of degraded eavesdropper [55]. The achievability technique is the randomized version of the Shannon's random coding approach, where the randomization is utilized to confuse the eavesdropper, in order to achieve security. Since the publication of Wyner's work, several studies in the network information theory domain have utilized this random coding approach to characterize the corresponding secrecy capacities. Yet, the design of secrecy achieving coding schemes with practical constraints such as low complexity and availability of channel state information remain as an important direction in the physical layer security.

Recently, polar codes have been utilized for communication over degraded wiretap channels [56][57][58][59]. These schemes are based on the behavior of the polarization of degraded channels, where the polarized channels for the degraded wiretap channels can be partitioned to one of the following sets:

1) Good for both receiver and eavesdropper;

2) Good for receiver and bad for eavesdropper;

3) Bad for both receiver and eavesdropper.

The fraction of type 2 channels approach to the secrecy capacity for the degraded (binary symmetric) wiretap channels, and the communication scheme utilizes this type of polarized indices to transmit information; whereas, type 1 channels are assigned to random bits to limit the eavesdropper's ability to obtain information about the messages. (Type 3 channels are frozen, i.e., set to a constant value and shared to receiver.) This scheme allows for achieving the secrecy capacity, while inheriting the low complexity nature of polar codes. In other words, this technique mimics the Wyner's random coding approach with practical encoding/decoding schemes. The main hurdle for most practical applications though is to have the eavesdropper channel state information (CSI) at the transmitter, e.g., in order to differentiate between type 1 and 2 channels in this coding scheme. Remarkably, an incorrect knowledge about the eavesdropper CSI would leak information, hence will not result in a meaningful security guarantee. In this work, we focus on relaxing the assumption on

the instantaneous CSI knowledge, and develop polar coding schemes for fading wiretap channels where only the statistics of CSI is known at the transmitter.

Recent studies on the design of polar coding schemes to achieve secrecy include [60][61][62][63][64][65][66], where strong security is considered in [57][61][64][66], key agreement/generation is studied in [57][62][63], and other channel models (different than discrete memoryless wiretap channel) are considered in [65][67][58]. Our model is similar to the fading models considered in [57][67] but differentiates from all these prior studies in that only a statistical (i.e., distribution) CSI for *both* receiver and eavesdropper channels is assumed at the transmitter. Polar coding schemes for fading wiretap channels are first studied in [57], where the transmitter has the knowledge of instantaneous CSI for the receiver's channel and statistical CSI for the eavesdropper's channel. With this setup, a key agreement scheme is proposed based on utilizing polar codes for each fading block, where the communicated bits over fading blocks are then used in a privacy amplification step to construct secret keys. This technique when combined with invertible extractors allows for secure message transmission but with the requirement of receiver CSI at the transmitter[57]. Recent work [67] proposes a polar coding scheme that utilizes artificial noise and multiple transmit antennas under the same assumption (instantaneous CSI for receiver and statistical CSI for eavesdropper) for the fading channels. However, a guarantee of secrecy rate with some probability (not the corresponding channel capacity) is achieved. In contrast, in this paper, we consider a fading channel model where the transmitter does not need to know any in-

stantaneous CSI, but only its distribution for *both* receiver and eavesdropper channels. The hierarchical polar coding scheme proposed in this paper, to the best of our knowledge, is the first provably secrecy capacity achieving coding scheme for fading (binary symmetric) wiretap channels. Considering that this type of binary channels model the AWGN channels with BPSK modulation and demodulation, our scheme covers a wide application scenarios in practice.

6.2 System Model of Fading Wiretap BSCs

We investigate the case where main channel and eavesdropper fade simultaneously. More precisely, consider the fading (binary symmetric) wiretap channel model (Figure 6.1): Alice wishes to send message to Bob through the main channel $\mathcal{W}$, where the channel experiences the following block fading phenomenon: with probability ϱ_1, channel $\mathcal{W}$ behaves as BSC(p_1) (in the superior state), and with the rest probability $\varrho_2 \triangleq 1 - \varrho_1$, channel $\mathcal{W}$ behaves as BSC(p_2) (in the degraded state). On the same time, the transmission also reaches to an adversary (Eve) through the wiretap channel $\mathcal{W}^*$, where $\mathcal{W}^*$ is degraded compared to the main channel, and experiences the same fading state as the main channel. In particular, when $\mathcal{W}$ behaves as BSC(p_1), $\mathcal{W}^*$ behaves as BSC(p_1^*); when $\mathcal{W}$ behaves as BSC(p_2), $\mathcal{W}^*$ behaves as BSC(p_2^*). Under this system model, we have $p_1 \leq p_2 \leq 0.5$, $p_1^* \leq p_2^* \leq 0.5$, $p_1 \leq p_1^*$, and $p_2 \leq p_2^*$.

Remark 6.1. *Simultaneous fading model consider the case where main channel and eavesdropper channel experience the fading states, and eavesdropper*

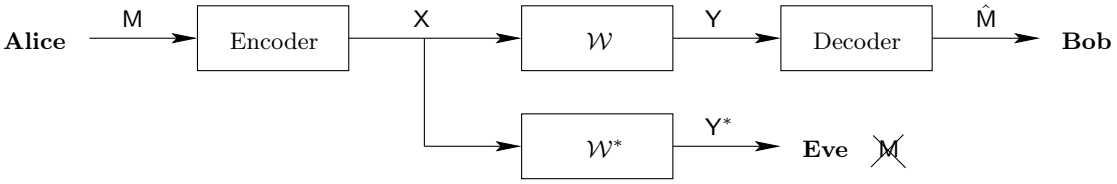

Figure 6.1: **System model for wiretap channels.** The target message M is demanded to be obtained by the main channel decoder, but not to be decodable at the eavesdropper.

channel is assumed to be degraded to the main channel over each fading block. The independent fading model is briefly discussed at the end of this chapter, where in some fading blocks, the eavesdropper can be stronger than the main channel (although in average sense the eavesdropper channel is degraded).

In general, fading coefficients vary at a much slower pace than the transmission symbol duration. For such cases, block fading model is considered, where the channel state is assumed to be constant within each coherence time interval, and follows a stationary ergodic process across fading blocks [49]. To this end, we consider the practical scenario where channel state information (CSI) is available only at the decoder (CSI-D), while the encoder only knows the statistics of channel states. Under this model, a secrete message M is encoded by an encoding function $\psi(\cdot)$ to generate transmitted symbols: $\mathsf{X}_{1:NB} = \psi(\mathsf{M})$, where N is the length of a fading block, and B is the number of blocks. At the receiver, a decoding function $\varphi(\cdot)$ gives an estimate of the estimate $\hat{\mathsf{M}}$, i.e., $\hat{\mathsf{M}} = \varphi(\mathsf{Y}_{1:NB})$. The reliability of transmission is satisfied if

$$P_{\mathrm{e}} \triangleq \Pr\{\mathsf{M} \neq \hat{\mathsf{M}} | \mathsf{Y}_{1:NB}, \mathsf{S}\} \to 0, \text{ as } N, B \to \infty \tag{6.1}$$

where S denotes CSI, and (weak) security is defined as achieving

$$\frac{1}{NB} I(\mathsf{M}; \mathsf{Y}^*_{1:NB} | \mathsf{S}) \to 0, \text{ as } N, B \to \infty. \tag{6.2}$$

Under the degraded assumption, the secrecy capacity of the wiretap system can be upper bounded by

$$\begin{aligned} SC_{\text{CSI-D}} &\overset{(a)}{\leq} SC_{\text{CSI-ED}} \\ &\overset{(b)}{=} \max_{p(x|s)} [I(\mathsf{X}; \mathsf{Y} | \mathsf{S}) - I(\mathsf{X}; \mathsf{Y}^* | \mathsf{S})] \\ &= \max_{p(x|1)} \varrho_1 [I(\mathsf{X}; \mathsf{Y} | \mathsf{S} = 1) - I(\mathsf{X}; \mathsf{Y}^* | \mathsf{S} = 1)] \\ &\qquad + \max_{p(x|2)} \varrho_2 [I(\mathsf{X}; \mathsf{Y} | \mathsf{S} = 2) - I(\mathsf{X}; \mathsf{Y}^* | \mathsf{S} = 2)] \\ &\overset{(c)}{=} \varrho_1 [H(p_1^*) - H(p_1)] + \varrho_2 [H(p_2^*) - H(p_2)], \end{aligned} \tag{6.3}$$

where

(a) follows by upper bounding the secrecy capacity with the case where encoder has CSI (and adapts its coding scheme according to the channel states);

(b) is due to the secrecy capacity of the degraded wiretap channel;

(c) is due to the secrecy capacity result for the degraded binary symmetric wiretap channel [50].

In this paper, assuming CSI is available only at the receivers, we provide a polar coding scheme that achieves this upper bound while satisfying reliability (6.1) and security (6.2) constraints. To this end, the upper bound

(6.3) gives the secrecy capacity of our model. For the moment, we assume $p_1 \leq p_2 \leq p_1^* \leq p_2^*$, and the remaining case ($p_1 \leq p_1^* \leq p_2 \leq p_2^*$) is detailed later in Section 6.7.

6.3 Hierarchical Polar Encoder

In this section, we combine the hierarchical polar coding scheme introduced in Chapter 4 with the polar coding scheme for wiretap channels [56][57][58][59].

The encoder works in two phases (see Figure 6.2), hierarchically using polar codes to generate an NB-length codeword.

6.3.1 Phase I: BEC Encoding

Here, we consider two sets of messages to be encoded using polar encoders designed for binary erasure channels (BECs). For the first set of messages, we generate $|\mathcal{M}_1|$ number of BEC polar codes, where

$$|\mathcal{M}_1| = N \cdot [H(p_2^*) - H(p_1^*)]. \tag{6.4}$$

Consider a set of blockwise messages $u_{1:|\mathcal{A}^c|}^{(j)}$ with $j \in \{1, \ldots, |\mathcal{M}_1|\}$, where $\mathcal{A}$ is the information set for BEC(ϱ_2), i.e.,

$$|\mathcal{A}| = B \cdot [\varrho_1 - \epsilon], \tag{6.5}$$

$$|\mathcal{A}^c| = B \cdot [\varrho_2 + \epsilon], \tag{6.6}$$

where ϵ is a positive number tending to 0 as N and B tend to infinity. For every $u_{1:|\mathcal{A}^c|}^{(j)}$, we combine it with $|\mathcal{A}|$ random bits to construct polar codeword $\tilde{u}_{1:B}^{(j)}$.

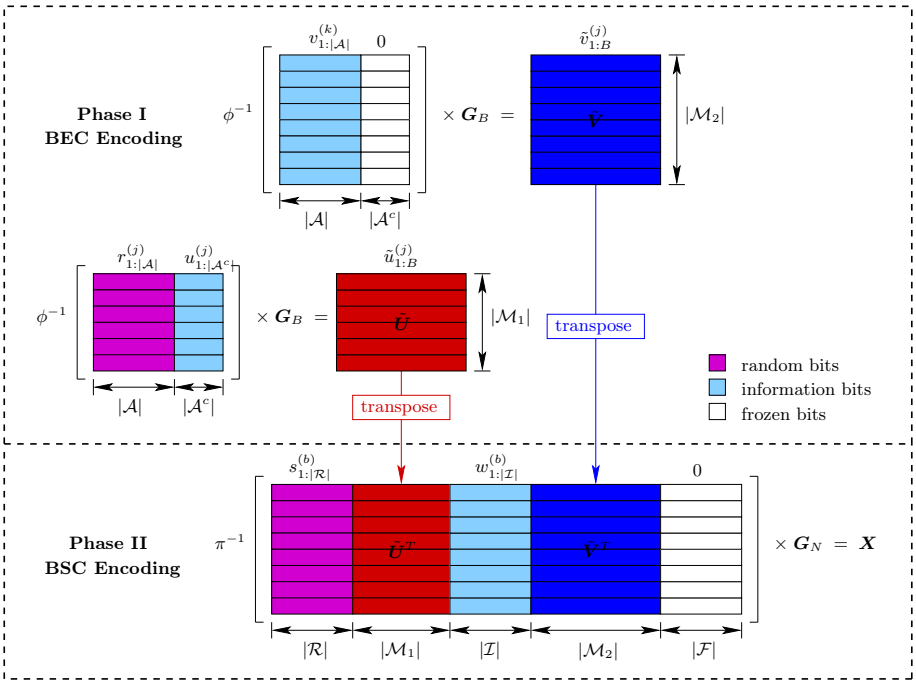

Figure 6.2: **Encoder of the polar coding scheme for wiretap channels.** The Encoder works in two phases, successively utilizing BEC and BSC polar encoders. The codewords encoded from Phase I are transposed and embedded into the massage of Phase II.

Denoting the permutation for BEC channel as ϕ, and the uniform random string as $r_{1:|\mathcal{A}|}^{(j)}$ (each bit is Ber(1/2) distributed), the encoding process is given by

$$\tilde{u}_{1:B}^{(j)} = \mu_{1:B}^{(j)} \times \boldsymbol{G}_B,$$
$$\phi\left(\mu_{1:B}^{(j)}\right) = \left[\begin{array}{c|c} r_{1:|\mathcal{A}|}^{(j)} & u_{1:|\mathcal{A}^c|}^{(j)} \end{array}\right],$$

for every $j \in \{1, \ldots, |\mathcal{M}_1|\}$, where $\boldsymbol{G}_B$ is the polar generator matrix with size B. By collecting all $\tilde{u}_{1:B}^{(j)}$ together, the encoder generates a $|\mathcal{M}_1| \times B$ matrix $\tilde{\boldsymbol{U}}$. We denote $\tilde{\boldsymbol{U}}_b^T$ as the b-th row of the transpose of $\tilde{\boldsymbol{U}}$, where $b \in \{1, \ldots, B\}$.

Secondly, we generate $|\mathcal{M}_2|$ number of BEC polar codes, where

$$|\mathcal{M}_2| = N \cdot [H(p_2) - H(p_1)]. \tag{6.7}$$

Consider another set of blockwise messages $v_{1:|\mathcal{A}|}^{(k)}$ with $k \in \{1, \ldots, |\mathcal{M}_2|\}$. Each message is set as information bits to construct polar codeword $\tilde{v}_{1:B}^{(k)}$. More formally, this encoding process is given by

$$\begin{aligned} \tilde{v}_{1:B}^{(k)} &= \nu_{1:B}^{(k)} \times \boldsymbol{G}_B, \\ \phi\left(\nu_{1:B}^{(k)}\right) &= \left[\begin{array}{c|c} v_{1:|\mathcal{A}|}^{(k)} & 0 \end{array} \right], \end{aligned}$$

for every $k \in \{1, \ldots, |\mathcal{M}_2|\}$. The collection of all $\tilde{v}_{1:B}^{(k)}$ together is denoted as a $|\mathcal{M}_2| \times B$ matrix $\tilde{\boldsymbol{V}}$. We denote $\tilde{\boldsymbol{V}}_b^T$ as the b-th row of the transpose of $\tilde{\boldsymbol{V}}$, where $b \in \{1, \ldots, B\}$.

6.3.2 Phase II: BSC Encoding

In this phase, we generate B number of BSC polar codes, each with length N. The encoded codewords from previous phase are embedded as messages of this phase. We consider a set of messages $w_{1:|\mathcal{I}|}^{(b)}$ with $b \in \{1, \ldots, B\}$, where

$$|\mathcal{I}| = N \cdot [H(p_1^*) - H(p_2)]. \tag{6.8}$$

For every $w_{1:|\mathcal{I}|}^{(b)}$, we introduce random bits $s_{1:|\mathcal{R}|}^{(b)}$, where

$$|\mathcal{R}| = N \cdot [1 - H(p_2^*) - \epsilon], \tag{6.9}$$

and combine the output from the previous phase as message to construct polar codeword $x_{1:N}^{(b)}$. More formally, if we denote the reordering permutation for BSC as π, then the encoder of this phase can be expressed as

$$\begin{aligned} x_{1:N}^{(b)} &= \omega_{1:N}^{(b)} \times \boldsymbol{G}_N, \\ \pi\left(\omega_{1:N}^{(b)}\right) &= \left[\begin{array}{ccccccc} s_{1:|\mathcal{R}|}^{(b)} & | & \tilde{\boldsymbol{U}}_b^T & | & w_{1:|\mathcal{I}|}^{(b)} & | & \tilde{\boldsymbol{V}}_b^T \quad | \quad 0 \end{array}\right], \end{aligned} \tag{6.10}$$

for every $b \in \{1, \ldots, B\}$, where $\boldsymbol{G}_N$ is the polar generator matrix with size N. That is, the codewords generated from BEC encoding phase are transposed and embedded into the messages of the BSC encoding process. We denote these codewords by a $B \times N$ matrix $\boldsymbol{X}$. The proposed encoder is illustrated in Figure 6.2.

6.4 Decoder for the Main Channel

The codewords $x_{1:N}^{(b)}$ are transmitted through both the main channel and the wiretap channel. After receiving the output sequence $y_{1:N}^{(b)}$ for all $b \in \{1, \ldots, B\}$, the task of the decoder of main channel is to make estimates for all the information and random bits. In particular, the decoder aims to recover $u_{1:|\mathcal{A}^c|}^{(j)}$, $v_{1:|\mathcal{A}|}^{(k)}$, $w_{1:|\mathcal{I}|}^{(b)}$, $r_{1:|\mathcal{A}|}^{(j)}$, and $s_{1:|\mathcal{R}|}^{(b)}$ successfully with high probability. As that of the encoding process, the decoding process also works in phases (see Figure 6.3).

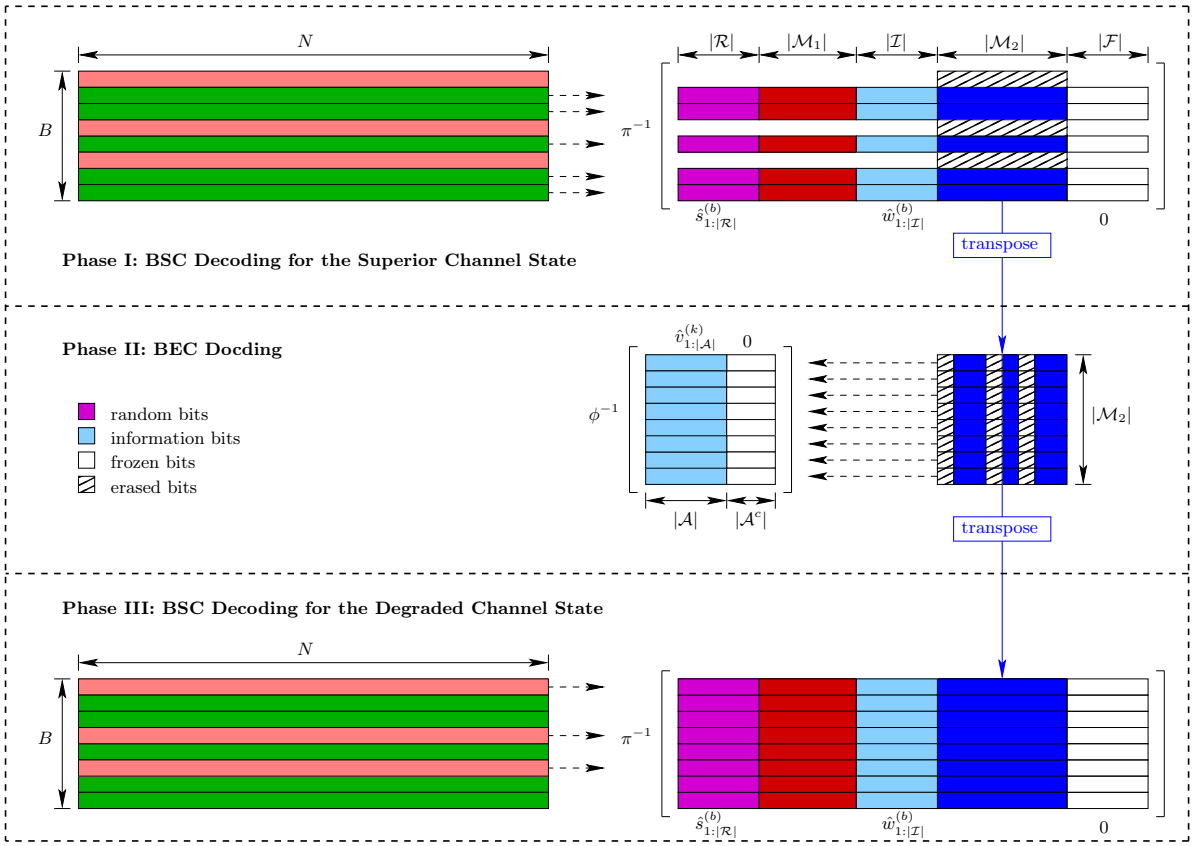

Figure 6.3: **Decoder at the main channel receiver given the knowledge of the channel states information.** The decoder also works in phases. After decoding blocks in the superior channel state, the decoder is enable to decode the blockwise information through BEC SC decoder. Finally, using the output from previous phase, blocks in the degraded channel state can also be decoded.

6.4.1 Phase I: BSC Decoding for the Superior Channel State

In this phase, using the BSC SC decoder, channels corresponding to the superior state are decoded. More precisely, since the receiver knows the channel states, it can adopt the correct SC decoder to obtain estimates $\hat{\omega}^{(b)}_{1:N}$ from $y^{(b)}_{1:N}$ for every b corresponding to the superior channel state. To this end,

the decoder adopted in this phase is the classical BSC SC polar decoder with parameter p_1, i.e.,

$$\hat{\omega}_i^{(b)} = \begin{cases} 1, & \text{if } i \notin \mathcal{F}, \text{ and } \frac{\mathcal{W}_{1,N}^{(i)}(y_{1:N}^{(b)}, \hat{\omega}_{1:i-1}^{(b)}|1)}{\mathcal{W}_{1,N}^{(i)}(y_{1:N}^{(b)}, \hat{\omega}_{1:i-1}^{(b)}|0)} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

in the order i from 1 to N, and $\mathcal{W}_{1,N}^{(i)}$ is the i-th polarized channel from BSC(p_1). Then, for every b corresponding to the superior channel state, the decoder can obtain the messages (with the knowledge of the frozen symbols corresponding to $\mathcal{F}$ indices)

$$\pi\left(\hat{\omega}_{1:N}^{(b)}\right) = \left[\begin{array}{c|c|c|c|c} \hat{s}_{1:|\mathcal{R}|}^{(b)} & \hat{\tilde{\boldsymbol{U}}}_b^T & \hat{w}_{1:|\mathcal{I}|}^{(b)} & \hat{\tilde{\boldsymbol{V}}}_b^T & 0 \end{array} \right].$$

However, for the blocks with degraded channel states, one cannot decode reliably because the frozen bits corresponding to set $\mathcal{M}_2$ are unknown at the decoder. At this point, we use the next phase to decode these frozen bits using a BEC SC decoder. To proceed, we construct a $B \times |\mathcal{M}_2|$ matrix $\hat{\tilde{\boldsymbol{V}}}^T$ such that its rows corresponding to the superior state are determined in previous decoding process, while the ones corresponding to the degraded states are all set to erasures.

6.4.2 Phase II: BEC Decoding

In this phase, we decode the frozen bits with respect to the degraded channel state. More precisely, each row of matrix $\hat{\tilde{\boldsymbol{V}}}$, denoted by $\hat{\tilde{\boldsymbol{V}}}_k$ for $k \in \{1, \ldots, |\mathcal{M}_2|\}$, is considered as the input to the decoder, and the receiver aims to obtain an estimate of the information bits from it using BEC SC decoder.

To this end, the decoder adopted in this phase is the classical BEC SC decoder with parameter ϱ_2, i.e.,

$$\hat{\nu}_b^{(k)} = \begin{cases} 1, & \text{if } b \in \mathcal{A}, \text{ and } \frac{\mathcal{W}_{\mathrm{e},B}^{(b)}(\hat{\tilde{\boldsymbol{V}}}_k, \hat{\nu}_{1:b-1}^{(k)}|1)}{\mathcal{W}_{\mathrm{e},B}^{(b)}(\hat{\tilde{\boldsymbol{V}}}_k, \hat{\nu}_{1:b-1}^{(k)}|1)} \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

in the order b from 1 to B, and $\mathcal{W}_{\mathrm{e},B}^{(b)}$ is the b-th polarized channel from BEC(ϱ_2). Then, for every k, the decoder can declare

$$\phi\left(\hat{\nu}_{1:B}^{(k)}\right) = \left[\begin{array}{c|c} \hat{v}_{1:|\mathcal{A}|}^{(k)} & 0 \end{array} \right].$$

At this point, the decoder can reconstruct all erased bits as well. More precisely, the erased rows in $\hat{\tilde{\boldsymbol{V}}}^T$ can be recovered, and they can be further utilized to decode the information bits in blocks with the degraded channel state in the next phase.

6.4.3 Phase III: BSC Decoding for the Degraded Channel State

In this phase, the remaining blocks from Phase I are decoded by using BSC SC decoders with respect to degraded channel states. In particular, bits in the frozen set for the degraded channel state (set $\mathcal{F}$ and set $\mathcal{M}_2$) are known due to the previous phases. Hence, the receiver can decode from $y_{1:N}^{(b)}$ using BSC SC decoder with parameter p_2, i.e.,

$$\hat{\omega}_i^{(b)} = \begin{cases} 1, & \text{if } i \notin \mathcal{F},\, n \notin \mathcal{M}_2, \text{ and } \frac{\mathcal{W}_{2,N}^{(i)}(y_{1:N}^{(b)}, \hat{\omega}_{1:i-1}^{(b)}|1)}{\mathcal{W}_{2,N}^{(i)}(y_{1:N}^{(b)}, \hat{\omega}_{1:i-1}^{(b)}|0)} \geq 1, \\ \hat{\tilde{v}}_{bi}^T, & \text{if } i \in \mathcal{M}_2, \\ 0, & \text{otherwise,} \end{cases}$$

in the order i from 1 to N, and $\mathcal{W}_{2,N}^{(i)}$ is the i-th polarized channel from BSC(p_2). Then, for every b corresponding to the degraded channel state, the

decoder declares

$$\psi\left(\hat{\omega}_{1:N}^{(b)}\right) = \left[\begin{array}{c|c|c|c|c} \hat{s}_{1:|\mathcal{R}|}^{(b)} & \hat{\boldsymbol{U}}_b^T & \hat{w}_{1:|\mathcal{I}|}^{(b)} & \hat{\boldsymbol{V}}_b^T & 0 \end{array}\right].$$

Hence, after this decoding procedure, the receiver makes an estimate $\hat{\tilde{\boldsymbol{U}}}$ of matrix $\tilde{\boldsymbol{U}}$, which further implies all information bits in $u_{1:|\mathcal{A}^c|}^{(j)}$ are decoded. Note that, in addition to information bits, all random bits are decoded reliably at Bob as well. However, in order to guarantee security, we set these bits random (instead of information).

6.5 Achievable Rate and Reliability

The proposed hierarchical scheme allows for recovering all information bits (represented by light blue in Figure 6.2) reliably, as long as the designed rates of polar codes do not exceed the corresponding channel capacities. Hence, the achievable rate is given by

$$\begin{aligned} R &= \frac{1}{NB}\left(|\mathcal{M}_2| \times |\mathcal{A}| + |\mathcal{M}_1| \times |\mathcal{A}^c| + B \times |\mathcal{I}|\right) \\ &= [H(p_2) - H(p_1)] \times [\varrho_1 - \epsilon] + [H(p_2^*) - H(p_1^*)] \times [\varrho_2 + \epsilon] + [H(p_1^*) - H(p_2)] \\ &= [H(p_1^*) - H(p_1)] \times \varrho_1 + [H(p_2^*) - H(p_2)] \times \varrho_2 - \delta(\epsilon), \end{aligned} \tag{6.11}$$

where we have used (6.4), (6.5), (6.6), (6.7), and (6.8), and $\delta(\epsilon) \to 0$ as N, $B \to \infty$. In this scheme, B number of N-length polar codes are decoded in Phase I and III in total, and $|\mathcal{M}_2|$ number of B-length polar codes are decoded in Phase II. Hence, the decoding error probability is upper bounded by

$$\Pr\{\mathsf{M} \neq \hat{\mathsf{M}} | \mathsf{Y}_{1:NB}, \mathsf{S}\} \leq B \cdot 2^{-N^\beta} + |\mathcal{M}_2| \cdot 2^{-B^\beta}, \tag{6.12}$$

where $\beta < 1/2$; and, M is the collection of random variables representing for all information bits (its realizations include $u_{1:|\mathcal{A}^c|}^{(j)}$, $v_{1:|\mathcal{A}|}^{(k)}$, and $w_{1:|\mathcal{I}|}^{(b)}$), and $\hat{\mathsf{M}}$ is the estimate of M obtained at the legitimate receiver. Noting that the right hand side of (6.12) tends to 0 when implemented with properly large B and N, the proposed scheme achieves the upper bound given by (6.3) reliably.

6.6 Security

Assume that, in addition to $y_{1:N}^{*(b)}$, a genie reveals Eve all information bits $u_{1:|\mathcal{A}^c|}^{(j)}$, $v_{1:|\mathcal{A}|}^{(k)}$, and $w_{1:|\mathcal{I}|}^{(b)}$. Under this condition, we show that all random bits can be reliably decoded at Eve. More precisely, the decoder designed for the eavesdropper also works in phases, similar to the one for the main channel (see Figure 6.4). We sketch the procedures for the decoder at Eve as follows:

1) Phase I (BSC Decoding for the Superior Channel State): The decoder still works over the blocks with the superior channel state. However, for the wiretap channel with superior channel state, the frozen set consists of bits not only in set $\mathcal{F}$, but also in sets $\mathcal{M}_2$ and $\mathcal{I}$. Since we have assumed the information bits are known at Eve, the classical BSC(p_1^*) SC decoder can be used to decode the random bits.

2) Phase II (BEC Decoding): In the second phase, we aim to recover the unknown frozen bits corresponding to the degraded channel state, where a similar scheme as that of the main receiver is adopted. More precisely, we utilize the BEC(ϱ_2) SC decoder over each row of the matrix after transpose.

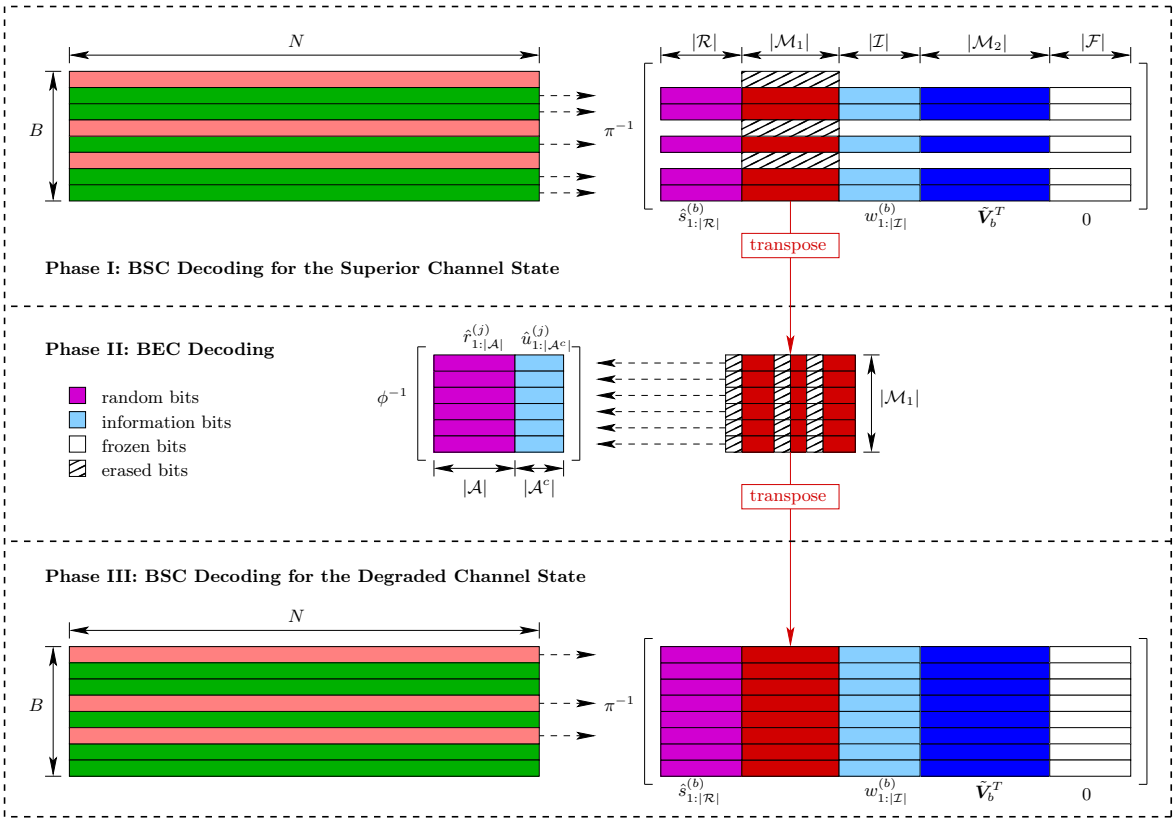

Figure 6.4: **Decoder at the eavesdropper given the knowledge of the channel states information and information bits.** The decoder at the eavesdropper works analogously to the one at the main channel.

This scheme successively recovers the erased elements, as the frozen bits for this BEC is the information bits $u^{(j)}_{1:|\mathcal{A}^c|}$ and they are assumed to be known.

3) Phase III (BSC Decoding for the Degraded Channel State): Finally, the decoded result from the BEC decoding phase is utilized at the BSC decoding for the degraded state, where the classical BSC(p_2^*) SC decoder is adopted.

By adopting this hierarchical polar decoder, Eve can decode all random bits with high probability, i.e.,

$$\Pr\{\mathsf{R} \neq \hat{\mathsf{R}} | \mathsf{Y}^*_{1:NB}, \mathsf{M}, \mathsf{S}\} \leq B \cdot 2^{-N^\beta} + |\mathcal{M}_1| \cdot 2^{-B^\beta}, \tag{6.13}$$

where R is the collection of random variables representing for random bits (its realization include $r^{(j)}_{1:|\mathcal{A}|}$ and $s^{(b)}_{1:|\mathcal{R}|}$), and $\hat{\mathsf{R}}$ is the estimate of R. Then, using Fano's inequality, together with (6.13), we have

$$\begin{aligned} &H(\mathsf{R}|\mathsf{Y}^*_{1:NB}, \mathsf{M}, \mathsf{S}) \\ &\quad \leq [B \cdot 2^{-N^\beta} + |\mathcal{M}_1| \cdot 2^{-B^\beta}] \cdot [|\mathcal{R}| \cdot B + |\mathcal{A}| \cdot |\mathcal{M}_1|] \\ &\qquad + H(B \cdot 2^{-N^\beta} + |\mathcal{M}_1| \cdot 2^{-B^\beta}). \end{aligned} \tag{6.14}$$

Based on this, the following steps provide an upper bound (omitting the subscript of Y^*):

$$\begin{aligned} I(\mathsf{M}; \mathsf{Y}^*|\mathsf{S}) &= I(\mathsf{M}, \mathsf{R}; \mathsf{Y}^*|\mathsf{S}) - [H(\mathsf{R}|\mathsf{M}, \mathsf{S}) - H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S})] \\ &\overset{(a)}{=} I(\mathsf{M}, \mathsf{R}; \mathsf{Y}^*|\mathsf{S}) - H(\mathsf{R}) + H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) \\ &\overset{(b)}{\leq} NB \cdot C_{\text{CSI-D}}(\mathcal{W}^*) - H(\mathsf{R}) + H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) \\ &\overset{(c)}{=} NB \cdot C_{\text{CSI-D}}(\mathcal{W}^*) - |\mathcal{A}| \cdot |\mathcal{M}_1| - B \cdot |\mathcal{R}| + H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) \\ &\overset{(d)}{=} NB \cdot C_{\text{CSI-D}}(\mathcal{W}^*) - B[\varrho_1 - \epsilon] \cdot N[H(p_2^*) - H(p_1^*)] \\ &\qquad - B \cdot N[1 - H(p_2^*) - \epsilon] + H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) \\ &= NB \cdot C_{\text{CSI-D}}(\mathcal{W}^*) - NB \cdot \varrho_1[1 - H(p_1^*)] \\ &\qquad - NB \cdot \varrho_2[1 - H(p_2^*)] + H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) - NB \cdot \delta'(\epsilon) \\ &\overset{(e)}{=} H(\mathsf{R}|\mathsf{Y}^*, \mathsf{M}, \mathsf{S}) - NB \cdot \delta'(\epsilon), \end{aligned}$$

where $\delta'(\epsilon) \to 0$ as N, $B \to \infty$, and

(a) follows as R is independent of M and S;

(b) is due to the definition of channel $\mathcal{W}^*$'s capacity with CSI-D;

(c) is due to the assumption that R is uniform;

(d) is due to equations (6.5), (6.4), and (6.9);

(e) is due to the ergodic capacity of the fading eavesdropper channel with channel state information known only at the decoder, i.e.,

$$C_{\text{CSI-D}}(\mathcal{W}^*) \leq C_{\text{CSI-ED}}(\mathcal{W}^*) = \varrho_1[1 - H(p_1^*)] + \varrho_2[1 - H(p_2^*)].$$

Finally, combining with (6.14), we have

$$\frac{1}{NB} I(\mathsf{M}; \mathsf{Y}_{1:NB}^* | \mathsf{S}) \to 0,$$

as N and B tends to infinity (with proper choice of the their scaling relationship). Hence, the proposed scheme achieves the secrecy constraint.

6.7 The Scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$

Here, we discuss the extension of the aforementioned coding scheme to the scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$. Combined with the result discussed earlier in this section, this completes the proof for all possible cases of simultaneous fading. Note that although $p_1^* \leq p_2$, in each fading block the main channel is

still stronger than the eavesdropper channel because of simultaneous fading. to this end, the upper bound of form (6.3) still holds in this scenario.

From the previous scenario, the key idea for hierarchical polar coding scheme is setting the size of random bits be $NB \cdot C_{\text{CSI-D}}(\mathcal{W}^*)$ and setting the size of information bits be $NB \cdot SC_{\text{CSI-D}}(\mathcal{W})$. Based on this observation, the encoder for the scenario discussed here is illustrated in Figure 6.5. Note that we still have five categories for channel indices after polarization. $\mathcal{R}$ and $\mathcal{F}$ remain the same as previous case, but we do not have pure information set in this scenario due to $p_1^* \leq p_2$. Instead, a new set $\mathcal{M}_3$ contains coding results from random bits and frozen bits. More precisely, parameters shown in the figure are defined as follow:

$$
\begin{aligned}
|\mathcal{R}| &= N \cdot [1 - H(p_2^*) - \epsilon], \\
|\mathcal{M}_1| &= N \cdot [H(p_2^*) - H(p_2)], \\
|\mathcal{M}_2| &= N \cdot [H(p_1^*) - H(p_1)], \\
|\mathcal{M}_3| &= N \cdot [H(p_2) - H(p_1^*)], \\
|\mathcal{F}| &= N \cdot H(p_1), \\
|\mathcal{A}| &= B \cdot [\varrho_1 - \epsilon], \\
|\mathcal{A}^c| &= B \cdot [\varrho_2 + \epsilon].
\end{aligned}
$$

Then, the encoding procedure works analog to the previous scenario, except that three sets of BEC encoding are performed and the resulting codewords are transposed and embedded into the second phase. In particular, the sketch of hierarchical coding scheme is sketched as follow:

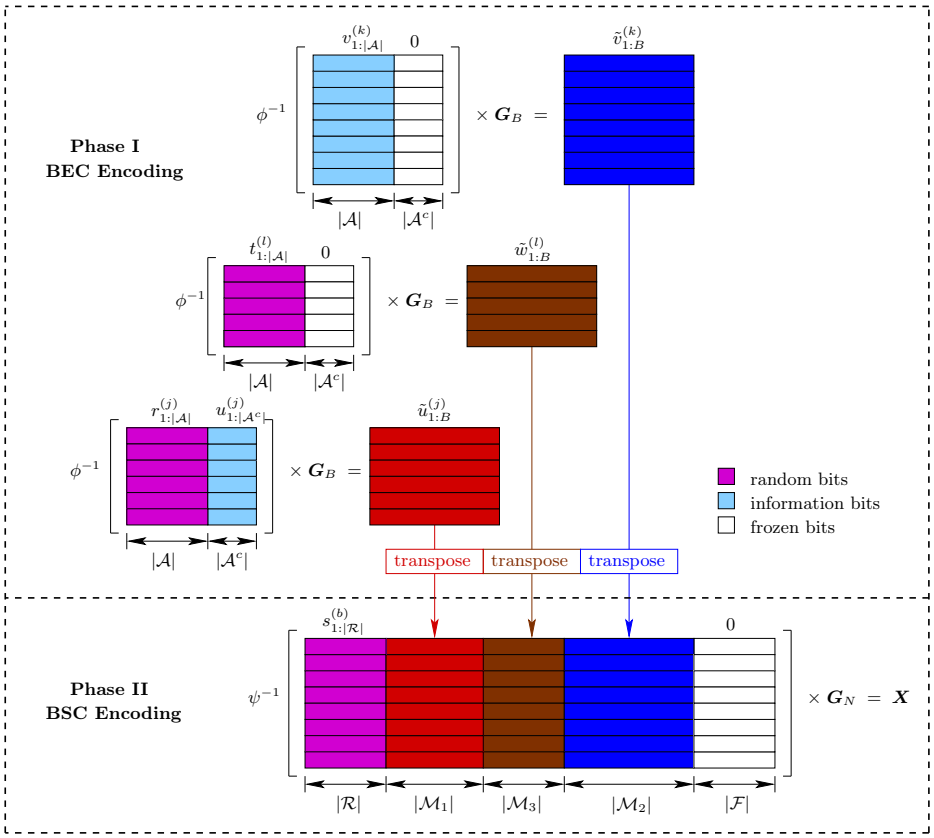

Figure 6.5: **Hierarchical polar encoder for the scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$.** For this scenario, there is no pure information index set, but a mixed index set with random bits and frozen bits ($\mathcal{M}_3$ in the figure).

1) Phase I (BEC Encoding): Three sets of BEC polar codes, with either random bits or information bits encoded, are considered in this phase:

- Random bits $r_{1:|\mathcal{A}|}^{(j)}$ combined with information bits $u_{1:|\mathcal{A}^c|}^{(j)}$ are encoded to generate BEC polar codes $\tilde{u}_{1:B}^{(j)}$, for each $j \in 1, \ldots, |\mathcal{M}_1|$;
- Information bits $v_{1:|\mathcal{A}|}^{(k)}$ combined with frozen bits 0 are encoded to generate BEC polar codes $\tilde{v}_{1:B}^{(k)}$, for each $k \in 1, \ldots, |\mathcal{M}_2|$;

- Random bits $t_{1:|\mathcal{A}|}^{(l)}$ combined with frozen bits 0 are encoded to generate BEC polar codes $\tilde{w}_{1:N}^{(l)}$, for each $l \in 1, \ldots, |\mathcal{M}_3|$.

2) Phase II (BSC Encoding): The encoded result from the previous phase are transposed and embedded into the message of this BEC encoding phase. more precisely, the encoded bits are combined with random bits $s_{1:|\mathcal{R}|}^{(b)}$ and frozen bits 0 to generate BSC polar codes $x_{1:N}^{(b)}$, for each $b \in 1, \ldots, B$.

The decoder at the main channel also works in phases. Quite similar to the previous case, the sketch of decoder is as follow (illustrated in Figure 6.6):

1) Phase I (BSC Decoding for the Superior Channel State): The Decoder can decode the block with respect to the superior state using BSC(p_1) SC decoder, because the frozen bits (set as 0) are known with respect to the superior channel state.

2) Phase II (BEC Decoding): In this phase, the decoder can recover the unknown frozen bits corresponding to the degraded channel state. More precisely, by adding erasures to the decoded bits in set $\mathcal{M}_3$ and $\mathcal{M}_2$ from the previous phase and forming the input to the BEC(ϱ_2) SC decoder, both the random bits and information bits can be decoded by choosing frozen bits as 0. Meanwhile, this scheme successively recovers the erased elements, which are utilized to help to decoding in the next phase.

3) Phase III (BSC Decoding for the Degraded Channel State): At last, by using the decoded frozen bits from the previous phase, the decoder can

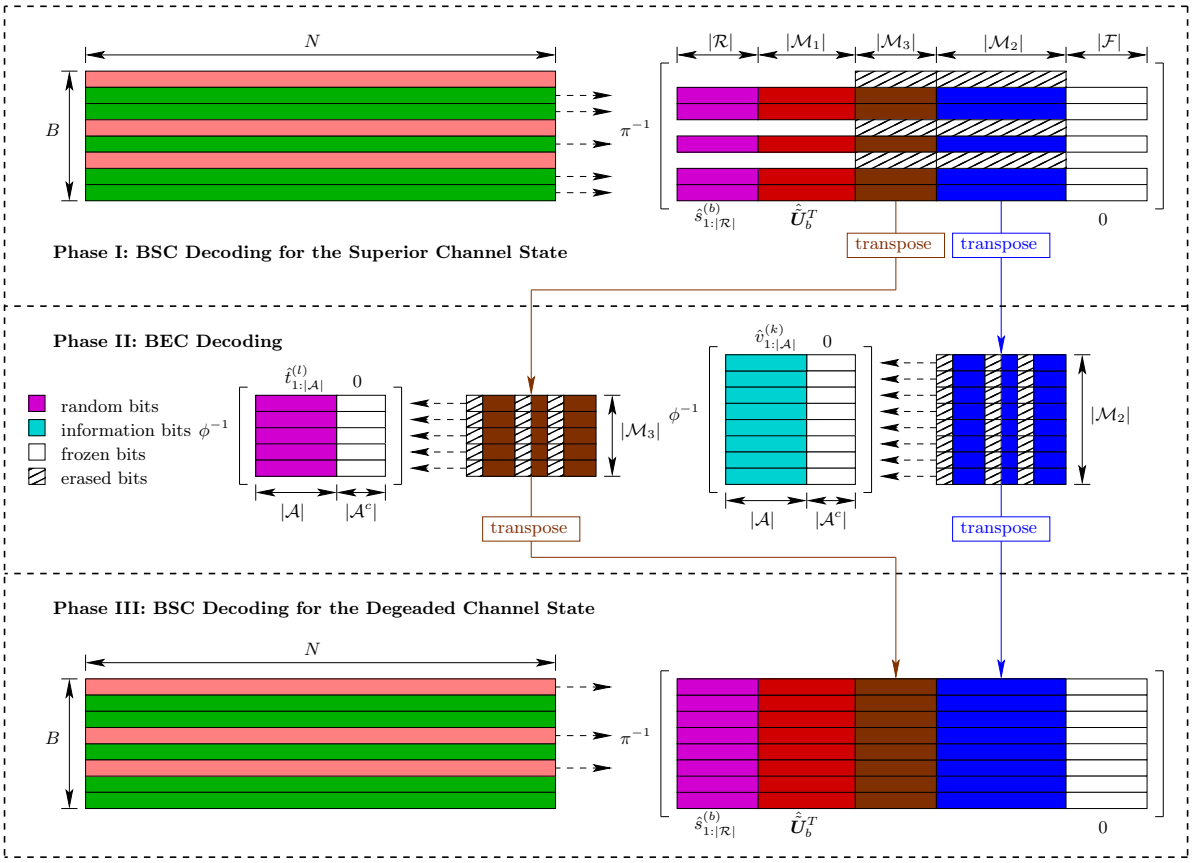

Figure 6.6: **Decoder at the main channel for the scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$.** The decoder here is similar to the one for the scenario of $p_1 \leq p_2 \leq p_1^* \leq p_2^*$, which also works in phases to hierarchically decode all information bits and random bits.

decode all blocks with respect to the degraded state using BSC(p_2) SC decoder.

In this way, all information bits and random bits can be recovered reliably, i.e., (6.12) still holds in this scenario. Meanwhile, we have

$$R = \frac{1}{NB}\left(|\mathcal{M}_2| \times |\mathcal{A}| + |\mathcal{M}_1| \times |\mathcal{A}^c|\right)$$

$$= [H(p_1^*) - H(p_1)] \times \varrho_1 + [H(p_2^*) - H(p_2)] \times \varrho_2 - \delta(\epsilon),$$

which means the upper bound (6.3) is also achieved in this scenario.

On the other hand, for the proof of security, we assume the receiver from the eavesdropper channel knows all the information bits, i.e., $u_{1:|\mathcal{A}^c|}^{(j)}$ and $v_{1:|\mathcal{A}|}^{(k)}$ in this scenario. Then, the eavesdropper can decode all random bits by following steps (also see Figure 6.7):

1) Phase I (BSC Decoding for the Superior Channel State): The eavesdropper can decode all random bits in the blocks with respect to the superior channel state using BSC(p_1^*) SC decoder, because the frozen bits for these blocks are given by $\mathcal{F}$ and $\mathcal{M}_2$, which are known by the assumption.

2) Phase II (BEC Decoding): By adding erasures blockwise to the decoded bits in set $\mathcal{M}_1$ and $\mathcal{M}_3$ from previous phase, eavesdropper can decode both the random bits using BEC(ϱ_2) SC decoder by choosing frozen bits as $u_{1:|\mathcal{A}^c|}^{(j)}$ and 0 respectively.

3) Phase III (BSC Decoding for the Degraded Channel State): Finally, for the degraded channel state, since the corresponding frozen bits are all known from the previous phase, eavesdropper can recover all random bits using BSC(p_2^*) SC decoder.

Hence, all random bits can be decoded reliably, i.e., (6.13) still holds in this scenario. Then, the same procedures as the previous scenario complete the proof of security.

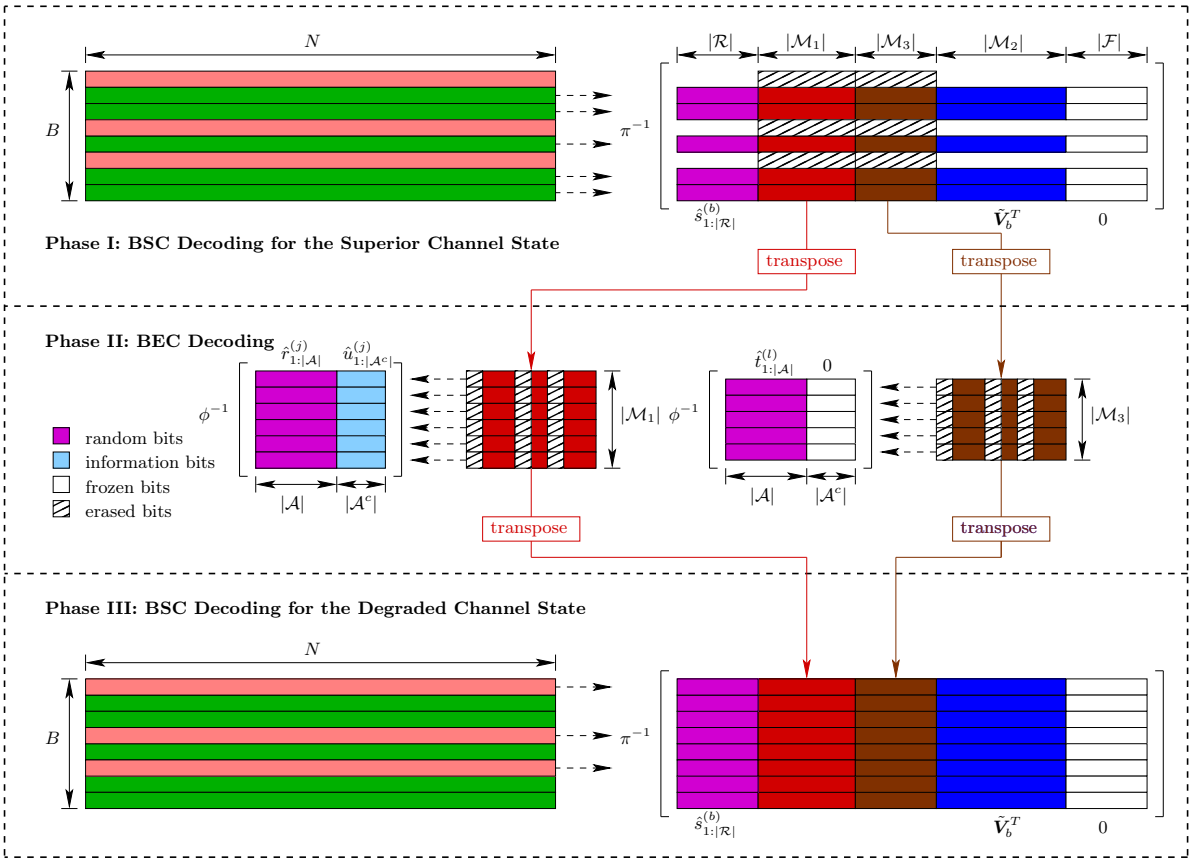

Figure 6.7: **Decoder at the eavesdropper for the scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$.** Similar to the one for the scenario of $p_1 \leq p_2 \leq p_1^* \leq p_2^*$, decoder at the eavesdropper here also works hierarchically to decode all random bits given knowledge of all information bits.

To this end, combining this section with the previous analysis in this chapter, the proposed polar coding scheme can achieve the secure capacity reliably and securely, when the main channel and wiretap channel fade simultaneously.

6.8 Discussion on Independent Fading Case

In this section, we discuss the case of independent fading for main channel and eavesdropper. More precisely, the main channel has probability ϱ_1 to behave in the superior fading state, while the eavesdropper channel has probability ϱ_1^* to behave in the superior state (independent of the main channel). The essential difference from the simultaneous fading case is that the main channel may be degraded comparing to the eavesdropper channel for certain fading blocks, which leads to challenges for coding. Still, we distinguish two scenarios based on the relation between parameters p_1^* and p_2, and we show that the performance evaluations behave quite different between the two cases.

1) In the scenario of $p_1 \leq p_2 \leq p_1^* \leq p_2^*$, for those fading blocks where the main channel is in the superior state and eavesdropper channel is in the degraded state, the main channel is still stronger than the eavesdropper channel due to $p_2 \leq p_1^*$. To this end, the upper bound for security capacity can be expressed as

$$\begin{aligned} SC_{\text{CSI-D}} &\leq SC_{\text{CSI-ED}} \\ &= \max_{p(x|s,s^*)} [I(\mathsf{X};\mathsf{Y}|\mathsf{S},\mathsf{S}^*) - I(\mathsf{X};\mathsf{Y}^*|\mathsf{S},\mathsf{S}^*)] \\ &= \varrho_1\varrho_1^*[H(p_1^*) - H(p_1)] + \varrho_1\varrho_2^*[H(p_2^*) - H(p_1)] \\ &\quad + \varrho_2\varrho_1^*[H(p_1^*) - H(p_2)] + \varrho_2\varrho_2^*[H(p_2^*) - H(p_2)] \\ &= \varrho_1^* H(p_1^*) + \varrho_2^* H(p_2^*) - \varrho_1 H(p_1) - \varrho_2 H(p_2), \end{aligned} \tag{6.15}$$

where random variables S and S^* are the fading states for the main channel and eavesdropper respectively.

Because the main channel is still stronger than the eavesdropper channel for all blocks, we can reuse Figure 6.2, Figure 6.3, and Figure 6.4) to illustrate the encoding and decoding scheme, except that the encoding and decoding for indices in set $\mathcal{M}_1$ should be substituted with respect to parameters for channel BEC(ϱ_2^*). To this end, all information bits and random bits can still be decoded reliably, which implies

$$\begin{aligned} R &= \frac{1}{NB}\left(|\mathcal{M}_2| \times |\mathcal{A}| + |\mathcal{M}_1| \times |\mathcal{A}^{*c}| + |\mathcal{I}| \times B\right) \\ &= [H(p_2) - H(p_1)] \times [\varrho_1 - \epsilon] + [H(p_2^*) - H(p_1^*)] \times [\varrho_2^* + \epsilon] \\ &\quad + [H(p_1^*) - H(p_2)] \\ &= \varrho_1^* H(p_1^*) + \varrho_2^* H(p_2^*) - \varrho_1 H(p_1) - \varrho_2 H(p_2) - \delta'(\epsilon). \end{aligned}$$

The reliability and security proof follows the same steps as simultaneous fading case. Hence, the achievable rate, matching the upper bound given by (6.15), approaches the secure capacity of the system.

2) In the scenario of $p_1 \leq p_1^* \leq p_2 \leq p_2^*$, for those fading blocks where the main channel is in the superior state and eavesdropper channel is in the degraded state, the eavesdropper channel is stronger. This situation contributes nothing to the secure capacity of the system, i.e.,

$$\begin{aligned} SC_{\text{CSI-D}} &\leq SC_{\text{CSI-ED}} \\ &= \max_{p(x|s,s^*)} [I(\mathsf{X};\mathsf{Y}|\mathsf{S},\mathsf{S}^*) - I(\mathsf{X};\mathsf{Y}^*|\mathsf{S},\mathsf{S}^*)] \\ &= \varrho_1\varrho_1^*[H(p_1^*) - H(p_1)] + \varrho_1\varrho_2^*[H(p_2^*) - H(p_1)] \\ &\quad + \varrho_2\varrho_1^* \cdot 0 + \varrho_2\varrho_2^*[H(p_2^*) - H(p_2)] \end{aligned}$$

$$= \varrho_1 \varrho_1^* H(p_1^*) + \varrho_2^* H(p_2^*) - \varrho_1 H(p_1) - \varrho_2 \varrho_2^* H(p_2). \tag{6.16}$$

Hence, if we still utilize the hierarchical polar encoding and decoding scheme, a gap may exist comparing with the upper bound given by (6.16). An effective polar coding scheme for the case of independent fading and in particular for this scenario is still open.

6.9 Summary

In this chapter, a hierarchical polar coding scheme is proposed for binary symmetric wiretap channels with block fading. By exploiting an erasure decoding approach at the receiver, this scheme utilizes the polarization of degraded binary symmetric channels to survive from the impact of fading. Meanwhile, to combat with eavesdropping, random bits are injected into the encoded symbols, and the resulting coding scheme is shown to achieve the secrecy capacity for the case of simultaneous fading of the main channel and eavesdropper channel. Although we consider binary symmetric channels in this paper, the hierarchical coding scheme can be applied as a general method to other scenarios (such as fading blocks with more states) for simultaneously resolving fading and security problems. Noting that AWGN channels with BPSK modulation and demodulation resembles a BSC, the proposed scheme covers a fairly large set of practically relevant channel models.

Chapter 7

Information-Theoretic Analysis of Haplotype Assembly

7.1 Background of Haplotype Assembly

Diploid organisms, including humans, have homologous pairs of chromosomes where one chromosome in a pair is inherited from mother and the other from father. The two chromosomes in a pair are structurally similar and basically carry the same type of information but are not identical. More specifically, chromosomes in a pair differ at a small fraction of positions (i.e., loci). Such variations are referred to as single nucleotide polymorphisms (SNPs); in humans, frequency of SNPs is approximately 1 base in 1000. A haplotype is the string of SNPs on a single chromosome in a homologous pair (see Figure 7.1). Haplotype information is essential for understanding genetic causes of various diseases and for advancement of personalized medicine. However, direct analysis and identification of a haplotype is generally challenging, costly, and time and labor intensive.

Alternatively, single individual haplotypes can be assembled from short reads provided by high-throughput sequencing systems. These systems rely on so-called shotgun sequencing to oversample the genome and generate a

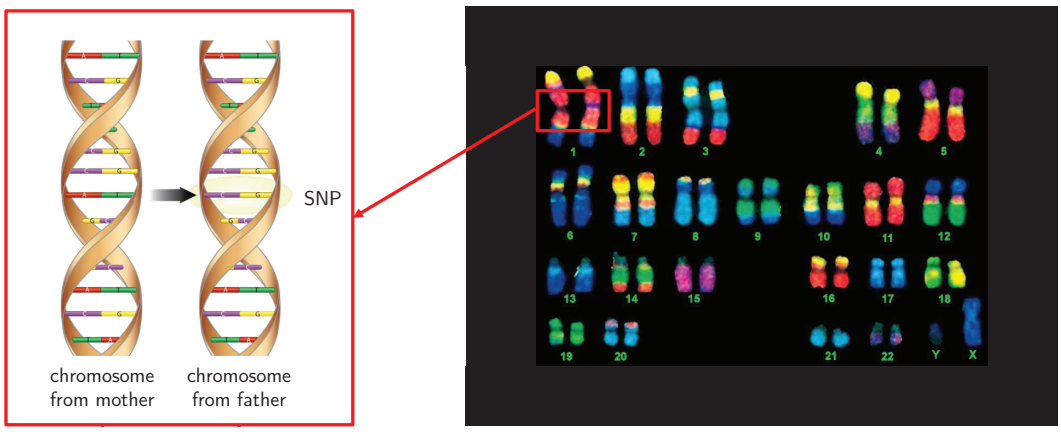

Figure 7.1: **Illustration of SNPs and haplotypes.** In a diploid cell (e.g. human cell in the figure), paired chromosomes are inherited from father and mother respectively. The collection of differences between these paired chromosomes, i.e., SNPs, is denoted as a haplotype.

redundant library of short reads. The reads are mapped to a reference and the individual genome is assembled following consensus of information provided by the reads. The length of each read (i.e., DNA fragment) in state-of-the-art sequencing systems is typically $100 - 1000$ base pairs [68]. Note that this length is comparable to the average distance between SNPs on chromosomes. Therefore, single reads rarely cover more than one variant site which is needed to enable haplotype assembly. Moreover, the origin of a read (i.e., to which chromosome in a pair the read belongs) is unknown and needs to be inferred [69].

Paired-end sequencing [70], also known as mate-paired sequencing [71], helps overcome these problems. This process generates pairs of short reads that are spaced along the target genome, where the spacing (so-called insert

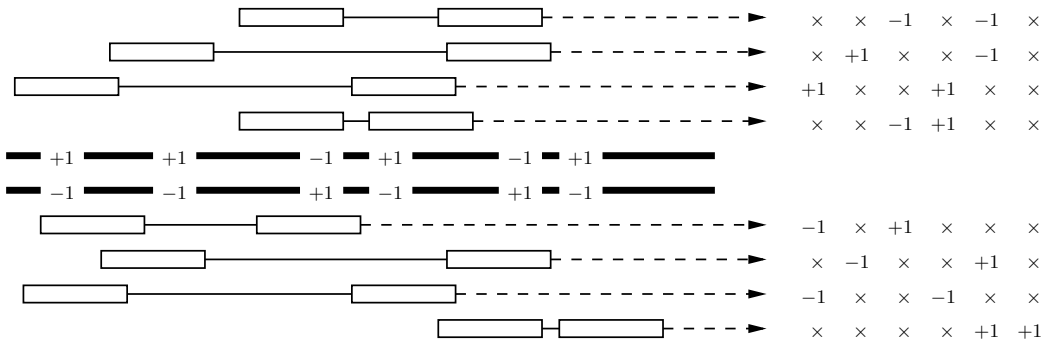

Figure 7.2: **Paired-end reads sampling two chromosomes in a homologous pair.** Rectangles linked by the lines above and below the target chromosome pair represent paired-end reads, and their relative positions indicate their location along the chromosomes. In this example, 6 SNPs and 8 reads are presented.

size) between the two reads in a pair is known. The mate-pairs allow acquisition of the information about distant SNPs on the same haplotype, and thus help assemble the haplotype. Figure 7.2 illustrates how paired-end reads may cover two or more variant sites along a homologous chromosome pair. The goal of haplotype assembly is to identify the chromosome from which fragments are sampled, and to reconstruct the haplotype sequences. When there are no sequencing errors, a fragment conflict graph framework [72] converts the original problem into partitioning of the set of reads into two subsets, each collecting the reads that belong to the same chromosome in a pair. For erroneous data, it poses haplotyping as an optimization problem of minimizing the number of transformation steps needed to generate a bipartite graph [73]. This leads to various formulations of the haplotype assembly problem including minimum fragment removal (MFR), minimum SNP removal (MSR), and minimum error

correction (MEC) [72]. The last one, MEC, has been the most widely used criterion for haplotype assembly, and is characterized by an inherent connection with the independent error model.

In this chapter, we analyze the haplotype assembly problem from information theoretic perspective. In particular, we determine necessary and sufficient conditions for haplotype assembly, both in the absence of noise as well as for the case where data is erroneous.

7.2 Problem Formulation

As detailed in the introduction, a single nucleotide polymorphism (SNP) is a variation in a DNA sequence where two corresponding bases at a specific location on the chromosomes in a homologous pair differ from each other. Typically, diploid organisms have only two possible variants at a SNP site, i.e., their SNPs are typically biallelic. For the sake of convenience, we denote one of the two variants as $+1$ while the other one we denote as -1. With this notation, a haplotype sequence $\boldsymbol{h}$ comprising information about all SNP sites on one of the chromosomes in a homologous pair can be represented by a string with elements in $\{+1, -1\}$, while the haplotype associated with the other chromosome in the pair is its additive inverse $-\boldsymbol{h}$, where we denote

$$\boldsymbol{h} = (h_1, h_2, \ldots, h_n),$$

and n is the length of haplotypes (i.e., the number of SNPs within each chromosome in a pair).

Each paired-end read acquired in a shotgun sequencing experiment contains partial information about either of these two haplotypes. Consider a set of discrete random variables c_i, where $i \in \{1, \ldots, m\}$ and m denotes the number of reads. Let c_i identify the origin of read i, i.e., c_i carries information about the chromosome membership for read i. More precisely,

$$c_i = \begin{cases} +1, & \text{if read } i \text{ is sampled from } \boldsymbol{h}, \\ -1, & \text{if read } i \text{ is sampled from } -\boldsymbol{h}. \end{cases} \tag{7.1}$$

Due to the limitation of read lengths and relatively rare occurrence of SNPs, only a small fraction of variant sites is covered by a read. Formally, the information about a haplotype provided by a paired-end read $\boldsymbol{r}_i$ can be represented by a sequence that consists of symbols from the alphabet $\{+1, -1, \times\}$, where "$\times$" indicates lack of information about a variant site. Let us collect the relevant information provided by the reads in an $m \times n$ matrix $\boldsymbol{R}$ having rows corresponding to paired-end reads and columns corresponding to SNP sites. The ith row of $\boldsymbol{R}$ (i.e., read i) is denoted as $\boldsymbol{r}_i$, and the jth element of $\boldsymbol{r}_i$ is denoted as r_{ij}. Typically, since the length of a haplotype is much larger than the number of SNPs covered by a read, only few entries in each row are numerical (ignoring the occurrence of bursty variations).

Note that, in the absence of sampling noise, every observed element r_{ij} can be represented as the product of the jth SNP and the variable indicating membership of the ith read [74]. Formally, this can be written as

$$r_{ij} = c_i \cdot h_j. \tag{7.2}$$

From (7.2), matrix $\boldsymbol{R}$ could be interpreted as being obtained from a rank 1 matrix $\boldsymbol{S}$ whose row $\boldsymbol{s}_i$ is either $\boldsymbol{h}$ or $-\boldsymbol{h}$ based on the value of c_i, while most of its entries are erased in the reading process. In particular, we have

$$\boldsymbol{R} = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{S}), \text{ and } \boldsymbol{S} = \boldsymbol{c}^T \cdot \boldsymbol{h}, \tag{7.3}$$

where $\boldsymbol{\Omega}$ is the collection of all observed locations, and the projection $\mathcal{P}$ is defined by

$$\mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{S})_{ij} = \begin{cases} s_{ij}, & \text{if } (i,j) \in \boldsymbol{\Omega}, \\ \times, & \text{if } (i,j) \notin \boldsymbol{\Omega}. \end{cases} \tag{7.4}$$

Hence, the task of haplotype assembly is to recover haplotype $\boldsymbol{h}$ and chromosome membership vector $\boldsymbol{c}$, or, equivalently, to find matrix $\boldsymbol{S}$ from matrix $\boldsymbol{R}$.

An example, illustrated by Figure 7.2, corresponds to the scenario where 6 SNP sites are covered by 8 paired-end reads. The first 4 reads are assumed to be (shotgun) sequenced from chromosome 1 and thus the chromosome membership vector is $\boldsymbol{c} = (+1, +1, +1, +1, -1, -1, -1, -1)$. The true haplotype associated with chromosome 1 is assumed to be $\boldsymbol{h} = (+1, +1, -1, +1, -1, -1)$. In the absence of errors, the acquired SNP fragment matrix is given by

$$\boldsymbol{R} = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{c}^T \cdot \boldsymbol{h}) = \begin{bmatrix} \times & \times & -1 & \times & -1 & \times \\ \times & +1 & \times & \times & -1 & \times \\ +1 & \times & \times & +1 & \times & \times \\ \times & \times & -1 & +1 & \times & \times \\ -1 & \times & +1 & \times & \times & \times \\ \times & -1 & \times & \times & +1 & \times \\ -1 & \times & \times & -1 & \times & \times \\ \times & \times & \times & \times & +1 & +1 \end{bmatrix}. \tag{7.5}$$

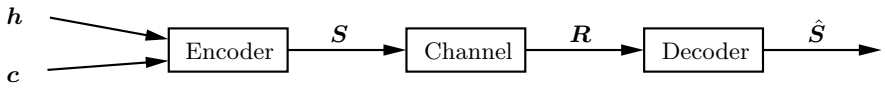

Figure 7.3: **Information theoretic model for the haplotype assembly problem.** Two messages, haplotype and membership vector, are passing through an erasure channel, characterizing the paired-end reading process.

7.3 Error-free Case

We first analyze haplotype assembly in the ideal scenario where the information provided by the sequencing reads is error-free. From a joint source-channel coding perspective, haplotype assembly aims to recover two sources being communicated through an erasure channel (see Figure 7.3). The first source is haplotype information, $\boldsymbol{h}$, and the second source is the chromosome membership vector $\boldsymbol{c}$. Both of these vectors are assumed to originate from a uniform distribution, i.e., their entries have $1/2$ probability to take values from $\{+1, -1\}$. These two sources are encoded jointly using the function $\psi : \{+1, -1\}^n \times \{+1, -1\}^m \to \{+1, -1\}^{m \times n}$, and hence the encoded codeword $\boldsymbol{S} = \psi(\boldsymbol{h}, \boldsymbol{c})$. In particular, each entry in $\boldsymbol{S}$ is given by $s_{ij} = c_i \cdot h_j$, which implies the encoder is a bijection.

After receiving the output from channel, $\boldsymbol{R}$, the decoder uses the decoding function to map its observations into an estimate of the message. Specifically, we consider the decoder (i.e., an algorithm for haplotype assembly) given by $\varphi : \{+1, -1, \times\}^{m \times n} \to \{+1, -1\}^{m \times n}$, such that $\hat{\boldsymbol{S}} = \varphi(\boldsymbol{R})$, where $\hat{\boldsymbol{S}}$ represents the estimate. Note that since the encoding function is a bijection, decoding $\boldsymbol{S}$ is equivalent to decoding both $\boldsymbol{h}$ and $\boldsymbol{c}$. We define the error

probability of decoding as

$$P_{\mathrm{e}} \triangleq \Pr\{\hat{\boldsymbol{S}} \neq \boldsymbol{S}|\boldsymbol{R}\}. \tag{7.6}$$

As in the conventional information-theoretic analysis of a communication channel, we consider all possible choices of matrix $\boldsymbol{S}$ and denote the resulting ensemble by $\mathcal{S}$. Let m and n be sufficiently large so that there exists at least one decoding function φ with small probability of error. The channel model reflects particular reading technique. For the paired-end sequencing technique without sampling errors, let us consider the channel $\mathcal{W} : \{+1, -1\}^{m\times n} \to \{+1, -1, \times\}^{m\times n}$ described as follows:

1) Erasures happen independently across rows.

2) In each row, only 2 entries remain and their positions are assumed to be uniformly placed. This can be easily extended to any number of (constant) entries within each row.

3) Unerased entries are observed correctly.

In other words, for the sake of simplicity we assume that precisely 2 entries are observed in each row of $\boldsymbol{S}$, and that the observations are correct and independent across different rows. Under these assumptions, the number of numerical entries in each column of $\boldsymbol{R}$ approximately obeys Poisson distribution. Moreover, the expected length of insert size between 2 sampled entries within a row is given by $(n-2)/3$. In practice, the insert size is limited and cannot be

made arbitrarily large – a constraint that we relax in our analysis by making the assumption 2) above.

Based on this model, we derive the necessary and sufficient conditions on the number of error-free reads needed for haplotype assembly.

Theorem 7.1. *Given the SNP fragment matrix* $\boldsymbol{R}$ *with* 2 *reliable observations at arbitrary positions in each row, the original haplotype matrix* $\boldsymbol{S}$ *can be reconstructed only if the number of reads satisfies*

$$m = \Omega(n),$$

where n *is the length of the target haplotype. Moreover, if* $m = \Theta(n \ln n)$*, a reconstruction algorithm, erasure decoding, could determine* $\boldsymbol{S}$ *accurately with high probability. Specifically, given a target small constant* $\epsilon > 0$*, there exists* n *large enough such that by choosing* $m = \Theta(n \ln n)$ *the probability of error* $P_e \leq \epsilon$*.*

We provide the proofs of necessary and sufficient conditions in the following two subsections.

7.3.1 Necessary Condition for Recovery

Using Fano's inequality [17], we find that

$$H(\boldsymbol{S}|\boldsymbol{R}) \leq P_{\mathrm{e}} \log |\mathcal{S}| \leq P_{\mathrm{e}}(m+n), \tag{7.7}$$

where the set of all possible $\boldsymbol{S}$, $\mathcal{S}$, has cardinality upper bounded by 2^{m+n}. Recall that $\boldsymbol{\Omega}$ specifies random locations where $\boldsymbol{S}$ is observed (i.e., sampled).

Note that $\boldsymbol{\Omega}$ is independent of $\boldsymbol{S}$ and that its rows are independent due to our assumption on the nature of the channel. The following simple steps provide a bound:

$$\begin{aligned}
H(\boldsymbol{S}) &\overset{(a)}{=} H(\boldsymbol{S}|\boldsymbol{\Omega}) \\
&= I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{\Omega}) + H(\boldsymbol{S}|\boldsymbol{\Omega},\boldsymbol{R}) \\
&= I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{\Omega}) + H(\boldsymbol{S}|\boldsymbol{R}) \\
&\overset{(b)}{\leq} I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{\Omega}) + P_{\mathrm{e}}(m+n) \\
&= H(\boldsymbol{R}|\boldsymbol{\Omega}) - H(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{\Omega}) + P_{\mathrm{e}}(m+n) \\
&\overset{(c)}{=} H(\boldsymbol{R}|\boldsymbol{\Omega}) + P_{\mathrm{e}}(m+n) \\
&\leq \sum_{i=1}^{m} H(\boldsymbol{r}_i|\boldsymbol{\omega}_i) + P_{\mathrm{e}}(m+n) \\
&\overset{(d)}{=} 2m + P_{\mathrm{e}}(m+n),
\end{aligned}$$

where

(a) follows from independence between $\boldsymbol{S}$ and $\boldsymbol{\Omega}$;

(b) from Fano's inequality, i.e., equation (7.7);

(c) from the fact that $\boldsymbol{R}$ is deterministic if $\boldsymbol{S}$ and $\boldsymbol{\Omega}$ are both known in the error-free case;

(d) from the assumption that every row has exactly 2 entries observed.

Finally, by noting that $H(\boldsymbol{S}) = m + n$, we clearly need

$$m \geq \frac{(1 - P_{\mathrm{e}})n}{1 + P_{\mathrm{e}}} \tag{7.8}$$

for accurate recovery. More precisely, we need $m = \Omega(n)$ for recovery with arbitrarily small probability of decoding error.

Remark 7.2. *Note that, in this proof, the channel model is only utilized when bounding* $H(\boldsymbol{R}|\boldsymbol{\Omega})$*. In fact, the necessary result is extendable to other channel models (i.e., reading techniques). In particular, the lower bound* $m = \Omega(n)$ *also holds in the case of deterministic choice of reading sites, paired-end reading with fixed insert size, and, more importantly, reading techniques with more than* 2 *observations in each read. The essential condition for the establishment of necessary condition is to ensure the number of observed entries in the matrix is* $\Theta(m)$*.*

7.3.2 Sufficient Condition for Recovery

The goal of a decoding algorithm is to recover $\boldsymbol{S}$ (or equivalently $\boldsymbol{h}$ and $\boldsymbol{c}$) from $\boldsymbol{R}$ with high confidence. Here, we show a simple and effective algorithm, called "erasure decoding", which requires only $\Theta(n \ln n)$ reads for reliable haplotype recovery. Detailed steps of this algorithm are described as follows:

1) Choose the "seed" s as an arbitrary non-erased entry in the first row, i.e., $s = r_{1j}$, where j is randomly chosen such that $r_{1j} \neq \times$. Set the chromosome membership variable of the first row to $c_1 = +1$.

2) Find all other rows with position j not erased, i.e., form a set

$$\mathcal{A} = \{k|\ r_{kj} \neq \times,\ k \neq 1\}. \tag{7.9}$$

3) Set the chromosome membership variables of the rows with indices in $\mathcal{A}$ to

$$c_k = \begin{cases} +1, & \text{if } r_{kj} = r_{1j}, \\ -1, & \text{otherwise,} \end{cases} \tag{7.10}$$

for every $k \in \mathcal{A}$.

4) Decode SNPs in the first row by evaluating

$$r_{1l} = c_k \cdot r_{kl}, \tag{7.11}$$

for every $k \in \mathcal{A}$ and $r_{kl} \neq \times$.

5) Delete all rows with indices in $\mathcal{A}$.

6) Arbitrarily choose another non-erased entry in the first row as the new seed $s = r_{1j}$ which has not been chosen as a seed in any of the previous steps. Repeat Step 2) to 6) until no row could be further erased.

7) If the first row is the only remaining one and its entries are all decoded, declare $\boldsymbol{h} = \boldsymbol{r}_1$; otherwise, declare a failure.

Remark 7.3. *In the previous algorithm, we arbitrarily set a chromosome membership variable of the first row, which may lead to incorrect association of the corresponding read with a haplotype. In fact, if the algorithm successfully decodes both* $\boldsymbol{h}$ *and* $\boldsymbol{c}$*, then all their components may be flipped due to an incorrect choice of the initial chromosome membership variable. However, matrix* $\boldsymbol{S}$ *would still be reconstructed correctly due to the particular product operation used to generate components of* $\boldsymbol{S}$*. Therefore, the choice of initial membership does not influence the decoding performance.*

Remark 7.4. *Erasure decoding is closely connected to the bipartite partitioning interpretation of the haplotype assembly problem [69]. Note that if our algorithm successfully recovers the message matrix* $\boldsymbol{S}$*, we can realign its rows such that the matrix could be partitioned into two sub-matrices with different chromosome memberships. Therefore, in the error-free case, the erasure decoding provides a computationally efficient method for partitioning reads into two sets.*

Figure 7.4 shows the details of the decoding procedure for the example illustrated in Figure 7.2, where the read matrix is given by (7.5).

Below we analyze the performance of the proposed algorithm. More precisely, we show that if the number of reads is large enough, i.e., $m = \Theta(n \ln n)$, the source matrix $\boldsymbol{S}$ can be recovered correctly with high probability. Observe that, in the absence of sampling errors, the erasure decoding algorithm ensures the output to be the correct haplotype if both of the following conditions are satisfied:

1. all rows except for the first one are deleted, and

2. all entries in the first row are decoded.

At this point, decoding error occurs if at least one of the following events happen:

1. Event E_1: at least one of the columns in $\boldsymbol{R}$ is erased and thus the corresponding SNP could not be decoded;

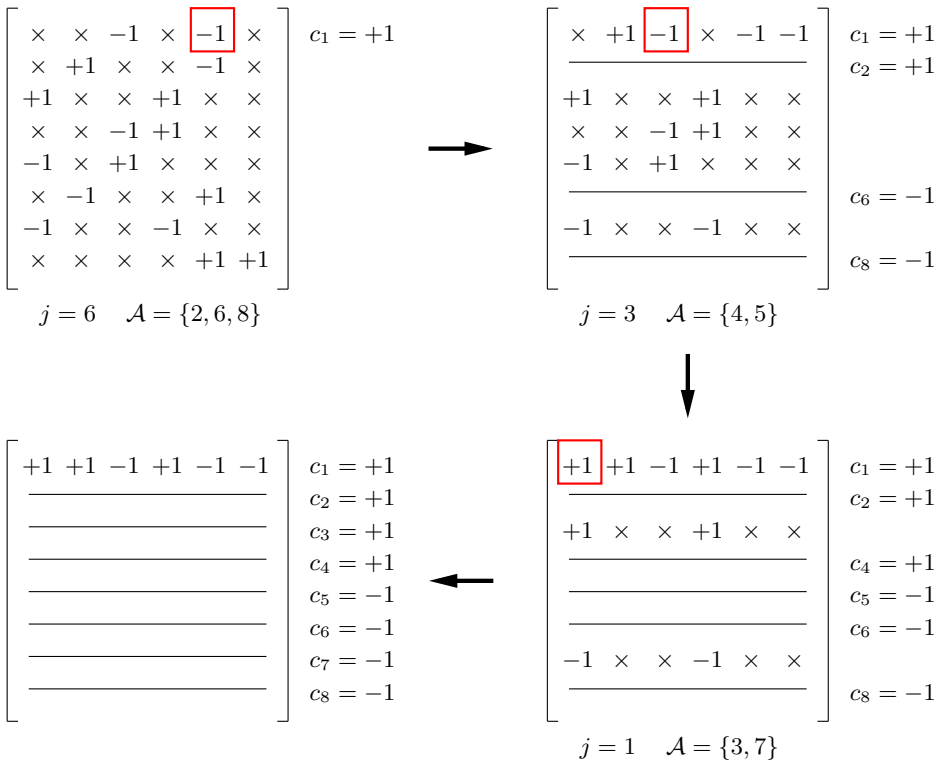

Figure 7.4: **Erasure decoding of the example illustrated in Figure 7.2.** In every round, the seed is marked in a rectangle, with its column index given by j. Rows that share the same positions as the seed are collected in the set $\mathcal{A}$. A straight line crossing a whole row of the matrix represents a deletion.

2. Event E_2: there exist a partition of row indices $\{1, \ldots, m\} = \mathcal{U}_1 \cup \mathcal{U}_2$ and a partition of column indices $\{1, \ldots, n\} = \mathcal{V}_1 \cup \mathcal{V}_2$ such that $|\mathcal{V}_1| \geq 2$ and $|\mathcal{V}_2| \geq 2$ (so that 2 entries could be sampled from each row), and $r_{ij} = \times$ for any $(i, j) \in (\mathcal{U}_1 \times \mathcal{V}_2) \cup (\mathcal{U}_2 \times \mathcal{V}_1)$. In other words, the sampled entries could be considered as originating from two disjoint subsets of target haplotypes and thus there is no hope for assembly due to the lack of information bridging these subsets.

We outline how to bound the probability of each of the two error events. First, note that by the coupon collector effect, if $m = \Theta(n \ln n)$ then every column is covered by at least one read with high probability. More precisely, by taking $m = n \ln n$, the error event (or, equivalently, the tail distribution for the coupon collector problem) is given by

$$\begin{aligned}
\Pr\{E_1\} &= \frac{\sum_{i=1}^{n-2} \binom{n}{i} \binom{n-i}{2}^m}{\binom{n}{2}^m} \\
&= \sum_{i=1}^{n-2} \binom{n}{i} \left[\frac{(n-i)(n-i-1)}{n(n-1)} \right]^m \\
&\leq \sum_{i=1}^{n-2} n^i e^{-m \frac{2in - i(i+1)}{n(n-1)}} \\
&= \sum_{i=1}^{n-2} O(n^{-i}) \\
&= O(n^{-1}).
\end{aligned} \tag{7.12}$$

On the other hand, the second error event E_2 could be further decomposed into sub-events $E_2^{u,v}$ which represent the type 2 error event with particular $u = |\mathcal{U}_1|$ and $v = |\mathcal{V}_1|$. Then, we have

$$\Pr\{E_2^{u,v}\} = \frac{\binom{n}{v} \binom{m}{u} \binom{v}{2}^u \binom{n-v}{2}^{m-u}}{\binom{n}{2}^m}. \tag{7.13}$$

Observe that by symmetry and monotonicity, the right hand side in (7.13) is maximized by two extreme points on the feasible $(u, v)-$region, i.e., for any u and v, $\Pr\{E_2^{u,v}\} \leq \Pr\{E_2^{1,2}\} = \Pr\{E_2^{m-1,n-2}\}$. In particular, we have

$$\Pr\{E_2^{1,2}\} = \frac{\binom{n}{2} \binom{m}{1} \binom{2}{2}^1 \binom{n-2}{2}^{m-1}}{\binom{n}{2}^m}$$

$$
\begin{aligned}
&= \frac{m[(n-2)(n-3)]^{m-1}}{[n(n-1)]^{m-1}} \\
&\leq n \ln n \left(1 - \frac{4n-6}{n(n-1)}\right)^{n \ln n - 1} \\
&\leq n \ln n e^{-\frac{4n-6}{n(n-1)}(n \ln n - 1)} \\
&= O(n^{-3} \ln n).
\end{aligned}
$$

Hence, the probability of the second error event is upper bounded by

$$
\begin{aligned}
\Pr\{E_2\} &= \sum_{u=1}^{m-1} \sum_{v=2}^{n-2} \Pr\{E_2^{u,v}\} \\
&\leq (m-2)(n-4) \Pr\{E_2^{1,2}\} \\
&\leq n^2 \ln n O(n^{-3} \ln n) \\
&= O(n^{-1} (\ln n)^2). \qquad (7.14)
\end{aligned}
$$

Combining these two bounds together, we obtain

$$
P_{\mathrm{e}} \leq \Pr\{E_1\} + \Pr\{E_2\} = O(n^{-1}) + O(n^{-1}(\ln n)^2) < \epsilon,
$$

for arbitrary $\epsilon > 0$ with sufficiently large n.

Remark 7.5. *Note that there is a log-factor gap between the lower and upper bounds. As analyzed in [75], this log-factor generally exists and reflects the need that sufficiently many entries should be sampled to facilitate accurate recovery. If a more systematic reading method, rather than random sampling, could be adopted to generate the observation matrix, the log-factor may not be essential for reconstruction. We will see in the next section that this log-factor gap between two bounds also exists for the erroneous case.*

7.4 Erroneous Case

When determining a component of the haplotype sequence at a particular position, we essentially need to perform a hypothesis test and decide between possible symbols in the corresponding column of the SNP fragment matrix. If sequencing errors are present, some of the entries in $\boldsymbol{R}$ are erroneously flipped. For the purpose of the following discussion, we assume such errors are independent and identically distributed (i.i.d.). More precisely, the errors are modeled as having originated by passing messages (i.e., the numerical entries in $\boldsymbol{R}$) through a collection of independent binary symmetric channels characterized by the parameter p, the probability of flipping the sign of a numerical entries of $\boldsymbol{R}$. Denoting the noise as matrix $\boldsymbol{N}$ with entries n_{ij} that are i.i.d., we can write

$$\boldsymbol{R} = \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{S} \oplus \boldsymbol{N}). \tag{7.15}$$

Hence, the model describing the erroneous case is as same as the one for the error-free case except for an additional noise term capturing the effects of "channel" (i.e., the effects of sequencing and data processing steps that precede haplotype assembly). The equivalent channel model $\mathcal{W} : \{+1, -1\}^{m \times n} \rightarrow \{+1, -1, \times\}^{m \times n}$ considered in this section is described as follows:

1) Erasures happen independently across rows.

2) In each row, only 2 entries remain and their positions are uniformly random.

3) The remaining entries are read incorrectly with probability p and the errors are independent.

We would like to reconstruct $\boldsymbol{S}$ from $\boldsymbol{R}$ with high probability. However, if no more than two numerical entries are observed in a row, solving this problem is not always feasible. Assume, for instance, that the observed numerical entries in $\boldsymbol{r}_i$ are $(+1, +1)$, and that only one sequencing error happened (i.e., either one of $\boldsymbol{r}_i$'s numerical entries is erroneous). Then, there is no hope to discover whether the true numerical entries in $\boldsymbol{s}_i$ are $(-1, +1)$ or $(+1, -1)$. For this reason, in the erroneous case we aim to recover (with high probability) only the row space, i.e., find the haplotype $\boldsymbol{h}$ from matrix $\boldsymbol{R}$. Let us denote the haplotype estimate found by an assembly algorithm by $\hat{\boldsymbol{h}}$. We define the probability of error as

$$P_{\mathrm{e}} = \Pr\{\hat{\boldsymbol{h}} \neq \boldsymbol{h} | \boldsymbol{R}\},$$

and use it to characterize the accuracy of assembly. We would like to make this probability arbitrarily small on average (averaged over all possible $\boldsymbol{h}$).

Based on the previously described model of the haplotype assembly problem, we next state the necessary and sufficient conditions on the number of reads required for assembly.

Theorem 7.6. *Given the SNP fragment matrix $\boldsymbol{R}$ with 2 unreliable observations at arbitrary positions in each row, the original haplotype vector $\boldsymbol{h}$ can be reconstructed only if the number of reads satisfies*

$$m = \Omega(n),$$

where n denotes the length of the target haplotype. Moreover, if $m = \Theta(n \ln n)$, a reconstruction algorithm, spectral partitioning, can determine $\boldsymbol{h}$ accurately with high probability. Specifically, given a target small constant $\epsilon > 0$, there exists n large enough such that by choosing $m = \Theta(n \ln n)$ the probability of error $P_e \leq \epsilon$.

The preceding theorem shows that although observations are not reliable due to sampling noise, the number of reads required for assembly still scales linearly with n. We provide the proofs of necessary and sufficient conditions in the following two subsections.

7.4.1 Necessary Condition for Recovery

From Fano's inequality,

$$H(\boldsymbol{h}|\boldsymbol{R}) \leq P_{\mathrm{e}} \cdot n.$$

Therefore, we can write

$$H(\boldsymbol{h}) \leq H(\boldsymbol{R}|\boldsymbol{\Omega}) - H(\boldsymbol{R}|\boldsymbol{h}, \boldsymbol{\Omega}) + P_{\mathrm{e}} \cdot n.$$

Unlike the error-free scenario analyzed in Section III.A, here $H(\boldsymbol{R}|\boldsymbol{h}, \boldsymbol{\Omega})$ does not vanish. In particular, since the noise is i.i.d., it holds that

$$H(\boldsymbol{R}|\boldsymbol{h}, \boldsymbol{\Omega}) \geq H(\boldsymbol{R}|\boldsymbol{S}, \boldsymbol{\Omega}) = \sum_{i=1}^{m} H(\boldsymbol{r}_i|\boldsymbol{s}_i, \boldsymbol{t}_i) = 2mH(p).$$

Since $H(\boldsymbol{R}|\boldsymbol{\Omega}) \leq 2m$ and $H(\boldsymbol{h}) = n$, we have

$$m \geq \frac{(1 - P_{\mathrm{e}})n}{2[1 - H(p)]}, \tag{7.16}$$

and thus the necessary number of reads is of the same order as in the error-free case, $m = \Omega(n)$.

7.4.2 Sufficient Condition for Recovery

Recall that, for the scenario where $\boldsymbol{R}$ is error-free, in Section III.B we proposed and analyzed the erasure decoding algorithm for the recovery of $\boldsymbol{S}$ (or, equivalently, $\boldsymbol{h}$ and $\boldsymbol{c}$). However, if the entries of $\boldsymbol{R}$ are potentially erroneous, erasure decoding may fail to find the correct solution. Effective methods for haplotype assembly from erroneous short reads are actively pursued in research community. Most state-of-the-art algorithms rely on graphical interpretation of the problem and consider optimization formulations focusing on different objective criteria [72].

Formulations of the haplotype assembly problem include minimum fragment removal (MFR), minimum SNP removal (MSR), and minimum error correction (MEC). MFR [72] formulation aims to identify the smallest number of fragments (i.e., reads) whose removal renders the graph representing the assembly problem bipartite. Since the resulting graph is conflict-free, algorithms for error-free case could be readily applied to assemble the haplotypes. However, solving the MFR formulation of the assembly problem is challenging since the resulting optimization is generally non-convex. MSR [72] is an alternative formulation focused on identifying the smallest possible number of SNP sites such that the graph representing remaining SNPs could be partitioned in two subgraphs corresponding to haplotypes. In graph-theoretic terms, MSR

aims to find the maximum independent set of the original graph. MEC [73] formulation seeks the smallest number of entries in matrix $\boldsymbol{R}$ whose flipping ensures that rows in $\boldsymbol{R}$ are consistent with having originated from two complementary haplotypes. In this formulation, the problem becomes the one of error-correction of binary data corrupted by i.i.d. noise. MEC is the most widely used formulation of the haplotype assembly problem, and a large number of algorithms have been developed for solving it (perhaps the most widely used one is HapCUT [76]).

The existing work on haplotype assembly focuses on the development of algorithms that treat the number of reads as a known parameter and do not explore the fundamental requirements for the assembly. In contrast, we rely on the information-theoretic framework to investigate the sufficient conditions on the number of reads needed for near-perfect recovery of the haplotype sequence. To this end, we present a low-rank matrix analysis formulation of the haplotype assembly problem. Intuitively, we aim to partition SNP sites into two sets, each corresponding to one of the two haplotypes in a homologous pair. By regarding the adjacency matrix of the original graphical representation of the problem as a perturbation of a planted model (which is inherently a low rank matrix), we claim that the partition is perfect as long as the parameters of the model are chosen appropriately. In what follows, we first describe the "spectral partitioning" algorithm that relies on the singular value decomposition (SVD) technique to obtain a weaker conclusion that the fraction of partition errors vanishes as n increases, and then propose a modified algorithm

for near-perfect haplotype recovery. The steps of the spectral partitioning algorithms are as follows:

1) Construct an adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n\times n}$ based on the observation matrix $\boldsymbol{R}$, such that for every $(u,v) \in \{1,\ldots,n\} \times \{1,\ldots,n\}$ with $u > v$,

$$a_{uv} = \begin{cases} 1, & \text{if } \sum_{i=1}^{m} \mathbf{1}_{\{r_{iu}\neq\times, r_{iv}\neq\times, r_{iu}=r_{iv}\}} > \sum_{i=1}^{m} \mathbf{1}_{\{r_{iu}\neq\times, r_{iv}\neq\times, r_{iu}\neq r_{iv}\}}, \\ 0, & \text{otherwise.} \end{cases} \tag{7.17}$$

Then, let $a_{uv} = a_{vu}$ for any $u > v$ to guarantee symmetry, and let $a_{uu} = 0$ to enforce zeros on the diagonal of $\boldsymbol{A}$.

2) Find the singular value decomposition (SVD) of $\boldsymbol{A}$, i.e., $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}$ such that $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times n}$ are unitary matrices and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ is diagonal.

3) Identify the eigenvector $\boldsymbol{v}_2(\boldsymbol{A})$ corresponding to the second largest eigenvalue of $\boldsymbol{A}$ and construct sets

$$\mathcal{C}_1 = \{j : v_{2j} < 0\}, \quad \mathcal{C}_2 = \{j : v_{2j} \geq 0\}.$$

The haplotype is then recovered as

$$h_j = \begin{cases} +1, & \text{if } j \in \mathcal{C}_1, \\ -1, & \text{if } j \in \mathcal{C}_2. \end{cases}$$

Remark 7.7. *As evident from* (7.17)*, elements of* $\boldsymbol{A}$ *are evaluated by examining all SNP position pairs and performing the majority voting operation over read components that cover the same SNP position pair. This procedure is*

equivalent to the MAP hypothesis testing that assumes uniform SNP prior distribution. If the distribution of SNPs is not uniform, or if error distributions are not identical across SNP sites, one should rely on weighted majority voting instead.

We analyze the performance of spectral partitioning by showing its relation to the classical partitioning problem on a planted model. This approach originates from the perturbation theory for eigenvectors and follows steps similar to those in [77], but additionally exploits structural features of the haplotype assembly problem to arrive at bounds that are much tighter than those provided in the general case.

7.4.2.1 Planted Model

Consider the planted model, i.e., a matrix $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ defined as

$$\boldsymbol{B} = \begin{bmatrix} [\alpha]_{n_1 \times n_1} & [\beta]_{n_1 \times n_2} \\ [\beta]_{n_2 \times n_1} & [\alpha]_{n_2 \times n_2} \end{bmatrix},$$

where $\alpha > \beta > 0$, $n_1 + n_2 = n$, and $[\alpha]_{n_1 \times n_1}$ denotes an $n_1 \times n_1$ sub-matrix with all entries equal to α. Clearly, such a matrix $\boldsymbol{B}$ is low-rank. More precisely, if we perform the SVD on $\boldsymbol{B}$, it becomes evident that the rank of $\boldsymbol{B}$ is 2 and that its first two singular values and the corresponding singular vectors are given by

$$\lambda_1(\boldsymbol{B}) = n_1\beta\mu_1 + n_2\alpha \ , \tag{7.18}$$

$$\lambda_2(\boldsymbol{B}) = n_1\beta\mu_2 + n_2\alpha \ , \tag{7.19}$$

$$\boldsymbol{v}_1(\boldsymbol{B}) = \left(\left[\frac{\mu_1}{\sqrt{n_1\mu_1^2 + n_2}} \right]_{1\times n_1}, \left[\frac{1}{\sqrt{n_1\mu_1^2 + n_2}} \right]_{1\times n_2} \right), \tag{7.20}$$

$$\boldsymbol{v}_2(\boldsymbol{B}) = \left(\left[\frac{\mu_2}{\sqrt{n_1\mu_2^2 + n_2}} \right]_{1\times n_1}, \left[\frac{1}{\sqrt{n_1\mu_2^2 + n_2}} \right]_{1\times n_2} \right), \tag{7.21}$$

where

$$\mu_1 = \frac{(n_1 - n_2)\alpha + \sqrt{(n_1 - n_2)^2\alpha^2 + 4n_1n_2\beta^2}}{2n_1\beta}, \tag{7.22}$$

$$\mu_2 = \frac{(n_1 - n_2)\alpha - \sqrt{(n_1 - n_2)^2\alpha^2 + 4n_1n_2\beta^2}}{2n_1\beta}. \tag{7.23}$$

Note that since $\mu_1 > 0$ and $\mu_2 < 0$ for any n_1 and n_2, it holds that $\lambda_1(\boldsymbol{B}) > \lambda_2(\boldsymbol{B})$. Moreover, since $\mu_2 < 0$, the first n_1 entries in $\boldsymbol{v}_2(\boldsymbol{B})$ have opposite signs from those of the last n_2 entries. Therefore, if we partition the indices into two sets with respect to their signs in $\boldsymbol{v}_2(\boldsymbol{B})$, the result naturally provides a classification corresponding to different blocks of matrix $\boldsymbol{B}$.

7.4.2.2 Generated Adjacency Matrix

As discussed above, eigenvector corresponding to the second largest eigenvalue of the planted model $\boldsymbol{B}$ enables partitioning, i.e., helps distinguish between different block indices. The next step is to relate the planted model $\boldsymbol{B}$ to the adjacency matrix $\boldsymbol{A}$ constructed according to (7.17). Note that the entries in the upper-triangular part of $\boldsymbol{A}$ are random and independent. In fact, the distribution of each entry in $\boldsymbol{A}$ is Bernoulli with parameters which only depend on whether the corresponding SNP sites belong to the same block or not (i.e., two parameters are sufficient to characterize the distribution of $\boldsymbol{A}$).

$\boldsymbol{A}$ and $\boldsymbol{B}$ are related through a series of permutations of rows and columns (note that permutations do not impact the eigenvectors). In particular, for any $(u, v) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ with $u > v$, we define

$$\Pr\{a_{uv} = 1\} = \pi(b_{uv}),$$
$$\Pr\{a_{uv} = 0\} = 1 - \pi(b_{uv}),$$

where π is the permutation of rows and columns. Let α denote the probability that two SNP sites from the same cluster are inferred correctly in the majority voting step, while β denotes the probability that two SNP sites from different clusters are inferred incorrectly. Clearly, α and β are closely related to the accuracy and redundancy in the sequencing data – more precisely, the parameters n, m, and p. In our case of unreliable paired-end sequencing, the probabilities α and β are given by

$$\begin{aligned}
\alpha &\triangleq \Pr\{\text{majority voting claims } a_{uv} = 1 |\ h_u = h_v\} \\
&= \sum_{i=1}^{m} \Pr\{\text{claims } a_{uv} = 1,\ i \text{ reads cover sites } u \text{ and } v |\ h_u = h_v\} \\
&= \sum_{i=1}^{m} \left\{ \binom{m}{i} \gamma^i (1-\gamma)^{m-i} \sum_{l=\lfloor i/2 \rfloor + 1}^{i} \binom{i}{l} [(1-p)^2 + p^2]^l [2p(1-p)]^{i-l} \right\},
\end{aligned}$$

where $\gamma \triangleq 2/n(n-1)$ is the probability that a read covers target SNP sites u and v; $(1-p)^2 + p^2$ is the probability that a read covers SNPs that are identical given $h_u = h_v$; and the second summation (ranging from $\lfloor i/2 \rfloor + 1$ to i) represents for the majority voting operation evaluated over i voters. Similarly, we have

$$\beta \triangleq \Pr\{\text{majority voting claims } a_{uv} = 1 |\ h_u \neq h_v\}$$

$$= \sum_{i=1}^{m} \left\{ \binom{m}{i} \gamma^i (1-\gamma)^{m-i} \sum_{l=\lfloor i/2 \rfloor + 1}^{i} \binom{i}{l} [2p(1-p)]^l [(1-p)^2 + p^2]^{i-l} \right\},$$

where $2p(1-p)$ is the probability that a particular read covers SNPs that are identical given $h_u \neq h_v$. Since neither α nor β is straightforward to compute, we seek more compact and manageable bounds on these probabilities that will enable analysis of the worst-case scenarios.

Lemma 7.8. *When the number of reads used to assemble a long haplotype of length n scales as $m = \Theta(n \ln n)$, there exist positive constants κ_1, κ_2, and κ_3, such that*

$$\alpha \geq \frac{2\kappa_1 \kappa_2 [(1-p)^2 + p^2] \ln n}{n-1}, \tag{7.24}$$

$$\beta \leq \frac{2\kappa_1 [2p(1-p)] \ln n}{(n-1)(1-\kappa_3^{-1})}, \tag{7.25}$$

where $\kappa_2 < 1$ and $\kappa_3 > 1$.

The lemma shows that both α and β have bounds which scale as $\Theta(n^{-1} \ln n)$ (for the proof, please see Appendix 7.A). Using these bounds, we next show that the signs of the corresponding entries of the eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$ are identical with high probability.

7.4.2.3 Matrix Eigenvector Perturbation

After establishing the relationship between the adjacency matrix $\boldsymbol{A}$ and the planted model $\boldsymbol{B}$, we proceed to explore the difference between their eigenvectors by relying on the matrix perturbation theory. In particular, we

show that for our choices of α and β, perturbation of the eigenvector of $\boldsymbol{A}$ associated with the second largest eigenvalue from the corresponding eigenvector of $\boldsymbol{B}$ (i.e., the difference between those two eigenvectors) vanishes as n increases. This result justifies performing spectral partitioning on $\boldsymbol{A}$, rather than $\boldsymbol{B}$, without a significant loss of performance.

The matrix perturbation theory allows one to determine sensitivity of matrix eigenvalues and eigenvectors with respect to slight perturbations. This area was pioneered in [78] where a general bound for the matrix eigenvalue perturbation effects was provided. More recently, [79] improved this bound under further assumptions on the matrix structure. Meanwhile, the famous Davis-Kahan sin-theta theorem [80] characterizes the rotation of eigenvectors after perturbation, and [81] focuses on random matrices to propose a probabilistic sin-theta theorem. Note that the observed matrices in the haplotype assembly problem are always characterized by a particular structures, for instance, independent and binary distributed entries, low rank, etc. To exploit the special structure, we follow the result from a recent perturbation study [82] which provides a much tighter bound for the perturbation effects with respect to binary random matrices, summarized in the following lemma.

Lemma 7.9 (Lemma 2 and 3 in [82])**.** *Consider a square $n \times n$ symmetric 0-diagonal random matrix $\boldsymbol{M}$ such that its elements $m_{uv} = m_{vu}$ are independent Bernoulli random variables with parameters $\mathbb{E}[m_{uv}] = \rho_{uv}\chi n^{-1}$, where ρ_{uv} are constants and $\chi = \Omega(\ln n)$. Then, with probability at least $1 - O(n^{-1})$, we have*

$$|\lambda_k(\boldsymbol{M}) - \lambda_k(\mathbb{E}[\boldsymbol{M}])| \leq O(\chi^{1/2}), \tag{7.26}$$

$$||\boldsymbol{v}_k(\boldsymbol{M}) - \boldsymbol{v}_k(\mathbb{E}[\boldsymbol{M}])|| \leq O(\chi^{-1/2}), \tag{7.27}$$

for any k not larger than the rank of $\mathbb{E}[\boldsymbol{M}]$, where $\lambda_k(\boldsymbol{M})$ is the k-th largest eigenvalue of $\boldsymbol{M}$, and $\boldsymbol{v}_k(\boldsymbol{M})$ is the corresponding k-th eigenvector.

We observe that the adjacency matrix $\boldsymbol{A}$ has the same structure as the matrix $\boldsymbol{M}$ in the statement of the lemma. In particular, note that $\boldsymbol{A}$ is a 0-diagonal random matrix with each entry being an independently distributed Bernoulli random variable. The parameters of the Bernoulli distributions, α and β, satisfy the scale constraints with $\chi = \ln n$ due to Lemma 7.8. Moreover, note that $\mathbb{E}[\boldsymbol{A}] = \pi(\tilde{\boldsymbol{B}})$, where $\tilde{\boldsymbol{B}} = \boldsymbol{B} - \alpha\boldsymbol{I}$, and that permutation π does not change the eigenvectors. Therefore, we can utilize Lemma 7.9 to study the haplotype assembly problem. In particular, from (7.27) it follows that

$$||\boldsymbol{v}_2(\boldsymbol{A}) - \boldsymbol{v}_2(\tilde{\boldsymbol{B}})|| \leq O(\ln^{-1/2} n).$$

By noting that an addition of the identity matrix does not influence the eigenvectors, we conclude that $\boldsymbol{v}_2(\tilde{\boldsymbol{B}}) = \boldsymbol{v}_2(\boldsymbol{B})$. Thus, we obtain

$$||\boldsymbol{v}_2(\boldsymbol{A}) - \boldsymbol{v}_2(\boldsymbol{B})|| \leq O(\ln^{-1/2} n). \tag{7.28}$$

Recall that $\boldsymbol{v}_2(\boldsymbol{B})$ has the form of (7.21), which implies that a particular entry perturbed to change its sign contributes at least $\Omega(n^{-1/2})$ to $||\boldsymbol{v}_2(\boldsymbol{A}) - \boldsymbol{v}_2(\boldsymbol{B})||$. Therefore, if n_e is the number of errors, we have

$$\sqrt{\frac{n_\mathrm{e}}{n}} \leq O(\ln^{-1/2} n). \tag{7.29}$$

By noting that n_e/n is the fraction of partition errors, we conclude that the haplotype can be recovered reliably with vanishing fraction of errors for sufficiently large number of reads n.

Remark 7.10. *As indicated by the analysis, spectral partitioning using SVD technique could only guarantee that the fraction of partition errors vanishes with high probability. For a stronger result, i.e., that the probability of partition error tends to zero, one may rely on another technique, "combinational projection" [77], instead of performing only the SVD. Essentially, the combinational projection gives another projection, after the one on the singular space, onto the span of characteristic vectors generated from a certain threshold. This way, the variances of target random variables are significantly reduced and the Chernoff-type argument could be adopted to arrive at a tighter bound on the distance of row spaces after the final projection. Note that* (7.26) *still holds in this case, and that by replacing the corresponding bounds in [77] it follows that* $\Theta(n \ln n)$ *reads are sufficient to exactly recover the haplotype with high probability.*

Remark 7.11. *Spectral partitioning is a very simple and computationally efficient algorithm that employs only the majority voting and the SVD techniques to perform haplotype assembly. In fact, we do not even require a full SVD calculation since only the second eigenvector is needed to determine the haplotype, as described in the algorithm. Therefore, by using the power method to discover the desired eigenvector, the complexity of spectral partitioning can*

be reduced from $O(n^3)$ in the general case to $O(n \ln n)$ for sparse adjacency matrix (since the number of total entries observed is roughly $O(n \ln n)$).

Remark 7.12. *Although the theoretical analysis presented in Section* **??**.*B is conducted under the assumption that there are precisely two entries observed in each row of the SNP fragment matrix, the results can easily be generalized to the case with multiple entries per row as long as reads may sample all pairs of SNP positions with non-trivial probability. If, however, the insert size is fixed or characterized by small variance, an alternative quantification of the minimum number of entries guaranteeing recovery of a low rank matrix may be needed. To this end, we note that a related matrix completion problem was studied in in [83] [84] [85], where an optimization approach was utilized to determine the necessary conditions and the recovery was facilitated by solving an appropriately formulated convex program. For our haplotype assembly problem, the observed fragments matrix $\boldsymbol{R}$ could be interpreted as a combination of the true haplotype matrix $\boldsymbol{S}$ and an independent sequencing error matrix $\boldsymbol{N}$. Moreover, the MEC criterion score is equivalent to the minimum l_1-norm of $\boldsymbol{N}$, and the associated optimization problem is given by*

$$\begin{aligned} min \quad & ||\boldsymbol{S}||_* + \gamma ||\boldsymbol{N}||_1 \\ s.t. \quad & \mathcal{P}_\Omega(\boldsymbol{S} \oplus \boldsymbol{N}) = \mathcal{P}_\Omega(\boldsymbol{R}), \end{aligned}$$

where $\|\boldsymbol{S}\|_*$ is the nuclear norm of $\boldsymbol{S}$ and γ denotes the balancing weight. [86] [87] report that the row space of the original matrix could be reliably recovered as long as the number of observed entries is large enough. Putting it more

precisely, the number of reads needed for recovery is as least $\Omega(n \cdot poly(\ln n))$*, which does not outperform the bound we obtained by relying on spectral partitioning. The kernel technique utilized in general for this type of proofs is the Golfing Scheme [86] [87], which requires a lower bound on the number of sampled entries to construct the dual certificate. If a new technique with a better performance guarantee could be used instead of the Golfing Scheme (at least for the case of the specific problem structure encountered in haplotype assembly), then the optimality method may also be able to provide the necessary condition that is characterized by a log-factor gap. Results utilizing this optimization method will be reported elsewhere in the future.*

7.5 Simulation Results and Analysis

7.5.1 Simulation on a Synthetic Data Set

We first test the performance of the two proposed algorithms – erasure decoding and spectral partitioning – on a synthetic data set. To this end, haplotypes are randomly generated according to a uniform distribution, followed by sampling paired-end fragments from haplotypes randomly and uniformly with i.i.d. sampling errors. For the moment, we enforce that 2 SNPs are observed in each fragment. The target of this simulation study is to empirically explore the relations among three key parameters featured in the algorithms, i.e., the length of the haplotype n, the probability of sampling errors p, and, most importantly, the number of sampled reads m. We show that the simulation results verify the conclusions of the theorems presented in the earlier

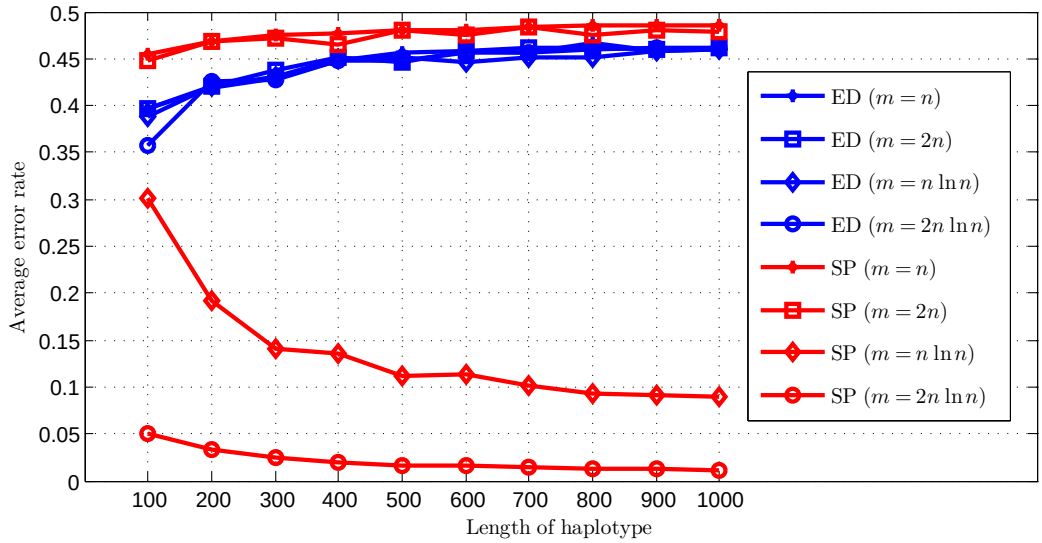

Figure 7.5: **Plot of average error rates from** 100 **random simulations where the probability of sampling errors is set to** $p = 0.1$**.** In this simulation, we illustrate how the accuracy of haplotype assembly depends on relationship between the number of reads m and the haplotype length n for both erasure decoding (ED) and spectral partitioning (SP).

sections of this paper, and also provide intuition for selecting appropriate parameters from the practical point of view.

To begin with, we set the probability of sampling error $p = 0.1$ (significantly larger than the typical value in practice), and study how the accuracy of haplotype assembly depends on relationship between the number of reads m and the haplotype length n. The results, shown in Figure 7.5, provide the following observations:

1) The erasure decoding algorithm fails to assemble the haplotype for all choices of m, which is basically due to large sampling noise. As indicated

in Section III, this algorithm is intuitively designed for the error-free case and it has no performance guarantees when adopted and applied for the erroneous case.

2) For spectral partitioning, choosing $m = \Theta(n)$ is not sufficient to ensure reliable recovery, while choosing $m = \Theta(n \ln n)$ is sufficient to guarantee that the error fraction vanishes for large n. This result is consistent with the conclusion of Theorem 7.6.

3) Spectral partitioning, when implemented with sufficiently large number of reads (i.e., $m = \Theta(n \ln n)$), provides better error rate for large haplotype lengths. This is predicted by the theoretical result provided by equation (7.28).

Next, motivated by the results of the theoretical analysis and the previously described initial simulation results, we scale the number of sampled reads as $m = 2n \ln n$ and empirically study how the performance of both algorithms depends upon sampling errors and haplotype lengths. The results of simulation are illustrated in Figure 7.6, leading to the following observations:

4) The erasure decoding algorithm performs extremely well in the error-free case when the number of fragments is sufficiently large. However, in the erroneous case, this algorithm fails to recover the original haplotypes with high confidence.

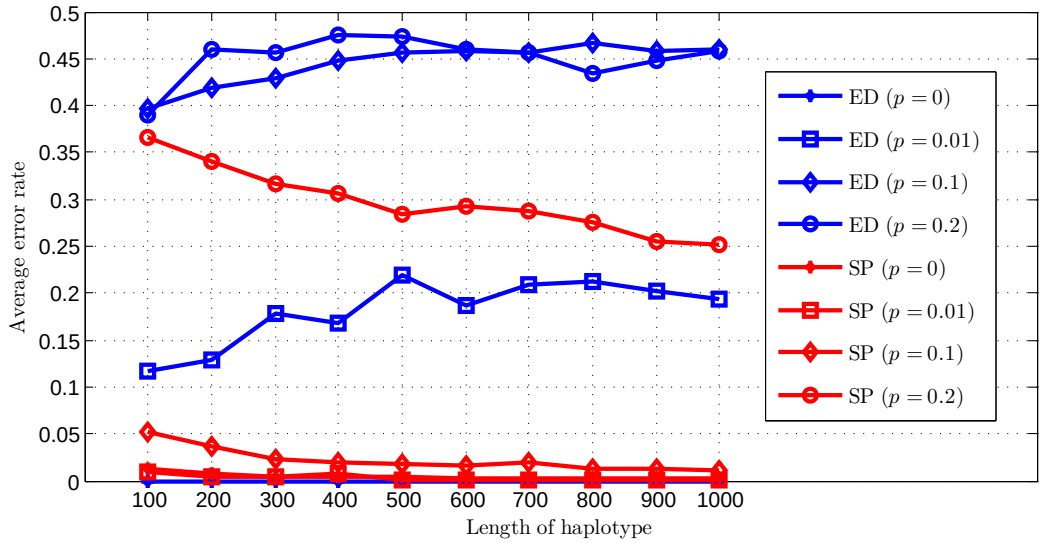

Figure 7.6: **Plot of the average error rates evaluated based on** 100 **random simulations where the number of reads is** $m = 2n \ln n$**.** Here we illustrate how the performance depends on sampling errors for both erasure decoding (ED) and spectral partitioning (SP).

5) The convergence rate for spectral partitioning highly depends on p. More specifically, spectral partitioning is well-suited for the low-noise scenario, i.e., $p \leq 0.1$, which is typical of practical applications.

These results on synthetic data verify the results of our theoretical analysis in Section 7.3 and Section 7.4, and the overall conclusions may be summarized as follows:

Erasure decoding is applicable only in the noise-free setting and it requires $m = \Theta(n \ln n)$ reads for a reliable assembly of a haplotype of length n.

Spectral partitioning proves useful in the low-noise scenario (e.g., $p \leq$

0.1). It, too, requires $m = \Theta(n \ln n)$ reads for a reliable assembly of a haplotype of length n. When these two conditions are met, spectral partitioning is capable of recovering the original haplotype with high accuracy, and the recovery rate is inversely proportional to the length of the haplotype.

7.5.2 Simulation on a Benchmark Database

Here we present the study of the performance of both algorithms on the database created in [88], generated from the Phase I of the HapMap project [89] and widely adopted for benchmarking the effectiveness of haplotype assembly algorithms. This database consists of all 22 chromosomes from 209 unrelated individuals; shotgun sequencing process has been simulated to obtain the SNP observation matrix. Note that only heterogeneous SNP sites are considered in our study and that the recovery rate is computed based on the haplotype block lengths after filtering out the homozygous sites. Moreover, note that here the number of SNPs covered by reads varies and is no longer fixed to 2 as was the case in Section V.A. Nevertheless, our algorithms can be directly applied since the assumption on having precisely 2 observations per read was only needed to allow theoretical analysis.

TABLE 7.1 shows the average recovery rate computed using 100 data sets from [88], where the free parameters include: 1) the haplotype length $n = 100, 350, 700$; 2) the coverage $c = 3, 5, 8, 10$; and 3) the sampling error rate $p = 0\%, 10\%, 20\%$. From the simulation results, we find that erasure decoding successfully assembles the haplotype with high probability when $p = 0$, but

fails to do so when $p > 0$. Moreover, sparse partitioning performs well in comparison with several recently proposed algorithms when the number of reads is sufficiently large. Therefore, our proposed algorithms, primarily meant to support theoretical results, also have practical significance.

7.6 Summary

In this chapter, we study the haplotype assembly problem from an information-theoretic perspective. To determine the chromosome membership of reads provided by high-throughput sequencing systems and thus enable haplotype assembly, we interpret the problem as the one of decoding data messages that are encoded and transmitted over a particular channel model. This channel model reflects the salient features of the paired-end sequencing technology and the haplotype assembly problem.

In the case of error-free sequencing, we find that the required number of reads needed for reconstruction is at least of the same order as the length of the haplotype sequence. To establish a sufficient condition, we analyze an erasure decoding algorithm that utilizes the common information across reads to iteratively recover haplotypes. We find that this algorithm ensures reconstruction with the optimal scaling of the number of reads.

In the case of erroneous sequencing, where errors are assumed to be generated independently and identically, we show that the number of reads needed to recover the haplotype is of the same order as in the error-free case. For the sufficient condition, we rephrase the original haplotype assembly problem as a

low-rank matrix recovery. Using matrix permutation theory, we illustrate that haplotype sequences could be recovered reliably when the number of reads scales as $\Theta(n \ln n)$, where n denotes the haplotype length.

Simulation results corroborate theoretical claims, and the information-theoretic view of the haplotype assembly problem is worth pursuing in other DNA-sequencing related applications (e.g., population haplotyping).

	Algorithms	$p = 0.0$				$p = 0.1$				$p = 0.2$			
		$c = 3$	$c = 5$	$c = 8$	$c = 10$	$c = 3$	$c = 5$	$c = 8$	$c = 10$	$c = 3$	$c = 5$	$c = 8$	$c = 10$
$n = 100$	SpeedHap	0.999	1.000	1.000	1.000	0.895	0.967	0.989	0.990	0.623	0.799	0.852	0.865
	Fast Hare	0.999	0.999	1.000	1.000	0.919	0.965	0.993	0.998	0.715	0.797	0.881	0.915
	2d-mec	0.990	0.997	1.000	1.000	0.912	0.951	0.983	0.988	0.738	0.793	0.873	0.894
	HapCUT	1.000	1.000	1.000	1.000	0.929	0.920	0.901	0.892	0.782	0.838	0.864	0.871
	MLF	0.973	0.992	0.997	0.998	0.889	0.970	0.985	0.995	0.725	0.836	0.918	0.938
	SHR-three	0.816	0.861	0.912	0.944	0.696	0.738	0.758	0.762	0.615	0.655	0.681	0.699
	ED	1.000	1.000	1.000	1.000	0.650	0.651	0.627	0.639	0.587	0.581	0.585	0.593
	SP	0.958	0.997	0.999	1.000	0.883	0.961	0.990	0.995	0.687	0.809	0.918	0.943
$n = 350$	SpeedHap	0.999	1.000	1.000	1.000	0.819	0.959	0.984	0.984	0.439	0.729	0.825	0.855
	Fast Hare	0.990	0.999	1.000	0.999	0.871	0.945	0.985	0.995	0.684	0.746	0.853	0.877
	2d-mec	0.965	0.993	0.998	0.999	0.837	0.913	0.964	0.978	0.675	0.729	0.791	0.817
	HapCUT	1.000	1.000	1.000	1.000	0.930	0.913	0.896	0.888	0.771	0.831	0.862	0.867
	MLF	0.864	0.929	0.969	0.981	0.752	0.858	0.933	0.962	0.642	0.728	0.798	0.831
	SHR-three	0.830	0.829	0.895	0.878	0.682	0.724	0.742	0.728	0.591	0.632	0.670	0.668
	ED	1.000	1.000	1.000	1.000	0.608	0.595	0.587	0.586	0.553	0.549	0.538	0.547
	SP	0.903	0.972	0.992	0.997	0.768	0.933	0.983	0.992	0.598	0.679	0.843	0.905
$n = 700$	SpeedHap	0.999	1.000	1.000	1.000	0.705	0.947	0.985	0.986	0.199	0.681	0.801	0.813
	Fast Hare	0.988	0.999	1.000	0.999	0.829	0.949	0.986	0.995	0.652	0.712	0.808	0.872
	2d-mec	0.946	0.976	0.992	0.997	0.786	0.880	0.948	0.965	0.647	0.697	0.751	0.778
	HapCUT	1.000	1.000	1.000	1.000	0.927	0.916	0.896	0.889	0.753	0.825	0.856	0.861
	MLF	0.787	0.854	0.919	0.933	0.698	0.809	0.863	0.884	0.624	0.682	0.747	0.765
	SHR-three	0.781	0.832	0.868	0.898	0.668	0.716	0.743	0.726	0.591	0.617	0.653	0.675
	ED	1.000	1.000	1.000	1.000	0.576	0.571	0.572	0.573	0.534	0.532	0.531	0.528
	SP	0.887	0.967	0.991	0.997	0.723	0.910	0.977	0.990	0.562	0.610	0.751	0.843

Table 7.1: **Comparisons of our algorithms, erasure decoding (ED) and spectral partitioning (SP), with existing algorithms.** Each entry in the table represents the average recovery rate from 100 randomly generated haplotype observation matrices, with respect to different n, c, and p.

7.A Proof of Lemma 7.8

Assume $m = \kappa_1 n \ln n$, where κ_1 is a positive constant. In order to provide a lower bound for α, we truncate the first summation by leaving only the term with $i = 1$. More precisely, we have by denoting $\gamma \triangleq 2/n(n-1)$,

$$\begin{aligned}
\alpha &= \sum_{i=1}^{m} \left\{ \binom{m}{i} \gamma^i (1-\gamma)^{m-i} \sum_{l=\lfloor i/2 \rfloor + 1}^{i} \binom{i}{l} [(1-p)^2 + p^2]^l [2p(1-p)]^{i-l} \right\} \\
&\geq \binom{m}{1} \left[\frac{2}{n(n-1)} \right] \left[1 - \frac{2}{n(n-1)} \right]^{m-1} \binom{1}{1} [(1-p)^2 + p^2][2p(1-p)]^0 \\
&\geq \frac{2\kappa_1 n \ln n}{n(n-1)} e^{-\frac{4\kappa_1 n \ln n}{n(n-1)}} [(1-p)^2 + p^2] \\
&= \frac{2\kappa_1 [(1-p)^2 + p^2] n^{-\frac{4\kappa_1}{n-1}} \ln n}{n-1}.
\end{aligned}$$

Note that $n^{-\frac{4\kappa_1}{n-1}}$ is an increasing function with n and tends to 1. Hence, for large enough n, there exists a constant $\kappa_2 < 1$ such that

$$n^{-\frac{4\kappa_1}{n-1}} \geq \kappa_2. \tag{7.30}$$

As a result, the lower bound becomes

$$\alpha \geq \frac{2\kappa_1 \kappa_2 [(1-p)^2 + p^2] \ln n}{n-1}. \tag{7.31}$$

Thus, α has a $\Theta(n^{-1} \ln n)$ scale lower bound. In fact, this bound is rather tight, because the first term ($i = 1$) dominates the overall value (analogue to the analysis of β that follows next).

In addition, we need to establish an upper bound on β. In particular, we show that the terms in the above summation are at least exponentially

decreasing, such that the first term dominates the value of β. For this purpose, we denote

$$\beta_i \triangleq \binom{m}{i}\left[\frac{2}{n(n-1)}\right]^i \gamma^i(1-\gamma)^{m-i}\sum_{l=\lfloor i/2\rfloor+1}^{i}\binom{i}{l}[2p(1-p)]^l[(1-p)^2+p^2]^{i-l},$$

where $\gamma = 2/n(n-1)$. Introducing

$$\beta_i^{(l)} \triangleq \binom{i}{l}[2p(1-p)]^l[(1-p)^2+p^2]^{i-l}$$

and

$$\beta = \sum_{i=1}^{m}\beta_i,$$

it follows that

$$\beta_i = \binom{m}{i}\left[\frac{2}{n(n-1)}\right]^i\left[1-\frac{2}{n(n-1)}\right]^{m-i}\sum_{l=\lfloor i/2\rfloor+1}^{i}\beta_i^{(l)}.$$

In order to derive a lower bound on β_i/β_{i+1} for any i, we focus on two cases:

1. For even i, write $i = 2k$ and note that

$$\begin{aligned}\frac{\beta_{2k}}{\beta_{2k+1}} &= \frac{\binom{m}{2k}\left[\frac{2}{n(n-1)}\right]^{2k}\left[1-\frac{2}{n(n-1)}\right]^{m-2k}\sum_{l=k+1}^{2k}\beta_{2k}^{(l)}}{\binom{m}{2k+1}\left[\frac{2}{n(n-1)}\right]^{2k+1}\left[1-\frac{2}{n(n-1)}\right]^{m-2k-1}\sum_{l=k+1}^{2k+1}\beta_{2k+1}^{(l)}}\\ &= \frac{(2k+1)[n(n-1)-2]\sum_{l=k+1}^{2k}\beta_{2k}^{(l)}}{2(\kappa_1 n\ln n-2k)\sum_{l=k+1}^{2k+1}\beta_{2k+1}^{(l)}}.\end{aligned}$$

   Note that there are $k+1$ terms for $\beta_{2k+1}^{(l)}$ in the denominator, but only k terms for $\beta_{2k}^{(l)}$ in the numerator. Hence, we duplicate the numerator to

compare it with the denominator. More precisely, for $k+1 \leq l \leq 2k$,

$$\frac{\beta_{2k}^{(l)}}{\beta_{2k+1}^{(l)}} = \frac{2k+1-l}{(2k+1)[(1-p)^2+p^2]} \geq \frac{1}{2k+1}, \tag{7.32}$$

where the last inequality holds due to $(1-p)^2+p^2 \leq 1$. Moreover,

$$\frac{\beta_{2k}^{(k+1)}}{\beta_{2k+1}^{(2k+1)}} = \frac{(2k)![(1-p)^2+p^2]^{k-1}}{(k+1)!(k-1)![2p(1-p)]^k} \geq \frac{1}{2k+1}, \tag{7.33}$$

where the last inequality holds due to $1 \geq (1-p)^2+p^2 \geq 2p(1-p)$. Combining these two expressions, we have

$$\begin{aligned}
\frac{2\beta_{2k}}{\beta_{2k+1}} &= \frac{(2k+1)[n(n-1)-2]\left\{\sum_{l=k+1}^{2k}\beta_{2k}^{(l)} + \sum_{l=k+1}^{2k}\beta_{2k}^{(l)}\right\}}{2(\kappa_1 n\ln n - 2k)\left\{\sum_{l=k+1}^{2k}\beta_{2k+1}^{(l)} + \beta_{2k+1}^{(2k+1)}\right\}} \\
&\geq \frac{(2k+1)[n(n-1)-2]\left\{\sum_{l=k+1}^{2k}\beta_{2k}^{(l)} + \beta_{2k}^{(k+1)}\right\}}{2(\kappa_1 n\ln n - 2k)\left\{\sum_{l=k+1}^{2k}\beta_{2k+1}^{(l)} + \beta_{2k+1}^{(2k+1)}\right\}} \\
&\geq \frac{(2k+1)[n(n-1)-2]}{2(\kappa_1 n\ln n - 2k)(2k+1)} \\
&= \frac{n(n-1)-2}{2(\kappa_1 n\ln n - 2)}.
\end{aligned}$$

Thus,

$$\frac{\beta_{2k}}{\beta_{2k+1}} \geq \frac{n(n-1)-2}{4(\kappa_1 n\ln n - 2)}. \tag{7.34}$$

2. For i odd, write $i = 2k-1$ and note that

$$\frac{\beta_{2k-1}}{\beta_{2k}} = \frac{\binom{m}{2k-1}\left[\frac{2}{n(n-1)}\right]^{2k-1}\left[1-\frac{2}{n(n-1)}\right]^{m-2k+1}\sum_{l=k}^{2k-1}\beta_{2k-1}^{(l)}}{\binom{m}{2k}\left[\frac{2}{n(n-1)}\right]^{2k}\left[1-\frac{2}{n(n-1)}\right]^{m-2k}\sum_{l=k+1}^{2k}\beta_{2k}^{(l)}}$$

$$= \frac{2k[n(n-1)-2]\sum_{l=k}^{2k-1}\beta_{2k-1}^{(l)}}{2(\kappa_1 n\ln n - 2k+1)\sum_{l=k+1}^{2k}\beta_{2k}^{(l)}}.$$

In this case, both numerator and denominator have k terms in summation. Hence, term-by-term comparison leads to

$$\frac{\beta_{2k-1}^{(l)}}{\beta_{2k}^{(l)}} = \frac{2k-l}{2k[(1-p)^2+p^2]} \geq \frac{1}{2k}. \tag{7.35}$$

Thus,

$$\frac{\beta_{2k-1}}{\beta_{2k}} \geq \frac{n(n-1)-2}{2(\kappa_1 n\ln n - 1)}. \tag{7.36}$$

Note that, in both cases, the lower bounds (7.34) and (7.36) tend to infinity as n increases. Therefore, there exists a constant $\kappa_3 > 1$ such that for large enough n

$$\min\left\{\frac{n(n-1)-2}{4(\kappa_1 n\ln n - 2)}\,,\ \frac{n(n-1)-2}{2(\kappa_1 n\ln n - 1)}\right\} \geq \kappa_3, \tag{7.37}$$

which further implies that for any value of i,

$$\frac{\beta_i}{\beta_{i+1}} \geq \kappa_3.$$

Based on this we obtain $\beta_i \leq \beta_1 \kappa_3^{1-i}$ and

$$\begin{aligned}\beta_1 &= \binom{m}{1}\left[\frac{2}{n(n-1)}\right]\left[1-\frac{2}{n(n-1)}\right]^{m-1}\binom{1}{1}[2p(1-p)][(1-p)^2+p^2]^0\\ &\leq \frac{2\kappa_1 n\ln n}{n(n-1)}e^{-\frac{2\kappa_1 n\ln n}{n(n-1)}}[2p(1-p)]\\ &= \frac{2\kappa_1[2p(1-p)]n^{-\frac{2\kappa_1}{n-1}}\ln n}{n-1}\end{aligned}$$

$$\leq \frac{2\kappa_1[2p(1-p)]\ln n}{n-1},$$

where we have used the fact that $n^{-\frac{2\kappa_1}{n-1}} \leq 1$. Hence, we obtain

$$\begin{aligned}\beta &= \sum_{i=1}^{m} \beta_i \\ &\leq \sum_{i=1}^{m} \beta_1 \kappa_3^{1-i} \\ &\leq \frac{2\kappa_1[2p(1-p)]\ln n}{(n-1)(1-\kappa_3^{-1})}. \end{aligned} \tag{7.38}$$

Thus, the upper bound for β is also $\Theta(n^{-1}\ln n)$ scale.

A point to clarify: in several places we have somewhat imprecisely used categorization "large enough n." One may be concerned with whether a particular choice of n satisfying the proof assumptions could match the practical haplotype assembly scenarios. As an illustration, for $\kappa_1 = 2$ that we use in the simulation setup, a simple choice of $\kappa_2 = 1/2$ and $\kappa_3 = 2$ implies that the minimum value of n needed to satisfy both assumptions (7.30) and (7.37) is given by

$$n \geq \max\{45, 69, 28\} = 69,$$

which is quite smaller than the commonly encountered value in the haplotype assembly problems. Therefore, our bounds are meaningful and useful in practical scenarios.

Chapter 8

Conclusion

In this thesis, we study the coding mechanisms for communication and compression, from the perspective of wireless communication and DNA sequencing respectively. Three important topics are considered, namely expansion coding for analog data transmission and compression, hierarchical polar coding scheme for fading channels, and information-theoretic analysis for haplotype assembly. In this chapter, we review our main contributions to each topic and point out a few possible directions for future research.

8.1 Summary of Main Results

8.1.1 Part I: Expansion Coding for Data Transmission and Compression

In Chapter 2, we propose expansion coding scheme to construct good codes for analog channel coding. With a perfect or approximate decomposition of channel noises, we consider coding over independent parallel representations, which provides a foundation for reducing the original problems to a set of parallel simpler subproblems. In particular, via expansion channel coding, we consider coding over q-ary channels for each expanded level. This approximation of the original channel together with capacity achieving codes

for each level (to reliably transmit messages over each channel constructed) and Gallager's method (to achieve desired communication rates for each channel) allow for constructing near-capacity achieving codes for the original channel. Both theoretical analysis and numeric result show the proposed coding scheme achieves the capacity of AEN channel at high SNR regime, when implemented with enough number of expanded levels.

In Chapter 3, similar to the case of channel coding, we utilize expansion coding to adopt discrete source codes achieving rate distortion limit on each level after expansion, and design codes achieving near-optimal performance for the original source. Exponential and Laplacian sources are concerned as examples to show effectiveness of the proposed scheme. Theoretical analysis and numerical results are provided to detail performance guarantees of the proposed expansion coding scheme.

To this end, expansion coding is proved to be an effective coding scheme framework for both data transmission over analog noise channel, and data compression of analog sources. The advantages of expansion coding scheme lie in its ability to shoot for near optimal rate, and to guarantee coding complexities tractable at the same time.

8.1.2 Part II: Hierarchical Polar Coding Scheme for Fading Channels

In Chapter 4, a hierarchical polar coding scheme is proposed for the fading BSC. This novel scheme, by exploiting an erasure decoding approach at

the receiver, utilizes the polarization results of different BSCs. (These BSCs are defined over channel uses at a given fading block and over fading blocks at a given channel use index.) This novel polar coding technique is shown to be capacity achieving for fading BSC.

In Chapter 5, we illustrate the utilization of hierarchical polar coding scheme for encoding over another fading channel model. For the fading AEN channel model, expansion coding is adopted to convert the problem of coding over an analog fading channel into coding over discrete fading channels. By performing this expansion approach and making the resulting channels independent (via decoding the underlying carries), a fading AEN channel is decomposed into multiple independent fading BSCs (with a reliable decoding of the carries). By utilizing the hierarchical polar coding scheme for fading BSC, both theoretical proof and numerical results show that the proposed approach achieves the capacity of this fading channel in the high SNR regime.

In Chapter 6, we move our focus to binary symmetric wiretap channels with block fading. By exploiting an erasure decoding approach at the receiver, this scheme utilizes the polarization of degraded binary symmetric channels to survive from the impact of fading. Meanwhile, to combat with eavesdropping, random bits are injected into the encoded symbols, and the resulting coding scheme is shown to achieve the secrecy capacity for the case of simultaneous fading of the main channel and eavesdropper channel.

Remarkably, for all of the fading channel concerned in this part, the proposed scheme does not assume channel state information at the transmitter,

and fading BSC models the fading additive white Gaussian noise (AWGN) channel with a BPSK modulation. Therefore, our results are quite relevant to the practical channel models considered in wireless communications.

8.1.3 Part III: Information-Theoretic Analysis for Haplotype Assembly

In Chapter 7, we study the haplotype assembly problem from an information-theoretic perspective. To determine the chromosome membership of reads provided by high-throughput sequencing systems and thus enable haplotype assembly, we interpret the problem as the one of decoding data messages that are encoded and transmitted over a particular channel model. This channel model reflects the salient features of the paired-end sequencing technology and the haplotype assembly problem.

In the case of error-free sequencing, we find that the required number of reads needed for reconstruction is at least of the same order as the length of the haplotype sequence. To establish a sufficient condition, we analyze an erasure decoding algorithm that utilizes the common information across reads to iteratively recover haplotypes. We find that this algorithm ensures reconstruction with the optimal scaling of the number of reads.

In the case of erroneous sequencing, where errors are assumed to be generated independently and identically, we show that the number of reads needed to recover the haplotype is of the same order as in the error-free case. For the sufficient condition, we rephrase the original haplotype assembly problem as a

low-rank matrix recovery. Using matrix permutation theory, we illustrate that haplotype sequences could be recovered reliably when the number of reads scales as $\Theta(n \ln n)$, where n denotes the haplotype length.

8.2 Future Directions

This thesis is motivated by our recognition of recent trend in information theory and coding theory, and represents our initial efforts to develop coding schemes for practical communication and compression problems. We take some first steps in developing and analyzing theoretical models and coding mechanisms for wireless communication and DNA sequencing. There are still many interesting open problems and unsolved issues in this broad area. Here, we summarize three such research directions as the closure of this thesis.

8.2.1 Expansion Coding for AWGN Channel

Although expansion coding has a good performance for AEN channel at high SNR regime, a more popular model concerned in communications is AWGN channel. To design an expansion coding method for AWGN, the challenges come from two aspects.

First, Gaussian distribution is not decomposable. More precisely, unlike exponential distribution, Gaussian distribution cannot be expressed a summation of independent random variables, however, it could be approximated by its binary expansion very closely, where each levels are forced to be independent. To this end, we are coding over a set of approximated discrete channels,

and proper argument should be concerned here in order to guarantee that the achievable scheme for the approximated channel also fit for the original channel.

Secondly, Gaussian distribution is two-sided. Again, unlike the exponential distribution, the support set for Gaussian distribution is the real set, which means the issue with sign should be taken into consideration in coding. More precisely, in addition to carries, we also have borrows in the design of coding scheme. Because borrows and carries happen arbitrarily with the same probability, they cannot be decoded corrected in any coding scheme.

8.2.2 Expansion Coding for Multi-User Channels

Deterministic model is an effective tool to study analog noise channel coding problems, where the basic idea is to construct an approximate channel for which the transmitted signals are assumed to be noiseless above the noise level. This approach is proved to be very effective in analyzing the capacity of networks. In particular, it has been testified that this framework perfectly represents and helps to characterize degrees of freedom of point-to-point AWGN channels, as well as some multi-user channels of concern.

In this sense, expansion coding scheme can be seen as a generalization of these deterministic approaches. Here, the effective noise in the channel is carefully calculated and the system takes advantage of coding over the noisy levels at any SNR. This generalized channel approximation approach can be useful in reducing the large gaps reported in the previous works, because the

noise approximation in our work is much closer to the actual distribution as compared to that of the deterministic model.

8.2.3 Information-Theoretic Analysis for Population Haplotyping

Population haplotyping is another important DNA sequencing related problem. Unlike haplotype assembly, population haplotyping aims to recover potential haplotypes from observed genotypes generated from group of diploid species. Similar information theoretic tool can be utilized to analyze the necessary condition of the number of observations (i.e., genotypes) for perfect or near-perfect recovery, however, the corresponding coding scheme that can achieve the necessary condition (with proper gap) is challenging and still open.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[2] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, Apr. 1950.

[3] M. J. E. Golay, "Notes on digital coding," *Proceeding of the Institute of Radio Engineers*, vol. 37, no. 6, pp. 657–657, Jun. 1949.

[4] D. E. Muller, "Application of boolean algebra to switching circuit design and to error detection," *IRE Transactions on Electronic Computers*, vol. 3, no. 3, pp. 6–12, Sep. 1954.

[5] I. S. Reed, "A class of multiple-error-correcting codes and the decoding scheme," *IRE Transactions on Information Theory*, vol. 4, no. 4, pp. 38–49, Sep. 1954.

[6] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial & Applied Mathematics*, vol. 8, no. 2, pp. 300–304, Jun. 1960.

[7] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. New York: Springer, 1988.

[8] F. J. McWilliams and N. J. A. Sloane, *The theory for error-correcting codes*. North-Holland, 1983.

[9] J. G. D. Forney, "Convolutional codes I: Algebraic structure," *IEEE Transactions on Information Theory*, vol. 16, no. 6, pp. 720–738, Nov. 1970.

[10] ——, *Concatenated codes*. Cambridge: MIT press, 1966.

[11] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1064–1070, Oct. 1996.

[12] D. J. C. MacKay and R. M. Neal, "Good codes based on very sparse matrices," in *Cryptography and Coding*. Springer, 1995, pp. 100–111.

[13] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1710–1722, Nov. 1996.

[14] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.

[15] J. Perry, H. Balakrishnan, and D. Shah, "Rateless spinal codes," in *Proc. of the 10th ACM Workshop on Hot Topics in Networks (HotNets 2011)*, New York City, New York, USA, Nov. 2011, pp. 1–6.

[16] H. Balakrishnan, P. Iannucci, D. Shah, and J. Perry, "De-randomizing Shannon: The design and analysis of a capacity-achieving rateless code," *arXiv:1206.0418*, Jun. 2012.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.

[18] A. J. Viterbi and J. K. Omura, "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 325–332, May 1974.

[19] Y. Matsunaga and H. Yamamoto, "A coding theorem for lossy data compression by LDPC codes," *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2225–2229, Sep. 2003.

[20] M. J. Wainwright, E. Maneva, and E. Martinian, "Lossy source compression using low-density generator matrix codes: Analysis and algorithms," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1351–1368, Mar. 2010.

[21] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1751–1768, Apr. 2010.

[22] E. Arıkan, "Source polarization," in *Proc. 2010 IEEE International Symposium on Information Theory (ISIT 2010)*, Austin, Texas, USA, Jun. 2010, pp. 899–903.

[23] M. Karzand and E. Telatar, "Polar codes for q-ary source coding," in *Proc. 2010 IEEE International Symposium on Information Theory (ISIT 2010)*, Austin, Texas, USA, Jun. 2010, pp. 909–912.

[24] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977.

[25] R. Staden, "A strategy of DNA sequencing employing computer programs," *Nucleic Acids Research*, vol. 6, no. 7, pp. 2601–2610, Mar. 1979.

[26] S. Anderson, "Shotgun DNA sequencing using cloned DNase I-generated fragments," *Nucleic Acids Research*, vol. 9, no. 13, pp. 3015–3027, May 1981.

[27] G. A. Churchill and M. S. Waterman, "The accuracy of DNA sequences: Estimating sequence quality," *Genomics*, vol. 14, no. 1, pp. 89–98, Sep. 1992.

[28] J. D. J. Costello and J. G. D. Forney, "Channel coding: The road to channel capacity," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1150–1177, Jun. 2007.

[29] J. G. D. Forney, M. D. Trott, and S. Y. Chung, "Sphere-bound-achieving coset codes and multilevel coset codes," *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 820–850, May 2000.

[30] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information flow: A deterministic approach," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 1872–1905, Apr. 2011.

[31] E. Abbe and A. Barron, "Polar coding schemes for the AWGN channel," in *Proc. 2011 IEEE International Symposium on Information Theory (ISIT 2011)*, St. Petersburg, Russia, Jul. 2011, pp. 194–198.

[32] E. Abbe and E. Telatar, "Polar codes for the m-user multiple access channel," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5437–5448, Aug. 2012.

[33] M. Seidl, A. Schenk, C. Stierstorfer, and J. B. Huber, "Polar-coded modulation," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4108–4119, Oct. 2013.

[34] S. Verdú, "The exponential distribution in information theory," *Problems of Information Transmission*, vol. 32, no. 1, pp. 100–111, Jan. 1996.

[35] A. Martinez, "Communication by energy modulation: The additive exponential noise channel," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3333–3351, Jun. 2011.

[36] S. Y. Le Goff, "Capacity-approaching signal constellations for the additive exponential noise channel," *IEEE Wireless Communications Letters*, vol. 1, no. 4, pp. 320–323, Aug. 2012.

[37] R. G. Gallager, *Information theory and reliable communication.* John Wiley & Sons, 1968.

[38] G. Marsaglia, "Random variables with independent binary digits," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1922–1929, Dec. 1971.

[39] R. Zamir, "Lattices are everywhere," in *Proc. 2009 Information Theory and Applications Workshop (ITA 2009)*, San Diego, California, USA, Feb. 2009, pp. 392–421.

[40] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[41] D. Baron and T. Weissman, "An MCMC approach to universal lossy compression of analog sources," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5230–5240, Oct. 2012.

[42] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[43] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *Proc. 2009 IEEE International Symposium on Information Theory (ISIT 2009)*, Seoul, Korea, Jun. 2009, pp. 1493–1495.

[44] E. Şaşoğlu, E. Arıkan, and E. Telatar, "Polarization for arbitrary discrete memoryless channels," in *Proc. 2009 IEEE Information Theory Workshop (ITW 2009)*, Taormina, Sicily, Italy, Oct. 2009, pp. 144–148.

[45] J. J. Boutros and E. Biglieri, "Polarization of quasi-static fading channels," in *Proc. 2013 IEEE International Symposium on Information Theory (ISIT 2013)*, Istanbul, Turkey, Jul. 2013, pp. 769–773.

[46] A. Bravo-Santos, "Polar codes for the rayleigh fading channel," *IEEE Communications Letters*, vol. 17, no. 12, pp. 2352–2355, Dec. 2013.

[47] E. Arıkan, "Systematic polar coding," *IEEE Communications Letters*, vol. 15, no. 8, pp. 860–862, Aug. 2011.

[48] B. Li, H. Shen, and D. Tse, "An adaptive successive cancellation list decoder for polar codes with cyclic redundancy check," *IEEE Communications Letters*, vol. 16, no. 12, pp. 2044–2047, Dec. 2012.

[49] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[50] A. El Gamal and Y. H. Kim, *Network information theory*. Cambridge University Press, 2011.

[51] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2009.

[52] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley, 2009.

[53] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*. Pearson Education, 2010.

[54] E. Perahia and R. Stacey, *Next generation wireless LANs: Throughput, robustness and reliability in 802.11n*. Cambridge University Press, 2008.

[55] A. Wyner, "The wire-tap channel," *The Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[56] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6428–6443, Oct. 2011.

[57] O. O. Koyluoglu and H. El Gamal, "Polar coding for secure transmission and key agreement," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1472–1483, Oct. 2012.

[58] M. Andersson, V. Rathi, R. Thobaben, J. Kliewer, and M. Skoglund, "Nested polar codes for wiretap and relay channels," *IEEE Communications Letters*, vol. 14, no. 8, pp. 752–754, Aug. 2010.

[59] E. Hof and S. Shamai, "Secrecy-achieving polar-coding," in *Proc. 2010 IEEE Information Theory Workshop (ITW 2010)*, Dublin, Ireland, Aug. 2010, pp. 1–5.

[60] E. Abbe, "Low complexity constructions of secret keys using polar coding," in *Proc. 2012 IEEE Information Theory Workshop (ITW 2012)*, Lausanne, Switzerland, Sep. 2012, pp. 1–5.

[61] E. Sasoglu and A. Vardy, "A new polar coding scheme for strong security on wiretap channels," in *Proc. 2013 IEEE International Symposium on Information Theory Proceedings (ISIT 2013)*, Istanbul, Turkey, Jul. 2013, pp. 1117–1121.

[62] D. Sutter, J. M. Renes, and R. Renner, "Efficient one-way secret-key agreement and private channel coding via polarization," *Advances in Cryptology - ASIACRYPT 2013*, vol. 8269, pp. 194–213, Dec. 2013.

[63] R. A. Chou, M. R. Bloch, and E. Abbe, "Polar coding for secret-key generation," in *Proc. 2013 IEEE Information Theory Workshop (ITW 2013)*, Sevilla, Spain, Spe. 2013, pp. 1–5.

[64] Y. P. Wei and S. Ulukus, "Polar coding for the general wiretap channel," *arXiv:1410.3812*, Oct. 2014.

[65] R. A. Chou and M. R. Bloch, "Polar coding for the broadcast channel with confidential messages and constrained randomization," *arXiv:1411.0281*, Nov. 2014.

[66] T. C. Gulcu and A. Barg, "Achieving secrecy capacity of the wiretap channel and broadcast channel with a confidential component," *arXiv:1410.3422*, Oct. 2014.

[67] M. Zheng, M. Tao, and W. Chen, "Polar coding for secure transmission in MISO fading wiretap channels," *arXiv:1411.2463*, Nov. 2014.

[68] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature*, vol. 200, no. 8, pp. 16–18, Jan. 2007.

[69] R. Schwartz, "Theory and algorithms for the haplotype assembly problem," *Communications in Information & Systems*, vol. 10, no. 1, pp. 23–38, Mar. 2010.

[70] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara *et al.*, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature Genetics*, vol. 40, no. 6, pp. 722–729, Apr. 2008.

[71] A. Edwards, H. Voss, P. Rice, A. Civitello *et al.*, "Automated DNA sequencing of the human HPRT locus," *Genomics*, vol. 6, no. 4, pp. 593–608, Apr. 1990.

[72] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, "SNPs problems, complexity, and algorithms," in *Proc. 9th Annual European Symposium (Algorithms - ESA 2001)*, Aarhus, Denmark, Aug. 2001, pp. 182–193.

[73] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail, "Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem," *Briefings in Bioinformatics*, vol. 3, no. 1, pp. 23–31, Mar. 2002.

[74] Z. Puljiz and H. Vikalo, "A message-passing algorithm for haplotype assembly," in *Proc. 2013 Asilomar Conference on Signals, Systems, and Computers (Asilomar 2013)*, Pacific Grove, California, USA, Nov. 2013, pp. 1–5.

[75] S. Vishwanath, "Information theoretic bounds for low-rank matrix completion," in *Proc. 2010 IEEE International Symposium on Information Theory (ISIT 2010)*, Austin, Texas, USA, Jun. 2010, pp. 1508–1512.

[76] V. Bansal and V. Bafna, "HapCUT: An efficient and accurate algorithm for the haplotype assembly problem," *Bioinformatics*, vol. 24, no. 16, pp. i153–i159, Aug. 2008.

[77] F. McSherry, "Spectral partitioning of random graphs," in *Proc. 42nd IEEE Annual Symposium on Foundations of Computer Science (FOCS 2001)*, Las Vegas, Nevada, USA, Oct. 2001, pp. 529–537.

[78] Z. Füredi and J. Komlós, "The eigenvalues of random symmetric matrices," *Combinatorica*, vol. 1, no. 3, pp. 233–241, Sep. 1981.

[79] V. H. Vu, "Spectral norm of random matrices," in *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC 2005)*, Baltimore, Maryland, USA, May 2005, pp. 423–430.

[80] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, Mar. 1970.

[81] V. H. Vu, "Singular vectors under random perturbation," *Random Structures & Algorithms*, vol. 39, no. 4, pp. 526–538, Dec. 2011.

[82] D. C. Tomozei and L. Massoulié, "Distributed user profiling via spectral methods," *Stochastic Systems*, vol. 4, no. 1, pp. 1–43, Jan. 2014.

[83] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[84] B. Recht, "A simpler approach to matrix completion," *The Journal of Machine Learning Research*, vol. 12, no. 104, pp. 3413–3430, Dec. 2011.

[85] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. 23th Annual Conference on Neural Information Processing Systems (NIPS 2009)*, Whistler, British Columbia, Canada, Dec. 2009, pp. 2080–2088.

[86] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3(11), pp. 1–37, May 2011.

[87] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, Jul. 2013.

[88] F. Geraci, "A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem," *Bioinformatics*, vol. 26, no. 18, pp. 2217–2225, Jul. 2010.

[89] International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, Oct. 2005.

Vita

Hongbo Si received his B.S. in Mathematics and Physics and M.S. in Electrical Engineering from Tsinghua University (Beijing, China). He was a research member of the Complicated Engineering System Laboratory (CESL) in the Department of Electronic Engineering at Tsinghua University. He pursued his doctoral degree in the Department of Electrical and Computer Engineering at the University of Texas at Austin (Austin, TX, USA), where he was a member of the Laboratory of Informatics, Networks and Communications (LINC) and Wireless Networking and Communication Group (WNCG). His research focused on information theory, coding theory and bio-informatics.

Permanent email: sihongbo.utexas@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.