Polar Codes for Channel and Source Coding

THÈSE N^{O} 4461 (2009)

PRÉSENTÉE LE 15 JUILLET 2009

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DE THÉORIE DES COMMUNICATIONS

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Satish Babu KORADA

acceptée sur proposition du jury:

Prof. M. Hasler, président du jury
Prof. R. Urbanke, Dr N. Macris, directeurs de thèse
Prof. E. Arikan, rapporteur
Prof. A. Montanari, rapporteur
Prof. E. Telatar, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2009

Abstract

The two central topics of information theory are the compression and the transmission of data. Shannon, in his seminal work, formalized both these problems and determined their fundamental limits. Since then the main goal of coding theory has been to find practical schemes that approach these limits.

Polar codes, recently invented by Arıkan, are the first "practical" codes that are known to achieve the capacity for a large class of channels. Their code construction is based on a phenomenon called "channel polarization". The encoding as well as the decoding operation of polar codes can be implemented with $O(N \log N)$ complexity, where N is the blocklength of the code.

We show that polar codes are suitable not only for channel coding but also achieve optimal performance for several other important problems in information theory. The first problem we consider is lossy source compression. We construct polar codes that asymptotically approach Shannon's rate-distortion bound for a large class of sources. We achieve this performance by designing polar codes according to the "test channel", which naturally appears in Shannon's formulation of the rate-distortion function. The encoding operation combines the successive cancellation algorithm of Arıkan with a crucial new ingredient called "randomized rounding". As for channel coding, both the encoding as well as the decoding operation can be implemented with $O(N \log N)$ complexity. This is the first known "practical" scheme that approaches the optimal rate-distortion trade-off.

We also construct polar codes that achieve the optimal performance for the Wyner-Ziv and the Gelfand-Pinsker problems. Both these problems can be tackled using "nested" codes and polar codes are naturally suited for this purpose. We further show that polar codes achieve the capacity of asymmetric channels, multi-terminal scenarios like multiple access channels, and degraded broadcast channels. For each of these problems, our constructions are the first known "practical" schemes that approach the optimal performance.

The original polar codes of Arıkan achieve a block error probability decaying exponentially in the square root of the block length. For source coding, the gap between the achieved distortion and the limiting distortion also vanishes exponentially in the square root of the blocklength. We explore other polar-like code constructions with better rates of decay. With this generalization,

we show that close to exponential decays can be obtained for both channel and source coding. The new constructions mimic the recursive construction of Arıkan and, hence, they inherit the same encoding and decoding complexity. We also propose algorithms based on message-passing to improve the finite length performance of polar codes.

In the final two chapters of this thesis we address two important problems in graphical models related to communications. The first problem is in the area of low-density parity-check codes (LDPC). For practical lengths, LDPC codes using message-passing decoding are still the codes to beat. The current analysis, using density evolution, evaluates the performance of these algorithms on a tree. The tree assumption corresponds to using an infinite length code. But in practice, the codes are of finite length. We analyze the message-passing algorithms for this scenario. The absence of tree assumption introduces correlations between various messages. We show that despite this correlation, the prediction of the tree analysis is accurate.

The second problem we consider is related to code division multiple access (CDMA) communication using random spreading. The current analysis mainly focuses on the information theoretic limits, i.e., using Gaussian input distribution. However in practice we use modulation schemes like binary phase-shift keying (BPSK), which is far from being Gaussian. The effects of the modulation scheme cannot be analyzed using traditional tools which are based on spectrum of large random matrices. We follow a new approach using tools developed for random spin systems in statistical mechanics. We prove a tight upper bound on the capacity of the system when the user input is BPSK. We also show that the capacity depends only on the power of the spreading sequences and is independent of their exact distribution.

Keywords : Polar codes, low-complexity schemes, channel coding, source coding, Wyner-Ziv problem, Gelfand-Pinsker problem, multi-terminal scenarios, exponent of polar codes, Reed-Muller Codes, belief propagation, LDPC codes, density evolution, CDMA communication, statistical mechanics, interpolation method.

Résumé

Les deux sujets centraux de la théorie de l'information sont la compression et la transmission des données. Shannon, dans son travail fondamental, formalisa ces deux problèmes et détermina les limites théoriques ultimes associées. Depuis, le but principal de la théorie du codage a été de trouver des schémas de faible complexité qui s'approchent de ces limites.

Les codes *polaires* (polar codes), inventés récemment par Arıkan, sont les premiers codes pratiques qui atteignent la capacité, pour une large classe de canaux. La construction de ces codes est basée sur un phénomène appelé "polarisation du canal". Les opérations de codage, et de décodage basé sur une méthode d'éliminations successives, peuvent être implémentées avec une complexité $O(N \log N)$ où N est la longueur du code.

Comme nous le montrons, les codes polaires sont non seulement bien adaptés pour le codage de canal, mais atteignent la performance optimale de plusieurs autres problèmes importants en théorie de l'information. Le premier problème que nous considérons est le codage de source avec pertes. Nous construisons des codes polaires qui atteignent la limite ultime donnée par la fonction de distorsion de Shannon pour une large classe de sources. Nous montrons que cette performance optimale est atteinte en choisissant un code polaire adapté au "canal-test" qui apparait naturellement dans l'expression de Shannon pour la fonction de distorsion. L'opération de codage combine l'algorithme d'éliminations successives d'Arıkan avec un nouvel ingrédient crucial appelé "l'arrondi aléatoire". Les deux opérations, le codage et le décodage, peuvent être implémentées avec une complexité de $O(N \log N)$. Il s'agit du premier schéma qui est de faible complexité tout en atteignant la limite ultime donnée par la fonction de distorsion.

Nous construisons aussi des codes polaires qui sont optimaux pour les problèmes de Wyner-Ziv et de Gelfand-Pinsker. Dans ces deux problèmes, la performance optimale est atteinte grâce à des codes "emboités", et les codes polaires s'appliquent de façon très naturelle dans ce contexte. Nous montrons aussi que les codes polaires sont optimaux pour les canaux asymétriques, pour des scénarios muti-terminaux comme le canal à accès multiple et le canal broadcast dégradé. Pour chacun de ces problèmes, nos constructions sont les premières connues qui soient de faible complexité tout en atteignant la perfor-

mance optimale.

Les codes polaires initiaux d'Arıkan, pour le codage de canal, atteignent une probabilité d'erreur de bloc décroissant exponentiellement avec la racine carrée de la taille du bloc. Pour le codage de source, la différence entre la distorsion atteinte et la limite ultime décroit aussi exponentiellement en fonction de la racine carrée de la longueur du bloc. Nous explorons d'autres constructions de codes polaires avec de meilleurs taux de décroissance. Avec ces généralisations, nous montrons que des décroissances quasi-exponentielles peuvent être obtenues pour ces deux situations. Les nouvelles constructions mimiquent la construction récursive d'Arıkan, et donc héritent de la même complexité de codage et décodage. Nous proposons aussi des algorithmes basés sur la propagation de messages pour améliorer la performance des codes polaires pour les longueurs finies.

Les deux chapitres finaux de cette thèse s'attachent à deux problèmes importants concernant les modèles sur les graphes pour les communications. Le premier est dans le domaine des codes de parité de basse densité (LDPC). Pour des longueurs utilisées dans la pratique, les codes LDPC avec le décodeur de propagation de messages, restent encore les codes à battre. L'analyse, utilisant l'évolution de densité, évalue la performance de ces algorithmes sur un arbre. L'hypothèse de l'arbre correspond à l'utilisation d'un code de longueur infinie, mais en pratique ceux-ci sont de longueur finie. Nous analysons la propagation de messages pour ce dernier scenario. Sans l'hypothèse de l'arbre des corrélations entre les messgaes sont introduites. Nous montrons que malgré ces corrélations, la prédiction de l'analyse sur l'arbre est correcte.

Le second problème que nous considérons est relié au canal à accès multiple par répartition en code (CDMA) avec étalement aléatoire. L'analyse usuelle porte essentiellement sur les entrées à distribution Gaussienne. Néanmoins, en pratique on utilise des schémas de modulation par déplacement de phase (par exemple BPSK) qui sont loins d'être Gaussiens. Les effets de ces modulations ne peuvent pas être analysés avec les outils traditionnels qui sont basés sur la théorie des grandes matrices aléatoires. Nous suivons une nouvelle approche utilisant les méthodes developpées pour les systèmes de spin aléatoires en mécanique statistique. Nous prouvons une borne supérieure optimale sur la capacité du système pour des entrées BPSK. Nous montrons aussi que la capacité dépend seulement de la puissance des séquences d'étalement et est indépendante du détail de la distribution.

Mots clés: codes polaires, schémas de faible complexité, codage de canal, codage de source, problème de Wyner-Ziv, problème de Gelfand-Pinsker, scénarios multi-terminaux, exposants des codes polaires, codes de Reed-Muller, propagation de messages, codes LDPC, évolution de densité, communication CDMA, mécanique statistique, méthode d'interpolation.

Acknowledgments

I am very fortunate to be advised by Rüdiger Urbanke and Nicolas Macris. This thesis would not have been possible without their guidance and support.

Rüdiger has been a constant source of inspiration, knowledge, and creativity. I feel I am one of the few lucky persons whose adviser is a friend + guide + teacher put together. His doors are always open for discussions which range from coding theory to foosball tactics. He is always ready to get his hands dirty on my research problems. I would also like to acknowledge his efforts to improve my presentation and writing skills which were non-existent before. For all the above and many other reasons I would like to express my deepest gratitude. Finally, I would be very fortunate if I have absorbed a fraction of his joie de vivre.

I am deeply indebted to Nicolas for his tutelage in the first two years of my PhD. He was very patient to answer my questions on statistical mechanics, about which I had no clue before. The things I learned from him also played a crucial role in the later stages of my PhD. It is always a lot of fun to hear stories about great physicists from him. I also enjoyed his common man explanations of string theory, quantum mechanics and many other topics.

I am also very grateful to Emre Telatar. He was always ready to answer my questions. I would like to thank him for sharing his insights about polar codes and also for his comments on my thesis. It was a pleasure to discuss with Andrea Montanari. Even though we had only a few discussions, they were very helpful especially for my work on CDMA. The title of my thesis would have been entirely different if Erdal Arıkan had not invented polar codes. I would like to thank him for such a beautiful invention. I am very grateful to Emre, Andrea, Erdal, and Martin Hasler for agreeing to be in my thesis committee.

My stay in Lausanne was comfortable mainly due to the care taken by Muriel. She has helped me during every stage of my PhD, from finding an apartment to printing my thesis. I enjoyed the gossip sessions in the mornings with her and Françoise. I would also like to thank Damir for solving many many problems (some of them embarrassingly silly) I had with computers. I would like to thank him for setting up the cluster which made it possible to run massive simulations.

I would also like to thank my friends for the wonderful time I spent with

them at EPFL, in particular, Shrinivas, Dinkar, Abhishek, Sanket, Vish, Jérémie, Ayfer, Irina, Eleni, Nadine, Marius, Vojislav, Jasper, Olivier, Christine, Amin, Mohammed, Eren, Etienne, Sibi, Soheil, and many others. Special thanks to my collaborators Dinkar, Shrinivas, Eren, and Nadine.

Of course, the acknowledgments would not be complete without thanking my parents and brother. Without their love, support and sacrifices none of this would have been possible.

Last but not the least, I am very grateful to NCCR-MICS for funding my PhD.

Contents

1 Introduction

The two central topics of information theory are the efficient compression as well as the reliable transmission of data. The applications of these topics are everywhere; consider mobile communication, MP3 players, the Internet, CDs, or any other modern day digital technology.

Data compression can be either lossless or lossy. If the data consists of bank records or personal details we cannot afford to lose any information. In such cases, the compression is achieved by exploiting patterns and redundancies in the data. The commonly used zip format for storing data in computers is a good example of a lossless compression scheme.

Lossy compression, as it literally means, involves loss of information. Lossy compression is commonly used for multimedia data like images, music or video. The JPEG format for images and the MP3 format for music are good examples of lossy compression schemes. Why is lossy compression used at all? On the one hand this is due to necessity. Most physical phenomena are real valued. Storing them in digital form must involve some form of quantization, and hence, some loss of information. On the other hand, once lets say a video is captured in high-quality digital form, it can typically be compressed substantially with very little loss in perceptual quality. Technically speaking, given a source and a measure of quality, there is a trade-off between the storage requirement and the quality. This trade-off can be exploited to adapt a source coding scheme to a given situation. For example, a user with small bandwidth may be happy to get a low quality video if the alternative is not to be able to watch the video at all.

The second central topic of information theory is concerned with the transmission of data through a noisy medium. To make communication reliable in the presence of noise, the common procedure is to add redundancy to the data before transmission. The intended receiver only has access to a noisy version

of the data. However, if the redundancy is added in a clever way, then it is possible to reconstruct the original data at the receiver. Adding redundancy is called coding. Coding is a central part of any communication system; e.g., consider wired phones, mobile phones, or the Internet. Coding is also used for storage on CDs and DVDs to prevent data loss due to scratches or errors during the reading process.

1.1 Source and Channel Model

Shannon, in his seminal work [1], formalized the above problems of storage and communication and determined their fundamental limits. He provided a mathematical framework to study the problems systematically which lead to the advances in the past 50 years. The generality of his approach allows us to study even modern day scenarios, like mobile communication or ad-hoc networks. The basic model addressed by Shannon consists of a source, which generates the information, a sink which receives the information, and a channel, which models the physical transfer of information.

Figure 1.1: The basic communication scenario.

The source output is modeled as a realization of a random process $\{X_n\}$. Shannon showed that we can associate to this process an entropy, call it $H(X)$. The operational significance of entropy is that, for any $R > H(X)$, there exists a scheme to represent the source using only R bits per source symbol.

The channel is modeled by a conditional probability distribution. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output alphabet of the channel. The channel $W : \mathcal{X} \to \mathcal{Y}$ is a conditional probability distribution $W(y \,|\, x)$. When x is transmitted through the channel, the output at the receiver is the realization of a random variable $Y \in \mathcal{Y}$ distributed as $W(y \,|\, x)$. Shannon showed that in spite of this randomness, by intelligently adding redundancy, the data can be reproduced exactly at the receiver with high probability. He computed the capacity of a channel, $C(W)$, which quantifies the maximum rate at which reliable transmission of information is possible. In other words, for any $R < C(W)$, there exists a scheme which transmits R bits per channel use with vanishing error probability.

From the above discussion it is clear that if the source entropy is less than the channel capacity, i.e., if $H(X) < C(W)$, then the source can be reliably transmitted over the channel. The other crucial result of Shannon is that if the source entropy is larger than the channel capacity, then reliable communication is not possible. Therefore, if $H(X) > C(W)$ a loss of information is unavoidable and the best one can hope for is to minimize this loss. In [2]

Shannon introduced the notion of distortion which quantifies the dissimilarity between the source output and its compressed representation. He then characterized the rate-distortion trade-off $R(D)$. This trade-off describes the minimum rate required to achieve an average distortion D.

A crucial by-product of Shannon's work is that splitting the source and channel coding operations as shown in Figure 1.2 does not incur any loss in performance (if measured only in terms of the achievable rates).[1] The source coding module in Figure 1.2 compresses the source data to as low a rate as possible given a desired upper bound on the distortion. The output of the source encoder is passed through the channel coding module. This module adds redundancy to protect the data against noise. At the decoder, we first perform the channel decoding to remove the noise resulting from the transmission. Then the source is decoded, i.e., we reconstruct the source data from its compressed representation.

Figure 1.2: The simplified communication scenario.

For both source and channel coding to approach their fundamental limits, the blocklengths have to be large. This in turn has implications on the complexity. Therefore, for practical applications, we require schemes that operate with low space and computational complexity.

1.2 Existing Low-Complexity Schemes

Since Shannon's seminal work [1] the main goal has been to construct low-complexity coding schemes that achieve the fundamental limits. The complexity issues arise in two different contexts.

The first issue is the amount of memory required to store the code. This refers to the memory required for storing the mapping from the input of the encoder to its output. A code of rate R and blocklength N consists of 2^{NR} codewords of length N. A naive representation of such a code requires $O(N2^{NR})$ bits of memory, which is not practical. A significant progress in this respect was the result of Elias [3] and Dobrushin [4],[5, Section 6.2] which shows that linear codes are sufficient to achieve the capacity of an important class of channels, known as symmetric channels. Linear codes are subspaces of vector spaces. Hence, a linear code can be specified in terms of a basis of this subspace. This in turn can be done by describing RN vectors of length N. Therefore, the

[1] Joint source-channel code designs are preferable if delay is also a concern. However, they are typically more complicated than the modular approach.

resulting memory requirement is $O(N^2)$ bits. This is an exponential improvement over the general case. The corresponding result for source coding was shown by Goblick [6],[7, Section 6.2.3].

The second issue is the computational complexity of the encoding and decoding operations. In the following we give a brief history of some of the important developments. For a detailed history of channel coding we refer the interested reader to the excellent article by Costello and Forney [8].

1.2.1 Channel Coding

The initial research in coding was based on an algebraic approach. More precisely, the focus was on developing linear binary codes with large minimum distance (the smallest distance between any two distinct codewords) and good algebraic properties. Recall that the purpose of these codes is to recover data that is corrupted by noise during transmission. At the receiver, to simplify the decoding process, the following two-step approach was adopted. First, using the channel outputs, the value of individual bits are set to either 0 or 1, depending on whatever is more likely. This is called hard-decision. At the second stage the code constraints are used. The decoder selects the codeword that is closest to the received word (after hard-decision). Therefore, it is desirable to employ codes with a large minimum distance, since they will allow the correction of a large number of errors. This is true since as long as the number of errors is less than half the minimum distance, the above procedure is guaranteed to output the correct codeword.

The first algebraic codes were developed by Hamming and are named after him. Hamming codes are single error correcting codes and they are optimal in the sense of sphere packings.[2] Other important algebraic codes are Golay codes, BCH codes [9, 10], Reed-Muller codes [11, 12] and Reed-Solomon codes [13]. For all these codes efficient algebraic decoding algorithms are known. These codes are prominently used today in CDs, DVDs and modems.

As mentioned in the previous section, to achieve optimal performance one has to consider large blocklengths. The improvement in computational resources made it feasible to consider larger and larger codes in practice. But the previously discussed algebraic codes either have vanishing rate or vanishing relative distance (ratio of minimum distance and blocklength) for increasing blocklengths.

Product codes, introduced by Elias [14], were the first constructions which asymptotically (in the blocklength) achieved both non-vanishing relative distance and rate. The idea is to construct large codes by combining two or more codes of smaller length. Consider two codes $\mathtt{C}_1$ and $\mathtt{C}_2$ of length n_1 and n_2. Each codeword of the product code can be viewed as an $n_1 \times n_2$ matrix such that each column is a codeword of $\mathtt{C}_1$ and each row is a codeword of $\mathtt{C}_2$. A

[2] In more detail, the entire Hamming space consisting of 2^N (N is the blocklength) N-tuples can be covered with spheres of radius $(d-1)/2$ (d is the minimum distance) placed at every codeword.

low-complexity but sub-optimal decoding algorithm is to decode each row and column separately. However, the performance of such a combination was far below the capacity of the channel.

Code concatenation, introduced by Forney [15], is another construction based on combining codes. The idea is to first encode the data using $\mathtt{C}_1$ and then encode the resulting output with $\mathtt{C}_2$. Forney showed that for any rate below the capacity, by an appropriate choice of the two component codes, the error probability can be made to decay almost exponentially with a decoding algorithm that has polynomial complexity.

The next big step in improving the decoding performance came from considering probabilistic decoding. In typical scenarios the capacity is significantly reduced by making hard-decisions at the decoder. E.g., for Gaussian channels, close to half the power is lost due to this first step. The idea of probabilistic decoding is to make use of the channel outputs directly in the decoding algorithm and to avoid this loss.

The first class of codes well suited for probabilistic decoding were convolutional codes, introduced by Elias in [3]. The code structure enabled the development of efficient decoding algorithms. In particular the Viterbi algorithm [16], which minimizes the block error probability, and the BCJR algorithm [17], which minimizes the bit error probability, both operate with complexity which is linear in the blocklength. For any rate strictly less than the capacity of the channel one can show that there exist convolutional codes whose probability of error vanishes exponentially in the "constraint length". However, the complexity of the decoding algorithm also scales exponentially with the constraint length. Therefore, for practical purposes Fano's sequential decoding algorithm [18] was considered. For rates less than the "computational cutoff rate", the complexity of this algorithm is linear in the blocklength and independent of the constraint length. The cutoff rate is a rate that can be computed easily and that is strictly smaller than the capacity. For rates above the cutoff rate, the complexity is unbounded. This lead to the belief that, using practical algorithms, rates above the cutoff rate could not be achieved.

Another class of codes which were introduced during the 60's were low-density parity-check (LDPC) codes [19]. As the name suggests, the parity-check matrices of these codes have very few non-zero entries. In fact, these matrices have a constant number of non-zero entries per row and column. Gallager showed that these codes have a non-zero relative distance. He also proposed a low-complexity iterative decoding algorithm. Unfortunately, due to the lack of computational resources at that time, the power of these codes and the decoding algorithms was not realised.

The invention of turbo codes by Berrou, Glavieux and Thitimajshima [20] was a breakthrough in the practice of coding. Turbo codes achieved rates close to the capacity, and far above the cutoff rate, using a linear complexity decoding algorithm. The code is constructed by concatenating two convolutional codes but with a random bit interleaver in between. The decoding algorithm operates in iterations. In each iteration the BCJR algorithm is performed on

each of the component codes and the reliabilities are exchanged. Since the complexity of the BCJR algorithm is linear in the blocklength, the resulting decoding algorithm is also of linear complexity. The original turbo code of [20] matched the error probability of the best existing schemes which were operating with twice the power. The interleaver and the iterative property of the decoding algorithm are the two crucial components which are recurrent in the capacity achieving schemes constructed later.

MacKay and Neal constructed codes based on sparse matrices [21] and observed that they perform very well using a low complexity belief propagation algorithm. It was later noticed that these codes were a special case of LDPC codes and that the decoding algorithm is equivalent to the probabilistic decoding suggested by Gallager. Around the same time Sipser and Spielman [22] constructed expander codes and came up with a simple decoding algorithm that could correct a linear fraction of adversarial errors.

Wiberg, Loeliger and Kötter [23, 24] unified turbo codes and LDPC codes under the framework of codes on graphs. Within this framework the turbo decoding algorithm, the belief propagation algorithm of MacKay and Neal and the probabilistic decoding of Gallager turn out to be different incarnations of the same algorithm. The framework also provided a bridge between sparse graph codes and other fields like machine learning, statistical mechanics and computer science.

The success of turbo codes and the subsequent rediscovery of LDPC codes revived the interest in LDPC codes and message passing algorithms. Some of the first and important contributions to the analysis of message-passing algorithms was done in a series of papers by Luby, Mitzenmacher, Shokrollahi, Spielman and Stemann [25, 26, 27, 28]. In [25, 27], the authors analyzed a suboptimal decoder known as "peeling decoder" for the binary erasure channel (BEC). They constructed codes for the BEC which achieve capacity using the peeling decoder. Later, in [29], the peeling decoder was formulated as a process on a tree. The analysis of the new formulation is significantly simpler than that of [25, 27].

In [30], Richardson and Urbanke developed density evolution which generalizes the analysis of the BEC [29] to any symmetric channel and a class of algorithms known as message-passing algorithms. This class includes the important belief propagation algorithm. Combining density evolution for belief propagation with optimization techniques, codes that approach capacity of Gaussian channel to within 0.0045dB [31] were constructed. However, unlike the BEC, no capacity achieving codes are known for general channels.

Until now, many turbo and LDPC codes have been proposed which empirically achieve rates close to capacity for various channels. However, none of these codes are proven to achieve capacity for channels other than the BEC. In this thesis we discuss polar codes. Polar codes, recently introduced by Arıkan [32], are a family of codes that provably achieve the capacity of symmetric channels with "low encoding and decoding complexity". This settles the long standing open problem of achieving capacity with low complexity.

1.2.2 Source Coding

Let us start with the lossless source compression problem. Huffman coding [33] achieves the optimal compression for memoryless sources with known source statistics. The Lempel-Ziv algorithm [34, 35] solves the problem for the wide class of stationary ergodic sources. It even provides compression guarantees for individual sequences. The algorithm is easily implementable and it is widely used in commercial compression software, e.g., gzip.

The lossy compression problem is more difficult and the progress in this field is less spectacular. The problem was formally defined by Shannon in 1959 [2]. The most commonly studied problem is the compression of a binary symmetric source (BSS). Here we recall some of the past and current approaches. In his paper [2], Shannon suggested to use error correcting codes and he even provided some achievable rate-distortion pairs using Hamming codes. Since then the progress in lossy source coding closely followed in the footsteps of channel coding.

Trellis based quantizers [36] were perhaps the first "practical" solution for the lossy compression problem. Their encoding complexity is linear in the blocklength of the code (Viterbi algorithm). Like in channel coding, for any rate strictly larger than $R(D)$ the gap between the resulting distortion and the design distortion D vanishes exponentially in the constraint length and as mentioned before, the complexity scales exponentially with the constraint length.

The success of sparse graph codes combined with low-complexity message-passing algorithms for channel coding spurred the interest of many researchers to investigate their performance for lossy source compression. In this respect, Matsunaga and Yamamoto [37] showed that if the degrees of an LDPC ensemble are chosen as large as $\Theta(\log(N))$, where N is the blocklength, then this ensemble saturates the rate-distortion bound if optimal encoding is employed. Even more promising, Martininian and Wainwright [38] proved that properly chosen MN codes (proposed by MacKay and Neal [21] for channel coding) with *bounded* degrees are sufficient to achieve the rate-distortion bound under optimal encoding.

Much less is known about the performance of sparse graph codes under *message-passing* encoding. Initial observations suggested that LDPC codes combined with message passing algorithms do not perform well in practice. To understand why, the authors in [39] considered binary erasure quantization (BEQ), the source-compression equivalent of the channel coding problem for the BEC. For this problem, they show that LDPC-based quantizers fail if the parity check density is $o(\log(N))$ but that properly constructed low-density generator-matrix (LDGM) based quantizers combined with message-passing encoders are optimal. The latter claim is based on a duality relationship between the BEQ and the BEC. Using this duality, the problem of finding optimal codes for the former problem is mapped to the problem of finding capacity achieving codes for the BEC. The code for the BEQ is then given by

the dual of the code designed for the BEC with appropriate erasure probability.

The above result is significant in two respects. First, it showed that LDGM codes are suitable for source coding. Second, the source coding problem is mapped to a channel coding problem. This made it possible to use the machinery developed for channel coding. Unfortunately the duality relationship between the BEQ and the BEC is very specific and it does not extend to other sources like the BSS.

Regular LDGM codes were considered for the BSS in [40] and [41]. Using non-rigorous methods from statistical physics the authors argue that these codes approach the rate-distortion bound for large degrees using optimal encoding. In [40], it was empirically shown that LDGM codes with degree 2 check nodes perform well under a variant of belief propagation (BP) algorithm known as reinforced BP (RBP).[3] The encoding algorithm in [41] is another variant known as belief propagation inspired decimation (BID).

Let us now discuss briefly why BP itself does not work well for source coding and why we need to employ variations of BP. The reason for the failure of standard BP is the abundance of solutions. Typically, for a source word there are many codewords that, if chosen, result in similar distortion. Let us assume that these "candidate" codewords are roughly uniformly spread around the source word to be compressed. It is then clear that a message-passing decoder which operates locally can easily get "confused," producing locally conflicting information with regards to the "direction" towards which one should compress.

A standard way to overcome this problem is to combine the message-passing algorithm with "decimation" steps. This works as follows; first run the iterative algorithm for a fixed number of iterations and subsequently decimate a small fraction of the bits. More precisely, this means that for each bit which we decide to decimate we choose a value. We then remove the decimated variable nodes and adjacent edges from the graph. One is hence left with a smaller instance of essentially the same problem. The same procedure is then repeated on the reduced graph and this cycle is continued until all variables have been decimated. In BID, the standard BP plays the role of the message-passing algorithm.

Another approach to make message-passing algorithms work is to use reinforcement of the "priors". Each bit is associated with a prior probability, equivalent to the priors obtained from channel observations in channel coding. In source coding, since there is no observation to start with, we initialize the priors to uniform distribution over 0 and 1. The algorithm reinforces these priors during the encoding process. More precisely, we run the message-passing algorithm for a few iterations and then update the priors using the current messages. The message-passing algorithm is then started with the new priors. This process is continued until we build strong biases in the priors. These pri-

[3]The author refers to this algorithm as Thouless-Anderson-Palmer approach, as it is known by that name in statistical physics literature.

ors are then used to decide the values of the bits. This is the principle behind RBP.

The authors in [41] observe that BID as well as RBP do not perform well when the check node degree is larger than 2. They argue that the reason for this failure is the linearity of the check node. In order to accommodate check nodes with larger degrees, the authors in [42] replace parity-check constraints with non-linear constraints. The resulting code is similar to the k-SAT problem in computer science. The encoding was therefore done using survey propagation inspired decimation (SID), an algorithm used for finding solutions of a k-SAT formula [43]. The empirical results show that SID combined with non-linear codes achieve good performance.

In channel coding, the power of LDPC codes was fully realized only after considering irregular codes. The code design for channel coding uses density evolution for BP combined with optimization methods. The algorithms used in source coding, namely RBP, BID, as well as SID, combine message-passing with some additional steps as mentioned above. These additional steps make their analysis difficult. In fact, the analysis of these algorithms is still an unsolved problem. This makes the design and optimization of codes a challenging task.

Motivated by the construction in [39], the authors in [44] consider those LDGM codes whose duals (LDPC) are optimized for the binary symmetric channel (BSC). More precisely, a source code for distortion D is taken to be the dual of a channel code designed for BSC(p) where $p = h_2^{-1}(1 - D)$ and $h_2(\cdot)$ is the binary entropy function. Surprisingly, even though no duality relationship is known between the two problems, the approach was successful. They empirically show that by using SID one can approach closely the rate-distortion bound. Recently, in [45] it was experimentally shown that even using BID it is possible to approach the rate-distortion bound closely. The key to make BID work is to properly choose the code and the parameters of the algorithm.

The current state-of-the-art code design is thus based on the heuristic approach of designing an LDPC code for a suitably defined channel and then taking its dual LDGM code. Such an approach does not extend to sources other than the BSS. In addition to the heuristic argument, the code design relies on finding capacity achieving codes for channel coding which itself is an open problem.

1.3 Polar Codes

Polar codes, introduced by Arıkan in [32], are the first provably capacity achieving codes for any symmetric binary-input discrete memoryless channel (B-DMC) that have low encoding and decoding complexity. The complexity of these algorithms scales like $O(N \log N)$, where N is the block length of the code.

The origin of the idea of polar codes can be traced back to Arıkan's earlier work on improving the computational cutoff rate of channels [46]. In this paper, Arıkan considers applying a simple linear transform to the channel inputs before transmission and a successive cancellation decoder at the output. Let W be the original channel and let W_1 and W_2 be the channels seen by the bit that is decoded first and second respectively. Such a transformation results in a larger average (over W_1 and W_2) cutoff rate compared to that of W.

The idea of polar codes is based on repeating the above process "recursively". The recursive process is equivalent to applying a linear transform on a larger number of bits at the encoder. At the decoder, the bits are decoded successively in a particular order. As a result the effective channels seen by some of the bits are better than W and some are worse. Interestingly, as the blocklength increases, these effective channels tend towards either a completely noisy channel or a clean channel with the fraction of clean channels approaching the capacity of W. Arıkan refers to this phenomenon as channel polarization. This suggests a simple scheme where we fix the inputs to the channels that are bad and transmit reliably over the clean channels without any coding. The rate of such a scheme approaches the capacity of the channel.

However, for the scheme to be practical it remains to show that the recursive transformation at the encoder and the successive cancellation decoder can be implemented with low-complexity. Using a Fast-Fourier-like transform Arıkan showed that both the encoding and decoding operations can be implemented with $O(N \log N)$ complexity. Effectively, the scheme achieves capacity using low-complexity algorithms.

Arıkan's input transformation matrix is given by $G_2^{\otimes n}$ where $G_2 = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$, and "$\otimes n$" denotes the n-th Kronecker product. Choosing some of the input bits to transmit information and fixing the rest is equivalent to choosing the rows of $G_2^{\otimes n}$ that form the generator matrix of the code. Such codes are referred to as polar codes. Polar codes are closely related to RM codes. The generator matrix of RM codes is also constructed from the rows of $G_2^{\otimes n}$. The crucial difference between the two codes lies in the rule used for picking the rows of their generator matrices. RM codes correspond to the choice which maximizes the minimum distance of the code. For polar codes, on the other hand, the choice is dependent on the channel that is used. The choice optimizes the performance under successive cancellation decoding.

1.4 Contribution of this Thesis

Most of this thesis is centered around polar codes. For the sake of completeness, we also include our results in the area of LDPC codes and code-division multiple access (CDMA) communication. But in order not to disturb the flow, we placed these two extra topics in the last two chapters.

Our contributions in the area of polar codes can be classified into two main categories. 1) We construct the first known low-complexity coding schemes

that are optimal for a variety of important problems in information theory. These include the lossy source coding problem as well as several problems involving source and channel coding. 2) We expand the notion of polar codes to include more general transforms. This allows us to construct polar codes with considerably better block error probabilities. In Sections 1.4.1 and 1.4.2 we review these contributions in more detail. In Section 1.4.3 we outline our contributions in the areas of LDPC codes as well as CDMA communication.

1.4.1 Polar Codes: Low-Complexity Schemes with optimal performance

The first problem we consider is lossy source coding. Recall from the previous section that currently there are a variety of low-complexity schemes for this problem, but that none of them achieve the rate-distortion bound. Our main contribution is to show that suitable polar code constructions achieve Shannon's rate-distortion bound for an important class of sources. This class includes in particular the well studied BSS. The codes are not only provably optimal, but as the simulations suggest, they perform well in practice. For the BSS with Hamming distortion the performance closely approaches the optimal trade-off for reasonable blocklengths, lets say 10000.

Let us review the source coding problem. Given a source sequence, the job of the encoder is to quantize it to a codeword with the aim of minimizing the distortion. In our setting of using polar codes, the encoding algorithm combines successive cancellation with "randomized rounding"; the latter is the crucial new ingredient. Randomized rounding refers to determining a variable by *sampling* according to its posterior probability rather than choosing the value corresponding to the largest probability (MAP rule). It is exactly this choice which makes the analysis possible. Randomized rounding can be viewed as picking one codeword from a set of codewords having a similar distortion.

Let us get back to the general description of source coding. Once the encoder chooses the codeword, it describes it to the decoder using an index. At the decoder, the codeword is reconstructed from the index. This operation is similar to the encoding operation of channel coding. Therefore, both the encoding and decoding operations can be implemented with $O(N \log N)$ complexity.

We also consider source coding problems with side information as well as channel coding problems with side information. The optimal codes for these problems require nested structures. For the former problem we require a good channel code which can be partitioned into cosets of a good source code. For the latter problem the role of the channel and source codes are reversed. In Chapter 4, we provide a brief history of the current approaches. These approaches are based on creating the required nested structures using sparse graph codes. However, as yet there do not exist any low-complexity schemes that achieve the optimum performance. Our main result is to show that ap-

propriately designed polar codes achieve optimal performance for each of these problems. We exploit the natural nested structure existing in polar codes to prove this result. The encoding and decoding operations are performed using successive cancellation algorithms. Hence they can be implemented with $O(N \log N)$ complexity. We also discuss applications to multi-terminal problems like the Slepian-Wolf problem and the one-helper problem.

We show that polar codes achieve capacity of asymmetric channels, degraded broadcast channels as well as multiple access channels. The common feature in all these problems is that we require codes which induce non-uniform marginals over the channel inputs. We map the code design for each of the afore-mentioned problems to the code design of an appropriately defined non-binary (q-ary) input channel. The required binary code with non-uniform marginals is then obtained by collapsing the q symbols to the binary alphabet using a surjective map. Once again, these are the first low-complexity schemes that achieve capacity. We also discuss application of polar codes for compound channels and show that the achievable rate is strictly less than the compound capacity.

1.4.2 Improved Polar Codes

As we will explain in detail in Chapter 2, for any fixed rate R below capacity and a sequence of polar codes of rate R, the block error probability decays at best as $O(2^{-(N)^{1/2}})$ [47]. Similarly, for the source coding problem, we show that the resulting distortion approaches the target distortion D as $D + O(2^{-(N)^{1/2}})$. Notice that the exponent of the blocklength N is $\frac{1}{2}$ in both the cases. However, from the random coding argument we know that there exist codes having exponent 1 under MAP decoding. We therefore ask the following two questions. Is it the decoding algorithm or the code itself that results in a smaller exponent? Are there polar codes which achieve exponent 1? With respect to the first question we show that the reason for the smaller exponent is the code itself. To answer the second question we explore other polar-like code constructions. This is a joint work with Şaşoğlu and Urbanke [48].

Let us explain this in more detail. Recall that the rows of the generator matrix of a polar code are chosen from the matrix $G_2^{\otimes n}$. Our generalization is based on choosing the rows from $G_\ell^{\otimes n}$ for a general $\ell \times \ell$ matrix G_ℓ. Our main result is to show that there exist families of codes that have an exponent arbitrarily close to 1 using SC decoding. The recursive structure of these codes enables to implement the SC algorithm with complexity $O(2^\ell N \log N)$.

In the process we show that almost all matrices are suitable for channel polarization. For a given matrix G_ℓ we characterize its exponents both for channel as well as source coding in terms of its "partial distances". These distances can be computed easily for a given matrix. With this characterization, the problem of designing polar codes with large exponent is transformed to a simpler problem of finding $\ell \times \ell$ matrices with some properties. We also show how to transform a matrix that achieves large exponent for source coding into

a matrix that achieves the same exponent for channel coding. Effectively, the two problems of finding good matrices for channel and source coding boil down to a single problem.

The characterization in terms of the partial distances also allows us to use tools from algebraic coding to design good matrices. Using tools like the Gilbert-Varshamov bound and Hamming bound we obtain bounds on the best possible exponent for any matrix G_ℓ. We also show that for $\ell < 15$ no matrix has an exponent larger than $\frac{1}{2}$. This is discouraging because it implies that to improve the asymptotic performance of polar codes one has to consider very large block lengths and also larger complexity. We further exploit the relationship with algebraic coding to provide an explicit family of matrices based on BCH codes. This family achieves exponent tending to 1 for large ℓ and an exponent larger than $1/2$ for $\ell = 16$, which is shown to be the best possible exponent for $\ell = 16$.

Even though polar codes promise good block error probability their performance at practical blocklengths, in particular for channel coding, is not record breaking. In fact this was the reason which prompted us to explore codes with larger exponent. Another approach to improve the performance is to investigate different decoding algorithms. We empirically show that the performance of polar codes improves by considering belief propagation and some variants of it.

1.4.3 Rigorous Results for Graphical Models

Polar codes form the core of this thesis. But we also include our results on LDPC codes and CDMA communication. Both these problems are conveniently presented in a graphical way. These graphs play an important role in their analysis. This explains the title for this section.

Exchange of Limits

As mentioned in Section 1.2.1, prior to the invention of polar codes, the best known low-complexity coding schemes were based on LDPC codes combined with message-passing decoding. For such a combination, both the code design and the error probability analysis is based on density evolution (DE) [30].

DE analyzes the performance of the algorithm when the size of the graph is infinite. As a consequence, the local neighborhood is a tree and the messages seen on different edges are independent. The analysis is done by tracking the evolution of the messages on a tree. Let $P_b(\ell, N)$ denote the probability of error for ℓ iterations of a code of blocklength N. Using DE, we can compute $\lim_{N\to\infty} P_b(\ell, N)$ for any fixed ℓ. To quantify the performance of the algorithm for unlimited number of iterations, the DE equations are analyzed in the limit of $\ell \to \infty$. Mathematically, this corresponds to computing the limit

$$\lim_{\ell\to\infty} \lim_{N\to\infty} P_b(\ell, N).$$

In practice the size of the code is fixed and we perform a large number of iterations (typically hundreds and limited only by the computational resources at our disposal) at the receiver. It has been noticed in practice that increasing the number of iterations only improves the performance of the algorithm. Therefore, it is interesting to understand the behavior of the algorithm for unlimited number of iterations. Mathematically, this corresponds to computing the limit

$$\lim_{N\to\infty}\lim_{\ell\to\infty} P_b(\ell, N).$$

The tree assumption is not true in this limit and hence the DE equations are no longer valid. The correlation between the messages on various edges of a node makes the analysis difficult. Our main result is to show that under some suitable technical conditions, which are satisfied by many message-passing decoders, the predictions given by the DE analysis are still accurate in the regime where the decoding is successful. Mathematically, we show that

$$\text{if} \quad \lim_{\ell\to\infty}\lim_{N\to\infty} P_b(N, \ell) = 0, \qquad \text{then} \quad \lim_{N\to\infty}\lim_{\ell\to\infty} P_b(N, \ell) = 0.$$

This suggests that in the regime where the DE analysis predicts zero error probability, the correlation between the variables decays rapidly with the distance and it does not degrade the performance of the algorithm.

Capacity of Binary input CDMA

The second problem that we consider is CDMA communication over an additive white Gaussian noise channel using randomly generated spreading sequences. CDMA is used in multi-user scenarios for simultaneous access of a common receiver in a wireless medium. A prominent example is the uplink segment of cellular communication. This problem has been studied in great detail when the input distribution of the users is Gaussian [49], which corresponds to the information theoretic limits. However, in practice we use constellations like binary phase-shift keying (BPSK), which is far from being Gaussian. In contrast to the Gaussian input case, not much is known for these inputs. The analysis used for Gaussian inputs, based on the spectrum of large random matrices, is not applicable for other inputs.

For simplicity we treat the binary input case (BPSK), but the results can be extended to other input constellations. We study the capacity of the system, which is defined as the maximum achievable rate per user. Our approach is based on treating the CDMA system as a random Ising model and using tools developed for the analysis of random spin systems in statistical mechanics. The main result is a tight upper bound on the capacity in the limit of large number of users. The bound matches with the conjectured capacity which is based on the replica method. We explain this conjecture in more detail in Chapter 8. We show that in the limit of large number of users, the capacity depends only on the power of the spreading sequences and is independent of

their distribution. We also show that the capacity concentrates with respect to the randomness in the spreading sequences in the large-user limit. In other words, the capacity is the same for almost all spreading sequences.

The mathematical methods we use are quite powerful and are easily applicable to other scenarios. We demonstrate this by extending the results to the case of users with unequal power constraints, channels with colored noise, and other input constellations as well. We also derive the upper bound for Gaussian input distribution which matches with the existing result and thus strengthening our belief that the bounds are tight for other scenarios too.

1.5 Organization of the Thesis

Chapters 2 to 6 are related to polar codes. Chapter 7 is about LDPC codes and Chapter 8 deals with the CDMA problem. The last two chapters are self-contained and are not related to the rest of the thesis.

Chapter 2 reviews Arıkan's work on polar codes [32]. We explain the main ideas behind the code construction and the successive cancellation (SC) decoding algorithm. We then show that polar codes achieve the capacity for symmetric channels using SC decoding. The complexity of the encoding and decoding algorithms is shown to be $O(N \log N)$, where N is the blocklength. We also include a discussion about the complexity of the code construction. The last section provides some simulation results.

Chapter 3 considers the lossy source coding problem. We introduce the SC algorithm combined with randomized rounding that is used for the encoding operation. We then show that properly designed polar codes using the SC encoding algorithm achieve the Shannon's rate-distortion bound for symmetric sources. This is perhaps one of the most important contributions of this thesis and it solves a long standing problem. Some simulation results for binary symmetric source are provided. The chapter ends with a discussion on the connection with channel coding.

Chapter 4 considers other communication scenarios including the source coding with side information as well as the channel coding with side information problems. We show that polar codes again achieve optimal performance for these problems using the low-complexity SC algorithm. Applications to other problems like the lossless compression problem, the Slepian-Wolf problem and the one-helper problem are also discussed. We then show that the capacities of asymmetric channels, degraded broadcast channels, and multiple access channels can be achieved using non-binary polar codes. Simulation results are provided for some of the problems.

Chapter 5 deals with the generalization of polar codes to $\ell \times \ell$ matrices.

We explain the encoding and SC decoding in this case. We provide the formal definition for the exponent of channel and source coding. We then characterize the exponents in terms of partial distances. A duality relationship between the two exponents is also provided. Using bounds from classical coding, bounds for the best possible exponent are derived. The chapter ends with a construction of a family of matrices based on BCH codes.

Chapter 6 deals with the relationship between polar codes and RM codes. We show that the minimum distance of a polar code can scale at most as $O(N^{1/2})$. We also discuss briefly about the relationship of SC decoder to Dumer's recursive algorithm. The empirical performance of polar codes under some variants of belief propagation decoding is also shown. We then treat the compound capacity of polar codes. Some open problems and future research directions are also discussed.

Chapter 7 is the first of the two chapters dealing with graphical models. In this chapter, we consider communication using LDPC codes over binary input symmetric channels. We show that the density evolution analysis is indeed correct for finite graphs and for noise values below the channel threshold. Our results are split into two categories based on the variable degree. We first discuss the results for large variable degrees (larger than 5) where we show the exchange of limits for a large class message passing algorithms. We then discuss the results for variable degree 3 and 4 which are based on different methods.

Chapter 8 considers CDMA communication using binary inputs and random spreading sequences. We provide concentration results with respect to the spreading sequences. We also show that the capacity is independent of the exact distribution of the spreading sequence. We provide an upper bound to the capacity which matches with the conjectured capacity and hence believed to be tight. We end this chapter with extensions to other channels and input distributions.

1.6 Notation

Throughout the manuscript, we stick to the following notation. We use $W : \mathcal{X} \to \mathcal{Y}$ to represent a binary input discrete memoryless channel (B-DMC) with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. We use upper case letters U, X, Y to denote random variables and lower case letters u, x, y to denote their realizations. Let $\bar{u}$ denote the vector $(u_0, \ldots, u_{N-1})$ and let u_i^j for $i < j$ denote the subvector $(u_i, \ldots, u_j)$. We use $u_{i,e}^j$ and $u_{i,o}^j$ to denote the subvector of u_i^j consisting of only the even indices and the odd indices respectively. For any set F, $|F|$ denotes its cardinality. Let u_F denote $(u_{i_1}, \ldots, u_{i_{|F|}})$, where $\{i_k \in F : i_k \leq i_{k+1}\}$. We use the equivalent notation for random vectors too.

All the logarithms are to the base 2. We use $h_2(\cdot)$ to denote the binary entropy function, i.e., $h_2(x) = -x \log x - (1-x)\log(1-x)$.

We use BEC(ϵ) to denote the binary erasure channel with erasure probability ϵ, BSC(p) to denote the binary symmetric channel with flip probability p and BAWGNC(σ) to denote an additive white Gaussian noise channel with noise variance σ^2 whose input is restricted to $\{\pm 1\}$ (BPSK). We let Ber(p) denote a Bernoulli random variable with $\Pr(1) = p$ and $\Pr(0) = 1 - p$.

1.7 Useful Facts

Let $I(W) \in [0,1]$ denote the mutual information between the input and output of W with uniform distribution on the inputs, i.e.,

$$I(W) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} W(y \,|\, x) \log \frac{W(y \,|\, x)}{\frac{1}{2}W(y \,|\, 0) + \frac{1}{2}W(y \,|\, 1)}. \tag{1.1}$$

Let $Z(W) \in [0,1]$ denote the Bhattacharyya parameter of W, i.e.,

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \tag{1.2}$$

Also, let $P_e(W)$ denote the probability of error for a uniform distribution over the input $\{0,1\}$, i.e.,

$$P_e(W) = \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} W(y \,|\, x) \mathbb{1}_{\{W(y \,|\, x) \leq W(y \,|\, x \oplus 1)\}}. \tag{1.3}$$

Definition 1.1 (Symmetric B-DMC)**.** *A B-DMC $W : \mathcal{X} \to \mathcal{Y}$ is said to be symmetric if there exists a permutation $\pi : \mathcal{Y} \to \mathcal{Y}$ such that $\pi = \pi^{-1}$ and $W(y|0) = W(\pi(y)|1)$ for all $y \in \mathcal{Y}$.*

Symmetric channels play an important role in information and coding theory. Some of the most commonly studied channels like binary symmetric channel (BSC) and binary erasure channel (BEC) are both symmetric. For such channels the symmetric mutual information $I(W)$ is equal to their capacity. An important and useful property of symmetric B-DMCs is the following.

Theorem 1.2 (Symmetric B-DMC as BSCs [50])**.** *Any symmetric B-DMC can be decomposed into BSCs.*

The above theorem implies that for any symmetric B-DMC W there exists a set of BSCs, say BSC(ϵ_1),...,BSC(ϵ_K), and a probability distribution $p_1, \ldots, p_K$ over the set $\{1, \ldots, K\}$ such that using the channel W is equivalent to using a BSC(ϵ_i), with index i known only to the decoder, with probability p_i. Since the two channels are equivalent we have

$$I(W) = \sum_{i=1}^{K} p_i I(\mathrm{BSC}(\epsilon_i)), \tag{1.4}$$

$$Z(W) = \sum_{i=1}^{K} p_i Z(\mathrm{BSC}(\epsilon_i)), \tag{1.5}$$

$$P_e(W) = \sum_{i=1}^{K} p_i P_e(\mathrm{BSC}(\epsilon_i)). \tag{1.6}$$

Definition 1.3 (Symmetrized B-DMC)**.** *For any B-DMC $W : \mathcal{X} \to \mathcal{Y}$, the symmetrized B-DMC $W_s : \mathcal{X} \to \mathcal{Y} \times \mathcal{X}$ is defined as*

$$W_s(y, z \,|\, x) = \frac{1}{2} W(y \,|\, z \oplus x).$$

Lemma 1.4 (Channel Symmetrization)**.** *For any B-DMC W, there exists a symmetric B-DMC W_s such that*

$$I(W) = I(W_s), \quad Z(W) = Z(W_s), \quad P_e(W) = P_e(W_s).$$

Proof. Let W_s be the symmetrized B-DMC of Definition 1.3. The channel W_s is symmetric where the permutation π is given by $\pi(y, z) = (y, z \oplus 1)$. It is easy to check that $I(W) = I(W_s)$, $Z(W) = Z(W_s)$ and $P_e(W) = P_e(W_s)$. □

It is intuitive that $I(W)$ is close to 0 if and only if $Z(W)$ is close to 1 and vice-versa. The following lemma makes this intuition rigorous.

Lemma 1.5 ($I(W)$ and $Z(W)$[32])**.** *For any B-DMC W,*

$$I(W) + Z(W) \geq 1,$$
$$I(W)^2 + Z(W)^2 \leq 1.$$

Proof. Note that Lemma 1.4 implies that it is sufficient to prove the inequalities for symmetric B-DMCs. Theorem 1.2 combined with (1.4) and (1.5) implies that it is sufficient to show the first inequality for BSC(ϵ). The claim follows by checking that $2\sqrt{\epsilon\bar{\epsilon}} > h_2(\epsilon)$.

For the BSC(ϵ), the second inequality follows from $h_2(\epsilon) \geq 2\epsilon$ which implies $(1 - h_2(\epsilon))^2 \leq (1 - 2\epsilon)^2 = 1 - 4\epsilon(1 - \epsilon)$. The result for symmetric B-DMCs follows from the concavity of $f(x) = \sqrt{1 - x^2}$ for $x \in [0, 1)$ and Theorem 1.2. □

Lemma 1.6 ($P_e(W)$ and $Z(W)$ [51])**.** *For any B-DMC W,*

$$\frac{1}{2}(1 - \sqrt{1 - Z(W)^2}) \leq P_e(W) \leq Z(W).$$

Proof. The upper bound follows from the inequality $\mathbb{1}_{\{W(y|x) \leq W(y|x\oplus 1)\}} \leq \sqrt{\frac{W(y|x\oplus 1)}{W(y|x)}}$. Note that Lemma 1.4 implies that it is sufficient to show the lower bound for symmetric B-DMCs. It is easy to check that the lower bound is satisfied with an equality for BSC. The result for symmetric B-DMCs follows from the convexity of $f(x) = \frac{1}{2}(1 - \sqrt{1 - x^2})$ for $x \in (-1, 1)$ and Theorem 1.2. □

Definition 1.7 (Degradation)**.** *Let $W_1 : \{0,1\} \to \mathcal{Y}_1$ and $W_2 : \{0,1\} \to \mathcal{Y}_2$ be two B-DMCs. We say that W_1 is degraded with respect to W_2, denoted as $W_1 \preceq W_2$, if there exists a DMC $W : \mathcal{Y}_2 \to \mathcal{Y}_1$ such that*

$$W_1(y_1 \,|\, x) = \sum_{y_2 \in \mathcal{Y}_2} W_2(y_2 \,|\, x) W(y_1 \,|\, y_2).$$

Lemma 1.8 (Bhattacharyya and Degradation)**.** *Let W_1 and W_2 be two B-DMCs and let $W_1 \preceq W_2$. Then,*

$$Z(W_1) \geq Z(W_2).$$

Proof.

$$\begin{aligned}
Z(W_1) &= \sum_{y_1 \in \mathcal{Y}_1} \sqrt{W_1(y_1 \,|\, 0) W_1(y_1 \,|\, 1)} \\
&= \sum_{y_1 \in \mathcal{Y}_1} \sqrt{\sum_{y_2 \in \mathcal{Y}_2} W_2(y_2 \,|\, 0) W(y_1 \,|\, y_2) \sum_{y_2 \in \mathcal{Y}_2} W_2(y_2 \,|\, 1) W(y_1 \,|\, y_2)} \\
&\overset{(a)}{\geq} \sum_{y_1 \in \mathcal{Y}_1} \sum_{y_2 \in \mathcal{Y}_2} \sqrt{W_2(y_2 \,|\, 0) W(y_1 \,|\, y_2) W_2(y_2 \,|\, 1) W(y_1 \,|\, y_2)} \\
&= \sum_{y_2 \in \mathcal{Y}_2} \sqrt{W_2(y_2 \,|\, 0) W_2(y_2 \,|\, 1)} \sum_{y_1 \in \mathcal{Y}_1} W(y_1 \,|\, y_2) = Z(W_2),
\end{aligned}$$

where (a) follows from Cauchy-Schwartz inequality. □

Channel Coding : A Review

2

In this chapter we discuss the construction of polar codes for channel coding. It is based entirely on the work of Arıkan [32]. This chapter lays the foundation and sets the notation for the rest of this thesis. For the sake of brevity we skip some of the proofs.

The polar code construction is based on the following observation: Let

$$G_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{2.1}$$

Apply the transform $G_2^{\otimes n}$ (where "$\otimes n$" denotes the n^{th} Kronecker power) to a block of $N = 2^n$ bits U_0^{N-1}. Transmit the result $X_0^{N-1} = U_0^{N-1} G_2^{\otimes n}$ through independent copies of a B-DMC W (see Figure 2.1).

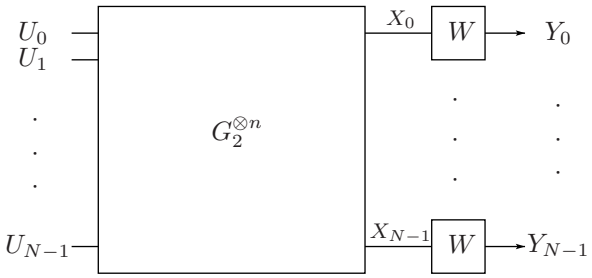

Figure 2.1: The transform $G_2^{\otimes n}$ is applied to the vector U_0^{N-1} and the resulting vector X_0^{N-1} is transmitted through the channel W.

Apply the chain rule to the mutual information between the input U_0^{N-1} and the output Y_0^{N-1}. This gives

$$I(U_0^{N-1}; Y_0^{N-1}) = \sum_{i=0}^{N-1} I(U_i; Y_0^{N-1} \,|\, U_0^{i-1}) = \sum_{i=0}^{N-1} I(U_i; Y_0^{N-1}, U_0^{i-1}),$$

where the last equality follows from the fact that $\{U_i\}$ are independent. The central observation of polar codes is that as n grows large, except for a negligible fraction, the terms in the summation either approach 0 (bad) or 1 (good). Moreover, the fraction of those terms tending to 1 approaches the symmetric mutual information $I(W)$. This phenomenon is referred to as *channel polarization*.

Note that we can associate the i-th row of $G_2^{\otimes n}$ with $I(U_i; Y_0^{N-1}, U_0^{i-1})$. More precisely, this term can be interpreted as the channel "seen" by the bit U_i assuming that we have already decoded all the previous bits. With this interpretation we choose the generator matrix of the polar code of rate R in the following way; choose those NR rows of $G_2^{\otimes n}$ with the largest $I(U_i; Y_0^{N-1} \,|\, U_0^{i-1})$. We will see later that such codes achieve rates close to $I(W)$, with vanishing block error probability, using a low-complexity SC decoder.

The following sections discuss polar codes in more detail. We start by applying the transform G_2 to a two dimensional vector in Section 2.1. We discuss the properties of the mutual information terms appearing in the chain rule. In Section 2.2 we recursively apply this transform to $N = 2^n$ dimensional vectors. In Section 2.3 we prove that channel polarization happens for the recursive transform introduced in the previous section. We discuss the construction of polar codes and analyze their block error probability in Section 2.4. The complexity issues are considered in Section 2.5. Finally, in Section 2.6 we end with some simulation results.

2.1 Basic Channel Transform

Let $\mathcal{X} = \{0, 1\}$ and $W : \mathcal{X} \to \mathcal{Y}$ be a B-DMC. Consider a random vector U_0^1 that is uniformly distributed over $\mathcal{X}^2$. Let $X_0^1 = U_0^1 G_2$ be the input to two independent uses of the channel W and let Y_0^1 be the corresponding outputs. The channel between U_0^1 and Y_0^1 is defined by the transition probabilities

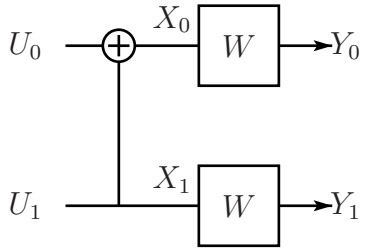

Figure 2.2: Channel combining of two channels.

$$W_2(y_0^1 \,|\, u_0^1) \triangleq \prod_{i=0}^{1} W(y_i \,|\, x_i) = \prod_{i=0}^{1} W(y_i \,|\, (u_0^1 G_2)_i). \tag{2.2}$$

In [32], this step is referred to as channel combining, which reflects the fact that two individual channels W are combined to create a new super channel $W_2 : \mathcal{X}^2 \to \mathcal{Y}^2$. Note that the linear transform between U_0^1 and X_0^1 is a bijection. Therefore,

$$I(U_0^1; Y_0^1) = I(X_0^1; Y_0^1) = 2I(W).$$

Using the chain rule of mutual information we can split the left hand side of the above equation as

$$\begin{aligned} I(U_0^1; Y_0^1) &= I(U_0; Y_0^1) + I(U_1; Y_0^1 \,|\, U_0) \\ &= I(U_0; Y_0^1) + I(U_1; Y_0^1, U_0). \end{aligned}$$

We can interpret the term $I(U_0; Y_0^1)$ as the mutual information of the channel between U_0 and the output Y_0^1, with U_1 considered as noise. Let us denote this channel by $W_2^{(0)}$. The transition probabilities of $W_2^{(0)} : \mathcal{X} \to \mathcal{Y}^2$ are given by marginalizing W_2 over U_1 as follows

$$W_2^{(0)}(y_0^1 \,|\, u_0) = \frac{1}{2} \sum_{u_1} W_2(y_0^1 \,|\, u_0^1) = \frac{1}{2} \sum_{u_1} W(y_0 \,|\, u_0 \oplus u_1) W(y_1 \,|\, u_1). \tag{2.3}$$

Similarly, the term $I(U_1; Y_0^1, U_0)$ can be interpreted as the mutual information of the channel between U_1 and Y_0^1 when U_0 is available at the decoder. Let us denote this channel by $W_2^{(1)}$. The transition probabilities of $W_2^{(1)} : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$ are given by

$$W_2^{(1)}(y_0^1, u_0 \,|\, u_1) = \frac{1}{2} W_2(y_0^1 \,|\, u_0^1) = \frac{1}{2} W(y_0 \,|\, u_0 \oplus u_1) W(y_1 \,|\, u_1). \tag{2.4}$$

In [32], this step is referred to as channel splitting, which reflects the fact that we split W_2 into two channels $W_2^{(0)}$ and $W_2^{(1)}$.

Let us define the following notation for the channels created by the above transformations. For any two B-DMCs $Q_1 : \mathcal{X} \to \mathcal{Y}_1, Q_2 : \mathcal{X} \to \mathcal{Y}_2$, let $Q_1 \boxast Q_2 : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2$ denote the B-DMC

$$(Q_1 \boxast Q_2)(y_1, y_2 \,|\, u) = \frac{1}{2} \sum_{x} Q_1(y_1 \,|\, u \oplus x) Q_2(y_2 \,|\, x),$$

and let $Q_1 \circledast Q_2 : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{X}$ denote the B-DMC

$$(Q_1 \circledast Q_2)(y_1, y_2, x \,|\, u) = \frac{1}{2} Q_1(y_1 \,|\, x \oplus u) Q_2(y_2 \,|\, u).$$

Using this notation, we can write $W_2^{(0)} = W \boxast W$ and $W_2^{(1)} = W \circledast W$. The channels $W_2^{(0)}$ and $W_2^{(1)}$ satisfy the following properties.

Lemma 2.1 (Transformation of Mutual Information)**.** *Let W be a B-DMC and let $W_2^{(0)}$ and $W_2^{(1)}$ be as defined above. Then*

$$\begin{aligned} I(W_2^{(0)}) + I(W_2^{(1)}) &= 2I(W), \\ I(W_2^{(0)}) \leq I(W) &\leq I(W_2^{(1)}). \end{aligned}$$

Proof. The first equality follows from the chain rule of mutual information. Since

$$I(W_2^{(0)}) + I(W_2^{(1)}) = I(U_0^1; Y_0^1) = I(X_0^1; Y_0^1) = 2I(W).$$

The second inequality follows by combining the first equality with

$$\begin{aligned} I(W_2^{(1)}) &= I(U_1; Y_0^1, U_0) \\ &= H(U_1) - H(U_1 \,|\, Y_0^1, U_0) \\ &\geq H(U_1) - H(U_1 \,|\, Y_1) = I(W). \end{aligned}$$

□

Lemma 2.2 (Transformation of Bhattacharyya Parameter)**.** *Let W be a B-DMC and let $W_2^{(0)}$ and $W_2^{(1)}$ be as defined before. Then*

$$\begin{aligned} Z(W_2^{(0)}) &\leq 2Z(W) - Z(W)^2, \\ Z(W_2^{(1)}) &= Z(W)^2. \end{aligned}$$

Proof. Note that $W_2^{(0)} = W \boxplus W$ and $W_2^{(1)} = W \circledast W$. The two inequalities follow by setting $W_1 = W_2 = W$ in Lemma 2.15 and Lemma 2.16 (see Appendix). □

2.2 Recursive Application of the Basic Transform

In the previous section, using the channel combining and splitting operation, we have created two new channels $(W_2^{(0)}, W_2^{(1)})$ from two identical channels (W, W). In Lemma 2.1 we have seen the first traces of polarization, with $W_2^{(0)}$ being a worse channel than W and $W_2^{(1)}$ being a better channel than W. In this section, we will show how to combine $N = 2^n$ channels W such that the channels $\{W_N^{(i)}\}_{i=0}^{N-1}$ resulting after the splitting operation polarize into either clean channels or completely noisy channels.

The channel combining is done in a recursive manner. The channel $W_N : \mathcal{X}^N \to \mathcal{Y}^N$ is defined as

$$W_N(y_0^{N-1} \,|\, u_0^{N-1}) = W_{N/2}(y_0^{N/2-1} \,|\, u_{0,e}^{N-1} \oplus u_{0,o}^{N-1}) W_{N/2}(y_{N/2}^{N-1} \,|\, u_{0,o}^{N-1}),$$

where $u_{0,o}^{N-1} = (u_1, u_3, \ldots, u_{N-1})$ and $u_{0,e}^{N-1} = (u_0, u_2, \ldots, u_{N-2})$. Similarly, the channel $W_{N/2}$ is defined in terms of $W_{N/4}$. And the channel W_2 is as defined before, i.e., W plays the role of W_1 in the recursion,

$$W_2(y_0^1 \,|\, u_0^1) = W(y_0 \,|\, u_0 \oplus u_1) W(y_1 \,|\, u_1).$$

Figure 2.3 shows one step of this recursion. First apply the transform G_2 to the 2^{n-1} pairs (U_{2i}, U_{2i+1}) and send the resulting output V_0^{N-1} through a

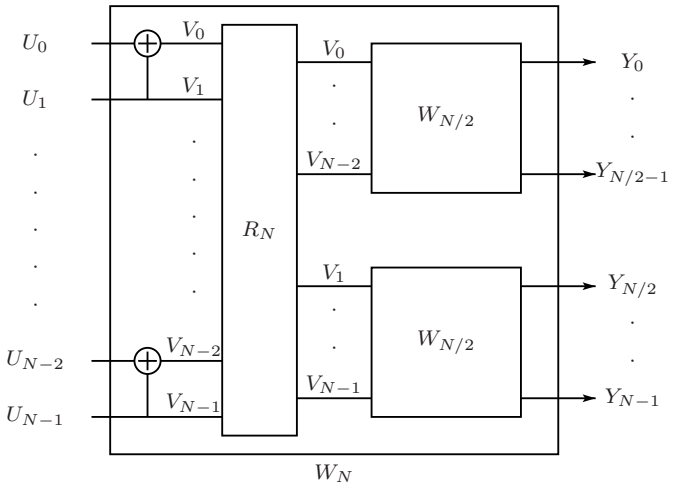

Figure 2.3: Recursive channel combining operation.

permutation R_N. The permutation R_N splits the output V_0^{N-1} into two sets $V_{0,o}^{N-1}$ and $V_{0,e}^{N-1}$. The two sets are then fed as inputs to two identical channels $W_{N/2}$. As an illustrative example, Figure 2.4 shows the recursive combining for 8 channels.

Let $P_{U_0^{N-1},X_0^{N-1},Y_0^{N-1}}$ denote the probability distribution induced on the set $\mathcal{X}^N \times \mathcal{X}^N \times \mathcal{Y}^N$ due to the channel combining operation. Since U_i are i.i.d. Ber($\frac{1}{2}$) random variables, the probability distribution can be expressed as

$$P_{U_0^{N-1},X_0^{N-1},Y_0^{N-1}}(u_0^{N-1}, x_0^{N-1}, y_0^{N-1}) = \frac{1}{2^N} \mathbb{1}_{\{x_0^{N-1} = u_0^{N-1} G_2^{\otimes n}\}} \prod_{i=0}^{N-1} W(y_i \,|\, x_i). \tag{2.5}$$

The channel between U_0^{N-1} and Y_0^{N-1} is defined by the transition probabilities

$$W_N(y_0^{N-1} \,|\, x_0^{N-1}) = P_{Y_0^{N-1} \,|\, U_0^{N-1}}(y_0^{N-1} \,|\, u_0^{N-1}) = \prod_{i=0}^{N-1} W(y_i \,|\, (u_0^{N-1} G_2^{\otimes n})_i).$$

From now on we will drop the subscript of P whenever it is clear from its arguments. The splitting operation is defined through the chain rule as before,

$$I(U_0^{N-1}; Y_0^{N-1}) = \sum_{i=0}^{N-1} I(U_i; Y_0^{N-1}, U_0^{i-1}).$$

The term $I(U_i; Y_0^{N-1}, U_0^{i-1})$ corresponds to the channel between U_i and (Y_0^{N-1}, U_0^{i-1}). Let us denote this channel by $W_N^{(i)} : \mathcal{X} \to \mathcal{Y}^N \times \mathcal{X}^{i-1}$. The transition probabilities are given by

$$W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, u_i) \triangleq P(y_0^{N-1}, u_0^{i-1} \,|\, u_i) = \sum_{u_{i+1}^{N-1}} \frac{P(y_0^{N-1} \,|\, u_0^{N-1}) P(u_0^{N-1})}{P(u_i)}$$

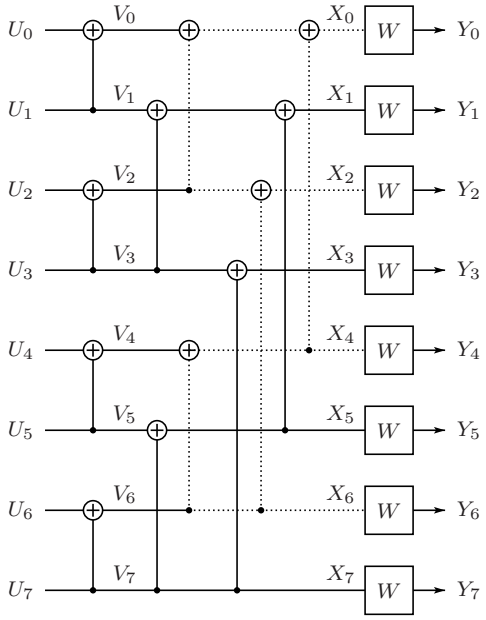

Figure 2.4: The channel W_8 obtained by combining 8 channels. The channels corresponding to the outputs $Y_{0,e}^7$ form the first W_4 channel (the dashed lines show the channel transform matrix). Similarly, the channels corresponding to the outputs $Y_{0,o}^7$ form the second W_4 channel.

$$= \frac{1}{2^{N-1}} \sum_{u_{i+1}^{N-1}} P(y_0^{N-1} \,|\, u_0^{N-1}) = \frac{1}{2^{N-1}} \sum_{u_{i+1}^{N-1}} W_N(y_0^{N-1} \,|\, u_0^{N-1}). \tag{2.6}$$

The following lemma shows the relationship between the channels $\{W_N^{(i)}\}$ and the channels $\{W_{N/2}^{(i)}\}$ which is crucial for the analysis later.

Lemma 2.3 (Relation Between $W_N^{(i)}$ and $W_{N/2}^{(i)}$)**.** *For $i = 0, \ldots, N/2 - 1$,*

$$\begin{aligned} &W_N^{(2i)}(y_0^{N-1}, u_0^{2i-1} \,|\, u_{2i}) \\ &\quad = \frac{1}{2} \sum_{u_{2i+1}} W_{N/2}^{(i)}(y_0^{N/2-1}, u_{0,o}^{2i-1} \,|\, u_{2i+1}) W_{N/2}^{(i)}(y_{N/2}^{N-1}, u_{0,e}^{2i-1} \oplus u_{0,o}^{2i-1} \,|\, u_{2i} \oplus u_{2i+1}), \end{aligned} \tag{2.7}$$

$$\begin{aligned} &W_N^{(2i+1)}(y_0^{N-1}, u_0^{2i} \,|\, u_{2i+1}) \\ &\quad = \frac{1}{2} W_{N/2}^{(i)}(y_0^{N/2-1}, u_{0,o}^{2i-1} \,|\, u_{2i+1}) W_{N/2}^{(i)}(y_{N/2}^{N-1}, u_{0,e}^{2i-1} \oplus u_{0,o}^{2i-1} \,|\, u_{2i} \oplus u_{2i+1}). \end{aligned} \tag{2.8}$$

Proof.

$W_N^{(2i)}(y_0^{N-1}, u_0^{2i-1} \,|\, u_{2i})$

$$= \frac{1}{2^{N-1}} \sum_{u_{2i+1}^{N-1}} W_{N/2}(y_0^{N/2-1} \,|\, u_{0,e}^{N-1} \oplus u_{0,o}^{N-1}) W_{N/2}(y_{N/2}^{N-1} \,|\, u_{0,o}^{N-1})$$

$$= \frac{1}{2} \frac{1}{2^{N/2-1}} \sum_{u_{2i+1,o}^{N-1}} W_{N/2}(y_{N/2}^{N-1} \,|\, u_{0,o}^{N-1}) \frac{1}{2^{N/2-1}} \sum_{u_{2i+1,e}^{N-1}} W_{N/2}(y_0^{N/2-1} \,|\, u_{0,e}^{N-1} \oplus u_{0,o}^{N-1}).$$

The second summation can be written as

$$\frac{1}{2^{N-1}} \sum_{u_{2i+1,e}^{N-1}} W_{N/2}(y_0^{N/2-1} \,|\, u_{0,e}^{2i} \oplus u_{0,o}^{2i}, u_{2i+1,e}^{N-1}).$$

The equation (2.7) now follows from the definition of $W_N^{(i)}$ given in (2.6). Using similar arguments equation (2.8) can also be derived. □

The identification

$$(y_0^{N/2-1}, u_{0,e}^{2i-1} \oplus u_{0,o}^{2i-1}) \to y_0, \quad (y_{N/2}^{N-1}, u_{0,o}^{2i-1}) \to y_1,$$
$$W_N^{(2i)} \to W_2^{(0)}, \; W_N^{(2i+1)} \to W_2^{(1)}, \; u_{2i} \to u_0, \; u_{2i+1} \to u_1,$$

suggests that the relationship shown in (2.7) and (2.8) is very similar to the equations (2.3) and (2.4) with one minor difference. For the 2×2 case, the output of $W_2^{(0)}$ is y_0^1 which is the joint output of the two channels W. This is not the case for the channel transformation of (2.7) and (2.8). However, the output of $W_N^{(2i)}$, namely (y_0^{N-1}, u_0^{2i-1}), and the outputs of the two $W_{N/2}^{(i)}$ channels, namely $(y_0^{N/2-1}, u_{0,e}^{2i-1} \oplus u_{0,o}^{2i-1})$ and $(y_{N/2}^{N-1}, u_{0,o}^{2i-1})$, are related by a one-to-one map. Let $f : \mathcal{Y}^N \times \mathcal{X}^{2i} \to \mathcal{Y}^N \times \mathcal{X}^{2i}$ denote the map

$$f(y_0^{N-1}, u_{0,e}^{2i-1}, u_{0,o}^{2i-1}) = (y_0^{N-1}, u_{0,e}^{2i-1} \oplus u_{0o}^{2i-1}, u_{0o}^{2i-1}).$$

Consider the channels $\tilde{W}_N^{(2i)}, \tilde{W}_N^{(2i+1)}$ defined as

$$\tilde{W}_N^{(2i)}(f(y_0^{N-1}, u_{0,e}^{2i-1}, u_{0o}^{2i-1}) \,|\, u_{2i}) \triangleq W_N^{(2i)}(y_0^{N-1}, u_{0,e}^{2i-1}, u_{0,o}^{2i-1} \,|\, u_{2i})$$
$$\tilde{W}_N^{(2i+1)}(f(y_0^{N-1}, u_{0,e}^{2i-1}, u_{0o}^{2i-1}), u_{2i} \,|\, u_{2i+1}) \triangleq W_N^{(2i+1)}(y_0^{N-1}, u_{0,e}^{2i-1}, u_{0,o}^{2i-1}, u_{2i} \,|\, u_{2i+1}).$$

Clearly, $\tilde{W}_N^{(2i)} = W_{N/2}^{(i)} \circledast W_{N/2}^{(i)}$. Moreover, up to a relabeling of the outputs, the channel $\tilde{W}_N^{(2i)}$ is equivalent to the channel $W_N^{(2i)}$. Therefore for all practical purposes, the two channels are equivalent and hence we say

$$W_N^{(2i)} = W_{N/2}^{(i)} \circledast W_{N/2}^{(i)}. \tag{2.9}$$

Based on similar reasoning, we say that

$$W_N^{(2i+1)} = W_{N/2}^{(i)} \circledast W_{N/2}^{(i)}. \tag{2.10}$$

Lemma 2.4 (Transformation of $I(W_N^{(i)})$ and $Z(W_N^{(i)})$)**.** *Let W be a B-DMC and let $W_N^{(i)}$ be as defined above. Then*

$$I(W_N^{(2i)}) \leq I(W_{N/2}^{(i)}) \leq I(W_N^{(2i+1)}),$$
$$I(W_N^{(2i)}) + I(W_N^{(2i+1)}) = 2I(W_{N/2}^{(i)}),$$

and

$$Z(W_N^{(2i)}) \leq 2Z(W_{N/2}^{(i)})^2 - Z(W_{N/2}^{(i)}),$$
$$Z(W_N^{(2i+1)}) = Z(W_{N/2}^{(i)})^2.$$

Proof. The lemma follows by combining (2.9), (2.10), with Lemma 2.1 and Lemma 2.2. □

2.3 Channel Polarization

In the previous section we have discussed a procedure to transform N copies of W to N distinct channels $\{W_N^{(i)}\}_{i=0}^{N-1}$. Using the recursive relations (2.9) and (2.10), we can express the channel $W_N^{(i)}$ as follows. Let $b_1 \dots b_n$ denote the n-bit binary expansion of i and let $W_{(b_1,\dots,b_n)} \triangleq W_N^{(i)}$. Then $W_N^{(i)}$ can be obtained by repeating the following operation,

$$W_{(b_1,\dots,b_{k-1},b_k)} = \begin{cases} W_{(b_1,\dots,b_{k-1})} \circledast W_{(b_1,\dots,b_{k-1})}, & \text{if } b_k = 0, \\ W_{(b_1,\dots,b_{k-1})} \circledcirc W_{(b_1,\dots,b_{k-1})}, & \text{if } b_k = 1. \end{cases} \tag{2.11}$$

where

$$W_{(b_1)} = \begin{cases} W \circledast W, & \text{if } b_1 = 0, \\ W \circledcirc W, & \text{if } b_1 = 1, \end{cases}$$

To analyze the behavior of these channels it is convenient to represent them through the following random process. Let $\{B_n : n \geq 1\}$ be a sequence of i.i.d. symmetric Bernoulli random variables defined over a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_0 = \{\phi, \Omega\}$ denote the trivial σ-field and let $\{\mathcal{F}_n, n \geq 1\}$ denote the σ-fields generated by the random variables $(B_1, \dots, B_n)$. Moreover, assume that $\mathcal{F}$ is such that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$. Let $W_0 = W$ and $\{W_n, n \geq 0\}$ denote a tree process with two branches defined as

$$W_{n+1} = \begin{cases} W_n \circledast W_n & \text{if } B_n = 0, \\ W_n \circledcirc W_n & \text{if } B_n = 1. \end{cases}$$

The output space of the random variable W_n is the set of channels $\{W_{2^n}^{(i)}\}_{i=0}^{2^n-1}$. Moreover, W_n is uniformly distributed over the channels $\{W_{2^n}^{(i)}\}_{i=0}^{2^n-1}$. For our purpose it is sufficient to track the Bhattacharyya parameters $\{Z(W_N^{(i)})\}$ and the mutual informations $\{I(W_N^{(i)})\}$. Therefore, we will restrict ourselves to

the analysis of the two processes $\{I_n : n \geq 0\} := \{I(W_n) : n \geq 0\}$ and $\{Z_n : n \geq 0\} := \{Z(W_n) : n \geq 0\}$. From the definitions of the random variables I_n and Z_n, we have

$$\Pr[I_n \in (a,b)] = \frac{|\{i : I(W_{2^n}^{(i)}) \in (a,b)\}|}{2^n},$$
$$\Pr[Z_n \in (a,b)] = \frac{|\{i : Z(W_{2^n}^{(i)}) \in (a,b)\}|}{2^n}.$$

The random processes $\{I_n\}$ and $\{Z_n\}$ satisfy the following properties.

Lemma 2.5 ($\{I_n\}$ and $\{Z_n\}$ [32])**.** *Let $\{I_n, Z_n, \mathcal{F}_n : n \geq 0\}$ be as defined before. Then*

(i) the sequence $\{(I_n, \mathcal{F}_n) : n \geq 0\}$ is a bounded martingale.

(ii) the sequence $\{(Z_n, \mathcal{F}_n) : n \geq 0\}$ is a bounded super-martingale.

Proof. Let $b_1 \dots b_n$ be a realization of the random variables $B_1, \dots, B_n$. Then

$$\begin{aligned}\mathbb{E}[I_{n+1} \,|\, B_1 = b_1, \dots, B_n = b_n] &= \frac{1}{2} I(W_{(b_1,\dots,b_n,0)}) + \frac{1}{2} I(W_{(b_1,\dots,b_n,1)}) \\ &= I(W_{(b_1,\dots,b_n)}),\end{aligned}$$

where the last equality follows from (2.11) and Lemma 2.1. Similarly, the proof for $\{Z_n\}$ follows from (2.11) and Lemma 2.2. The boundedness for both the processes follows from the fact that for any B-DMC W, $0 \leq I(W) \leq 1$ and $0 \leq Z(W) \leq 1$. □

Lemma 2.6 (Channel Polarization [32])**.** *Let $\{I_n, Z_n, \mathcal{F}_n : n \geq 0\}$ be as defined before. Then*

(i) the sequence $\{I_n\}$ converges almost surely to a random variable I_∞ and

$$I_\infty = \begin{cases} 1 & w.p.\ I(W), \\ 0 & w.p.\ 1 - I(W). \end{cases}$$

(ii) the sequence $\{Z_n\}$ converges almost surely to a random variable Z_∞ and

$$Z_\infty = \begin{cases} 1 & w.p.\ 1 - I(W), \\ 0 & w.p.\ I(W). \end{cases}$$

Proof. It is sufficient to prove one of the two statements. The second then follows by applying Lemma 1.5. Since $\{I_n\}$ is a bounded martingale, the limit $\lim_{n\to\infty} I_n$ converges almost surely and in $\mathcal{L}^1$ to a random variable I_∞. Similarly, since $\{Z_n\}$ is a bounded super-martingale, the limit $\lim_{n\to\infty} Z_n$ converges almost surely and in $\mathcal{L}^1$ to a random-variable Z_∞. The convergence in

$\mathcal{L}^1$ implies, $\mathbb{E}[|Z_{n+1} - Z_n|] \overset{n\to\infty}{\to} 0$. Since $Z_{n+1} = Z_n^2$ with probability $\frac{1}{2}$, we have

$$\mathbb{E}[|Z_{n+1} - Z_n|] \geq \frac{1}{2}\mathbb{E}[Z_n(1 - Z_n)] \geq 0.$$

Thus, $\mathbb{E}[Z_n(1-Z_n)] \overset{n\to\infty}{\to} 0$, which implies $\mathbb{E}[Z_\infty(1-Z_\infty)] = 0$. Therefore, $Z_\infty \in \{0, 1\}$ *a.s.* This fact combined with Lemma 1.5, implies that $I_\infty \in \{0, 1\}$ *a.s.* Since $\{I_n\}$ is a martingale, we have $P(I_\infty = 1) = \mathbb{E}[I_\infty] = \mathbb{E}[I_0] = I(W)$. □

From Lemma 2.6, we have for any $\delta > 0$

$$\lim_{n\to\infty} \frac{|\{i : I(W_{2^n}^{(i)}) \in (1-\delta, 1]\}|}{2^n} = \lim_{n\to\infty} \Pr[I_n \in (1-\delta, 1]] = I(W),$$
$$\lim_{n\to\infty} \frac{|\{i : I(W_{2^n}^{(i)}) \in [0, \delta)\}|}{2^n} = \lim_{n\to\infty} \Pr[I_n \in [0, \delta)] = 1 - I(W).$$

This implies that as the length gets larger almost all the channels $\{W_N^{(i)}\}$ get polarized to either clean channels (mutual information close to 1) or noisy channels (mutual information close to 0). For obvious reasons, we refer to the channels with mutual information close to 1 as "good channels" and the remaining channels as "bad channels".

2.4 Polar Codes Achieve Channel Capacity

In the previous section we have seen how to create channels that are polarized. Having such a collection of channels between the encoder and the decoder suggests a natural coding scheme where information bits are transmitted on the good channels. The inputs to the remaining channels are fixed and declared to the decoder in advance. Since the fraction of good channels is tending to $I(W)$, we can achieve rates close to $I(W)$.

The mutual information $I(U_i; Y_0^{N-1}, U_0^{i-1})$ corresponds to decoding U_i with the knowledge of U_0^{i-1} and the output Y_0^{N-1}. To create such a channel, the decoder should know U_0^{i-1} when decoding U_i. But the decoder knows in advance only those U_j that are fixed (indices corresponding to bad channels) and for the rest, the decoder can only have its estimate $\hat{U}_j$, which need not be correct. We therefore consider an SC decoder which decodes the bits in the order $U_0, \ldots, U_{N-1}$, so that the decoder at least has an estimate of U_0^{i-1} while decoding U_i. In the following we show that, using channel polarization, by an appropriate choice of the indices that are fixed the block error probability of the SC decoder vanishes for rates below $I(W)$.

Let us first provide a few definitions regarding the coding scheme.

Definition 2.7 (Polar Code)**.** *The polar code $\mathtt{C}_N(F, u_F)$, defined for any $F \subseteq \{0, \ldots, N-1\}$ and $u_F \in \mathcal{X}^{|F|}$, is a linear code given by*

$$\mathtt{C}_N(F, u_F) = \{x_0^{N-1} = u_0^{N-1} G_2^{\otimes n} : u_{F^c} \in \mathcal{X}^{|F^c|}\}.$$

In other words the code $\mathtt{C}_N(F, u_F)$ is constructed by fixing the indices in F to u_F and varying the indices in F^c over all possible values. Let us refer to the set F as frozen set and the indices belonging to it as frozen indices. Using a code $\mathtt{C}_N(F, u_F)$ is equivalent to transmitting U_0^{N-1} through the channel W_N with the indices in F fixed to u_F.

Definition 2.8 (Polar Code Ensemble)**.** *The polar code ensemble $\mathtt{C}_N(F)$, defined for any $F \subseteq \{0, 1 \ldots, N-1\}$, denotes the ensemble*

$$\mathtt{C}_N(F) = \{\mathtt{C}_N(F, u_F), \forall u_F \in \mathcal{X}^{|F|}\}.$$

Let $L_N^{(i)}(y_0^{N-1}, v_0^{i-1})$ denote the function

$$L_N^{(i)}(y_0^{N-1}, u_0^{i-1}) = \frac{W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, 0)}{W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, 1)}, \tag{2.12}$$

which is the likelihood ratio of U_i given the output y_0^{N-1} and the past $U_0^{i-1} = u_0^{i-1}$. Let $\hat{U}_i(y_0^{N-1}, u_0^{i-1})$ denote the decision based on this likelihood, i.e.,

$$\hat{U}_i(y_0^{N-1}, u_0^{i-1}) = \begin{cases} 0, & \text{if } L_N^{(i)}(y_0^{N-1}, u_0^{i-1}) > 1, \\ 1, & \text{if } L_N^{(i)}(y_0^{N-1}, u_0^{i-1}) \leq 1. \end{cases} \tag{2.13}$$

Let $\mathtt{C}_N(F, u_F)$ be the code used for transmission. The vector u_F is kept fixed during the course of transmission and is declared to the decoder. At the decoder consider an SC algorithm which operates as follows. For each i in the range 0 till $N-1$:

(i) If $i \in F$, then set $\hat{u}_i = u_i$.

(ii) If $i \in F^c$, then compute $L_N^{(i)}(y_0^{N-1}, \hat{u}_0^{i-1})$ and set $\hat{u}_i = \hat{U}_i(y_0^{N-1}, \hat{u}_0^{i-1})$.

Let the function $\hat{U}(y_0^{N-1}, u_F)$ denote the output of the above decoding algorithm. Let $P_B(F, u_F)$ denote the block error probability of the code $\mathtt{C}_N(F, u_F)$ with a uniform probability over the codewords. Let $P_B(F)$ denote the average block error probability of the ensemble $\mathtt{C}_N(F)$, i,e., average of $P_B(F, u_F)$ with a uniform choice over all $u_F \in \mathcal{X}^{|F|}$.

Lemma 2.9 (Bounds on Block Error Probability)**.** *For a given B-DMC W and a set of frozen indices F, the average block error probability $P_B(F)$ is bounded as*

$$\max_{i \in F^c} \frac{1}{2}\left(1 - \sqrt{1 - Z(W_N^{(i)})^2}\right) \leq P_B(F) \leq \sum_{i \in F^c} Z(W_N^{(i)}).$$

Proof. Since the codewords are chosen with uniform distribution, the block error probability for $\mathtt{C}_N(F, u_F)$ can be expressed as

$$P_B(F, u_F) = \frac{1}{2^{|F^c|}} \sum_{u_{F^c}, y_0^{N-1}} W_N(y_0^{N-1} \,|\, u_0^{N-1}) \mathbb{1}_{\{\hat{U}(y_0^{N-1}, u_F) \neq u_0^{N-1}\}}$$

$$= \frac{1}{2^{|F^c|}} \sum_{u_{F^c}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \mathbb{1}_{\{\exists i : \hat{U}_i(y_0^{N-1}, \hat{u}_0^{i-1}) \neq u_i\}}$$

$$= \frac{1}{2^{|F^c|}} \sum_{u_{F^c}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \mathbb{1}_{\{\exists i : \hat{u}_0^{i-1} = u_0^{i-1}, \hat{U}_i(y_0^{N-1}, \hat{u}_0^{i-1}) \neq u_i\}}$$

$$= \frac{1}{2^{|F^c|}} \sum_{u_{F^c}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \mathbb{1}_{\{\exists i : \hat{U}_i(y_0^{N-1}, u_0^{i-1}) \neq u_i\}}.$$

Averaging over the ensemble $\mathtt{C}_N(F)$, we get

$$P_B(F) = \frac{1}{2^{|F|}} \sum_{u_F} P_B(F, u_F)$$

$$= \frac{1}{2^N} \sum_{u_0^{N-1}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \mathbb{1}_{\{\exists i : \hat{U}_i(y_0^{N-1}, u_0^{i-1}) \neq u_i\}}.$$

For any $j \in F^c$, P_B can be lower bounded by

$$P_B(F) \geq \frac{1}{2^N} \sum_{u_0^{N-1}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \mathbb{1}_{\{\hat{U}_j(y_0^{N-1}, u_0^{j-1}) \neq u_j\}}$$

$$= \frac{1}{2} \sum_{u_j \in \mathcal{X}} \sum_{u_0^{j-1}, y_0^{N-1}} W_N^{(j)}(y_0^{N-1} u_0^{j-1} \mid u_j) \mathbb{1}_{\left\{ \frac{W_N^{(j)}(y_0^{N-1} u_0^{j-1} \mid u_j)}{W_N^{(j)}(y_0^{N-1} u_0^{j-1} \mid u_j \oplus 1)} \leq 1 \right\}}$$

$$= P_e(W_N^{(j)}),$$

where $P_e(W)$ is the probability of error function defined in (1.3). Using the union bound, P_B can be upper bounded as

$$P_B(F) \leq \frac{1}{2^N} \sum_{u_0^{N-1}, y_0^{N-1}} W_N(y_0^{N-1} \mid u_0^{N-1}) \sum_{i \in F^c} \mathbb{1}_{\{\hat{U}_i(y_0^{N-1}, u_0^{i-1}) \neq u_i\}}$$

$$\leq \sum_{i \in F^c} P_e(W_N^{(i)}).$$

The statement follows by applying Lemma 1.6. □

For the block error probability to vanish, we should show that for the indices that are used to transmit information, i.e., for $i \in F^c$, $Z(W_N^{(i)})$ decays faster than $1/N$. This corresponds to showing that the random process $\{Z_n\}$, when tends to 0, decays at a sufficiently fast rate. In [47] Arıkan and Telatar obtained the following result which yields a stretched-exponential bound on the block error probability.

Theorem 2.10 (Rate of Convergence [47])**.** *Let* $\{X_n : n \geq 0\}$ *be a positive random process satisfying*

$$\begin{array}{ll} X_n \leq X_{n+1} \leq q X_n & w.p.\ \frac{1}{2}, \\ X_{n+1} = X_n^2 & w.p.\ \frac{1}{2}. \end{array}$$

Let $X_\infty := \lim_{n\to\infty} X_n$ *exist almost surely and* $\Pr(X_\infty = 0) = P_\infty$. *Then for any* $\beta < \frac{1}{2}$,

$$\lim_{n\to\infty} \Pr(X_n < 2^{-2^{n\beta}}) = P_\infty, \tag{2.14}$$

and for any $\beta > \frac{1}{2}$,

$$\lim_{n\to\infty} \Pr(X_n > 2^{-2^{n\beta}}) = 0. \tag{2.15}$$

Our process $\{Z_n\}$ satisfies the conditions of Theorem 2.10 with $q = 2$ and $P_\infty = I(W)$. This implies the following theorem. Though the result follows directly from Theorem 2.10, we state it as a theorem due to its significance.

Theorem 2.11 (Rate of Z_n Approaching 0 [47])**.** *Given a B-DMC* W*, and any* $\beta < \frac{1}{2}$,

$$\lim_{n\to\infty} \Pr(Z_n \leq 2^{-2^{n\beta}}) = I(W).$$

Now we are ready to prove the main theorem of this chapter which shows that polar codes achieve the symmetric capacity using SC decoder.

Theorem 2.12 (Polar Codes Achieve the Symmetric Capacity [32])**.** *Given a B-DMC* W *and fixed* $R < I(W)$, *for any* $\beta < \frac{1}{2}$ *there exists a sequence of polar codes of rate* $R_N > R$ *such that*

$$P_N = O(2^{-(N)^\beta}).$$

Proof. Fix any $0 < \beta < \frac{1}{2}$ and $\epsilon > 0$. Choose the set of frozen indices as

$$F_N = \{i : Z(W_N^{(i)}) > \frac{1}{N} 2^{-(N)^\beta}\}.$$

Theorem 2.11 implies that for N sufficiently large,

$$\frac{|F_N^c|}{N} \geq I(W) - \epsilon.$$

The block error probability of such a scheme under SC decoding is

$$P_B(F_N) \leq \sum_{i\in F_N^c} Z(W_N^{(i)}) \leq 2^{-(N)^\beta}. \tag{2.16}$$

Recall that $P_B(F_N)$ is the block error probability averaged over all choices of u_F. Since the average block error probability fulfills (2.16), it follows that there must be at least one choice of u_{F_N} for which

$$P_B(F_N, u_{F_N}) \leq 2^{-(N^\beta)}$$

for any $0 < \beta < \frac{1}{2}$. □

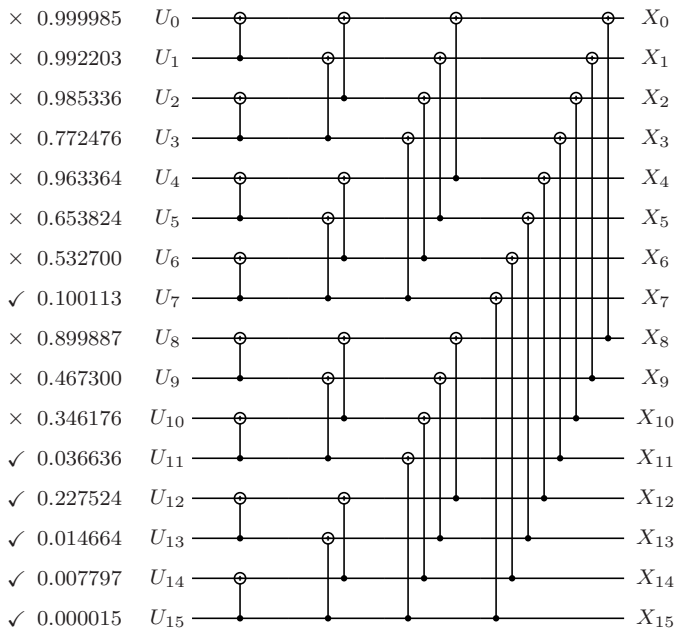

Figure 2.5: Example of a polar code of length 16. The Bhattacharyya parameters $Z(W_N^{(i)})$ are shown for the BEC(0.5). The frozen indices are marked as $\times$ and the remaining indices are marked as ✓. The rate of the resulting code is $6/16$.

Theorem 2.10 also provides a lower bound on the block error probability. For any rate $R > 0$ and $\beta > \frac{1}{2}$, (2.15) implies that for n sufficiently large, any set F_N^c of size NR will satisfy $\max_{i \in F_N^c} Z_N^{(i)} > 2^{-2^{n\beta}}$. Combining this with the lower bound in Lemma 2.9 we obtain

$$P_N > 2^{-(N)^\beta}.$$

Remark 2.13. *In [32] Arıkan showed that*

$$\lim_{n \to \infty} \Pr(Z_n \leq 2^{-5n/4}) = I(W).$$

The resulting bound on the block error probability is $P_B(N, R) \leq \frac{1}{N^{1/4}}$.

Example 2.14 (Polar Code for $N = 16$, BEC(0.5))**.** *Consider a polar code of length* $N = 16$*. Let the code be designed for the BEC*(0.5)*. The Bhattacharyya parameters are shown in Figure 2.5. To design a code of rate* $6/16$*, we choose the frozen set* F *to consist of the 10 indices with the largest Bhattacharyya parameters. We get* $F = \{0, 1, 2, 4, 8, 3, 5, 6, 9, 10\}$.

In [32, Section VI] it was shown that for symmetric B-DMCs, all codes in the ensemble $\mathtt{C}_N(F)$ have the same performance. In other words, the value of u_F does not influence the block error probability. Therefore, a convenient choice is to set it to zero.

2.5 Complexity

2.5.1 Encoding and Decoding

The main advantage of polar codes is that both the encoding and the SC decoding can be implemented with $O(N \log N)$ complexity. Let us first consider the encoding complexity. Let $\chi_E(N)$ denote the encoding complexity for blocklength N. The recursive channel combining operation shown in Figure 2.3 implies that

$$\chi_E(N) = \frac{N}{2} + 2\chi_E(N/2).$$

The encoding complexity for block length 2 is $\chi_E(2) = 1$ because we need just one XOR operation to compute $u_0^1 G_2$. The above recursive relation implies

$$\begin{aligned}\chi_E(N) &= \frac{N}{2} + 2\chi_E(N/2) = \frac{N}{2} + 2(N/4 + 2\chi_E(N/4)) \\ &= \frac{N}{2} + \frac{N}{2} + 4(N/8 + 2\chi_E(N/8)) = \cdots = \frac{N}{2}\log N.\end{aligned}$$

This implies that the encoding complexity is $O(N \log N)$. For example, consider the factor graph representation for $N = 8$ as shown in Figure 2.6. To compute the vector x_0^7 from u_0^7, we need $\frac{N}{2} = 4$ XOR operations in each layer and there are $\log N = 3$ layers in total.

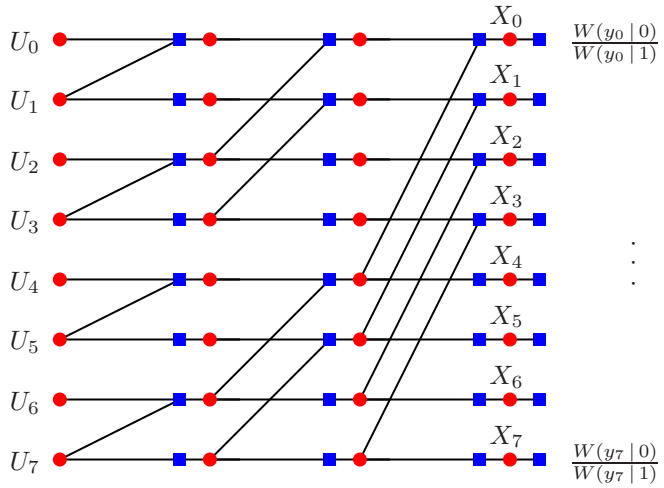

Figure 2.6: Factor graph representation used by the encoder and the SC decoder. The quantities $\frac{W(y_i \mid 0)}{W(y_i \mid 1)}$ are the initial likelihoods of the variables X_i, when y_i is received at the output of a B-DMC W.

Let us now look at the SC decoding complexity. For a detailed explanation with an example please refer to [32]. Using (2.7) and (2.8) we can prove that

$$L_N^{(2i)}(y_0^{N-1}, \hat{u}_0^{2i-1}) = \frac{1 + L_{N/2}^{(i)}(y_0^{N/2-1}, \hat{u}_{0,e}^{2i-1} \oplus \hat{u}_{0,o}^{2i-1}) L_{N/2}^{(i)}(y_{N/2}^{N-1}, \hat{u}_{0,o}^{2i-1})}{L_{N/2}^{(i)}(y_0^{N/2-1}, \hat{u}_{0,e}^{2i-1} \oplus \hat{u}_{0,o}^{2i-1}) + L_{N/2}^{(i)}(y_{N/2}^{N-1}, \hat{u}_{0,o}^{2i-1})},$$

$$L_N^{(2i+1)}(y_0^{N-1}, \hat{u}_0^{2i}) = L_{N/2}^{(i)}(y_0^{N/2-1}, \hat{u}_{0,e}^{2i-1} \oplus \hat{u}_{0,o}^{2i-1})^{1-2\hat{u}_{2i}} L_{N/2}^{(i)}(y_{N/2}^{N-1}, \hat{u}_{0,o}^{2i-1}).$$

Therefore, the likelihoods $\{L_N^{(i)}\}$ can be computed with $O(N)$ computations from the two sets of likelihoods

$$\{L_{N/2}^{(i)}(y_0^{N/2-1}, \hat{u}_{0,e}^{2i-1} \oplus \hat{u}_{0,o}^{2i-1}) : i \in \{0, \ldots, N/2-1\}\},$$
$$\{L_{N/2}^{(i)}(y_{N/2}^{N-1}, \hat{u}_{0,o}^{2i-1}) : i \in \{0, \ldots, N/2-1\}\}.$$

Now the problem is reduced to two similar problems of length $N/2$. Let $\chi_D(N)$ denote the decoding complexity of a code of block length N. Then

$$\chi_D(N) = O(N) + 2\chi_D(N/2).$$

The decoding complexity for blocklength 2 is $O(1)$, assuming that infinite precision arithmetic can be implemented with unit cost. By the same reasoning as before this implies that the decoding complexity is $O(N \log N)$.

2.5.2 Code Construction

Another issue where complexity is of concern is the code design. Recall that the code construction requires the knowledge of either $\{Z(W_N^{(i)})\}$ or $\{I(W_N^{(i)})\}$. If W is a BEC, then the resulting channels $\{W_N^{(i)}\}$ are all BECs and hence it is sufficient to track their erasure probabilities. These in turn are equal to the Bhattacharyya parameters. For the BEC, the recursions (2.11) simplify to

$$Z(W_{(b_1,\ldots,b_{k-1},b_k)}) = \begin{cases} 2Z(W_{(b_1,\ldots,b_{k-1})}) - Z(W_{(b_1,\ldots,b_{k-1})})^2 & \text{if } b_k = 0, \\ Z(W_{(b_1,\ldots,b_{k-1})})^2 & \text{if } b_k = 1, \end{cases} \tag{2.17}$$

where $b_1 \ldots b_n$ is the n-bit binary expansion of i and $Z(W_N^{(i)}) = Z(W_{(b_1,\ldots,b_n)})$. Clearly, the computation of each $Z(W_N^{(i)})$ can be accomplished using $O(\log N)$ arithmetic operations. Hence, the total complexity is $O(N \log N)$, under the assumption that infinite precision arithmetic can be implemented at unit cost. As mentioned in [52], by reusing the intermediate Bhattacharyya parameters the complexity can be reduced to $O(N)$. Let $\chi_C(N)$ denote the code construction complexity for the BEC. The equations (2.17) imply that $\{Z_N^{(i)}\}$ can be computed from $\{Z_{N/2}^{(i)}\}$ with an additional $2N$ operations. Therefore,

$$\chi_C(N) = 2N + \chi_C(N/2),$$

which implies $\chi_C(N) = O(N)$.

For channels other than the BEC, the complexity grows exponentially with N. This is because, the output alphabet of $W_N^{(i)}$ is $\mathcal{Y}^N \times \mathcal{X}^i$. Arıkan suggested to use Monte-Carlo method for estimating the Bhattacharyya parameters.

$$Z(W_N^{(i)}) = \sum_{y_0^{N-1}, u_0^{i-1}} W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, 1) \sqrt{L_N^{(i)}(y_0^{N-1}, u_0^{i-1})}$$

$$= \sum_{y_0^{N-1}, u_0^{i-1}, u_{i+1}^{N-1}} \frac{1}{2^{N-1}} W_N(y_0^{N-1} \,|\, u_0^{i-1}, 1, u_{i+1}^{N-1}) \sqrt{L_N^{(i)}(y_0^{N-1}, u_0^{i-1})}.$$

To estimate the above quantity, we generate $(u_0^{i-1}, u_{i+1}^{N-1})$ uniformly at random, and y_0^{N-1} according to $W_N(y_0^{N-1} \,|\, u_0^{i-1}, 1, u_{i+1}^{N-1})$ and compute $L_N^{(i)}(y_0^{N-1}, u_0^{i-1})$. We repeat this process over many samples and take the average. It is clear that, as we increase the number of samples, the estimates converge to their actual values. By replacing $\sqrt{L_N^{(i)}}$ with $\frac{1}{1+L_N^{(i)}}$ we can estimate the probability of error $P_e(W_N^{(i)})$. We use the latter quantity for our simulations.

Another alternative is to use density evolution tools developed for the analysis of message passing algorithms. It is well known that density evolution [30] computes the effective channel seen by the root of a tree code. The channel $W_N^{(i)}$ can be represented as the effective channel seen by the root variable of an appropriately defined tree code. First restrict W to be symmetric. Let $b_1 \dots b_n$ be the n-bit binary expansion of i. The tree code consists of $n+1$ levels, namely $0, \dots, n$. The root is at level n. At level i we have 2^{n-i} nodes, all of which are either check nodes if $b_i = 0$ or variable nodes if $b_i = 1$. An example for $W_8^{(5)}$ is shown in Figure 2.7.

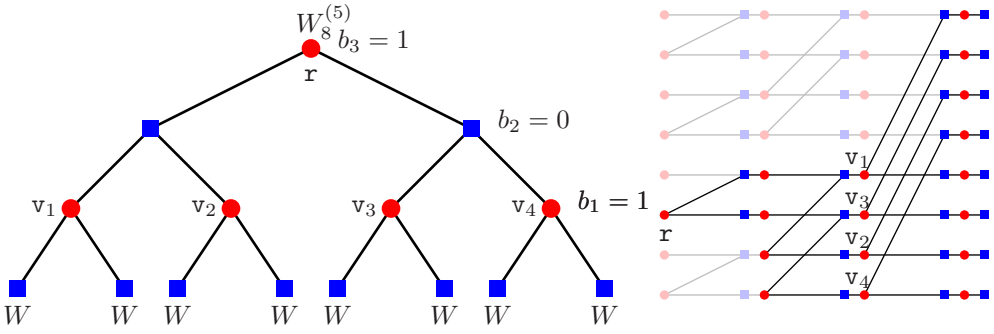

Figure 2.7: Tree representation of the channel $W_8^{(5)}$. The 3-bit binary expansion of 5 is $b_1 b_2 b_3 = 101$. The right figure shows the tree embedded in the factor graph of the code. The root variable $\mathtt{r}$ and the variables $\mathtt{v}_1, \dots, \mathtt{v}_4$ are marked for reference. The shaded part of the factor graph is removed because U_0^4 is known but U_6^7 is not known.

The recursions in (2.11), imply that the channel law of $W_N^{(i)}$ is equal to the effective channel seen by the root node when the leaves are transmitted through the channel W. The tools developed for density evolution [51, Appendix B] can be used for an efficient computation of the channel law of $W_N^{(i)}$.

To use density evolution it is convenient to represent the channel in the log-likelihood domain. We refer the reader to [51] for a detailed description of density evolution. The B-DMC W is represented as a probability distribution over $\mathbb{R} \cup \{\pm\infty\}$. The probability distribution is the distribution of the variable $\log(\frac{W(Y|0)}{W(Y|1)})$ where Y is distributed as $W(y \,|\, 0)$. During implementation, in

order to keep the complexity in check, the space $\mathbb{R}$ is quantized. If the number of quantization levels are fixed, then using similar arguments as in the case of the BEC, we claim that the complexity of estimating the log-likelihood densities of $\{W_N^{(i)}\}$ is $O(N)$ [52]. As n increases (the number of levels in the tree) the quantization errors accumulate and hence, the number of quantization levels should scale with n. It is an interesting open question to find what the scaling should be.

If W is not symmetric, then we use the symmetrized channel W_s at the leaves. Recall from Definition 1.3, the channel W_s is defined as

$$W_s(y, z \,|\, x) = \frac{1}{2} W(y \,|\, z \oplus x).$$

It can be shown that $I(W_{sN}^{(i)}) = I(W_N^{(i)})$ and $Z(W_{sN}^{(i)}) = Z(W_N^{(i)})$. Therefore, we can accomplish the task by using density evolution for the channel W_s.

2.6 Simulation Results and Discussion

Let us now look at the finite-length performance of polar codes for communication over the binary input additive white Gaussian noise channel (BAWGNC) and the BEC. Even though we considered only discrete channels throughout our discussion, the results can be extended to channels with continuous outputs by replacing the summations over the output alphabet with integrals.

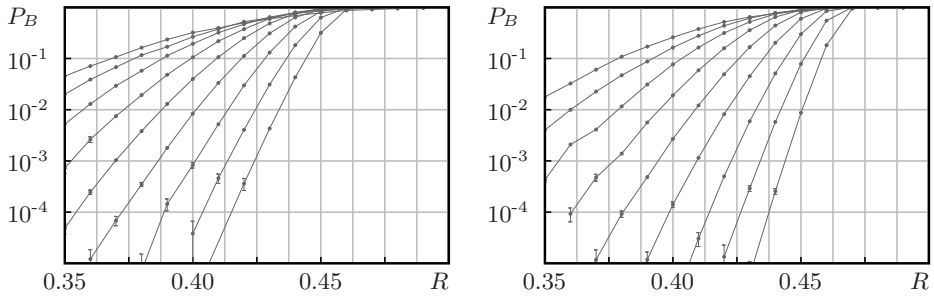

Figure 2.8: Performance of SC decoder in terms of block error probability, when transmission takes place over the BAWGNC($\sigma = 0.97865$) (left) and BEC(0.5) (right). The performance curves are shown for $n = 10, 11, \ldots, 18$.

Recall that the length N of the code is always a power of 2, i.e., $N = 2^n$. Figure 2.8 shows the performance of polar codes under SC decoding for BAWGNC(0.97865) and BEC(0.5). Both the channels have capacity $\frac{1}{2}$. The frozen set for BAWGNC are chosen using the estimates for $P_e(W_N^{(i)})$. These estimates are done using the Monte-Carlo method described in Section 2.5.2. Since the channel is symmetric, we can restrict u_0^{N-1} to zero.

For all the simulation points in the plots, the 95% confidence intervals are shown. In most cases these confidence intervals are smaller than the point size

and are therefore not visible. From the simulations we see that the performance at small lengths is not impressive. In Chapter 6 we discuss some approaches based on BP which significantly improve the empirical performance. However, the resulting performance is still not comparable to that of the best known LDPC codes.

2.A Appendix

The following lemmas have appeared in [51] in the context of the analysis of the BP decoder for LDPC codes. These lemmas were also proved by Arıkan in [32].

Lemma 2.15 (Upper Bound on $Z(W_1 \circledast W_2)$ [51, 32])**.** *Let $W_1 : \mathcal{X} \to \mathcal{Y}_1$ and $W_2 : \mathcal{X} \to \mathcal{Y}_2$ be two B-DMCs. Then*

$$Z(W_1 \circledast W_2) \leq Z(W_1) + Z(W_2) - Z(W_1)Z(W_2).$$

Proof. Let $Z = Z(W_1 \circledast W_2)$ and $Z_i = Z(W_i)$. Z can be expanded as follows.

$$\begin{aligned}
Z &= \sum_{y_1,y_2} \sqrt{W_1 \circledast W_2(y_1, y_2 \,|\, 0) W_1 \circledast W_2(y_1, y_2 \,|\, 1)} \\
&= \frac{1}{2} \sum_{y_1,y_2} \Big[W_1(y_1 \,|\, 0) W_2(y_2 \,|\, 0) W_1(y_1 \,|\, 0) W_2(y_2 \,|\, 1) \\
&\qquad + W_1(y_1 \,|\, 0) W_2(y_2 \,|\, 0) W_1(y_1 \,|\, 1) W_2(y_2 \,|\, 0) \\
&\qquad + W_1(y_1 \,|\, 1) W_2(y_2 \,|\, 1) W_1(y_1 \,|\, 0) W_2(y_2 \,|\, 1) \\
&\qquad + W_1(y_1 \,|\, 1) W_2(y_2 \,|\, 1) W_1(y_1 \,|\, 1) W_2(y_2 \,|\, 0) \Big]^{\frac{1}{2}} \\
&= \frac{Z_1 Z_2}{2} \sum_{y_1,y_2} P_1(y_1) P_2(y_2) \\
&\qquad \sqrt{\frac{W_1(y_1 \,|\, 0)}{W_1(y_1 \,|\, 1)} + \frac{W_1(y_1 \,|\, 1)}{W_1(y_1 \,|\, 0)} + \frac{W_2(y_2 \,|\, 0)}{W_2(y_2 \,|\, 1)} + \frac{W_2(y_2 \,|\, 1)}{W_2(y_2 \,|\, 0)}}
\end{aligned}$$

where $P_i(y_i)$ denotes

$$P_i(y_i) = \frac{\sqrt{W_i(y_i \,|\, 0) W_i(y_i \,|\, 1)}}{Z_i}.$$

Note that P_i is a probability distribution over $\mathcal{Y}_i$. Let $\mathbb{E}_i$ denote the expectation with respect to P_i and let

$$A_i(y) \triangleq \sqrt{\frac{W_i(y \,|\, 0)}{W_i(y \,|\, 1)}} + \sqrt{\frac{W_i(y \,|\, 1)}{W_i(y \,|\, 0)}}.$$

Then Z can be expressed as

$$Z = \frac{Z_1 Z_2}{2} \mathbb{E}_{1,2} \left[\sqrt{(A_1(Y_1))^2 + (A_2(Y_2))^2 - 4} \right].$$

If $a \geq c$ and $b \geq c$, then $\sqrt{a+b-c} \leq \sqrt{a} + \sqrt{b} - \sqrt{c}$. Applying this inequality with $a = (A_1(Y_1))^2$, $b = (A_2(Y_2))^2$, and $c = 4$, we get

$$Z \leq \frac{Z_1 Z_2}{2} \left(\mathbb{E}_1[A_1(Y_1)] + \mathbb{E}_2[A_2(Y_2)] - 2 \right)$$

The claims follow by substituting $\mathbb{E}_i[A_i(Y_i)] = \frac{2}{Z_i}$. □

Lemma 2.16 ($Z(W_1 \circledast W_2)$ [51, 32])**.** *Let $W_1 : \mathcal{X} \to \mathcal{Y}_1$ and $W_2 : \mathcal{X} \to \mathcal{Y}_2$ be two B-DMCs. Then*

$$Z(W_1 \circledast W_2) = Z(W_1) Z(W_2).$$

Proof. Let $Z = Z(W_1 \circledast W_2)$ and $Z_i = Z(W_i)$. Z can be expanded as follows.

$$\begin{aligned}
Z &= \sum_{y_1, y_2, x} \sqrt{W_1 \circledast W_2(y_1, y_2, x \,|\, 0) W_1 \circledast W_2(y_1, y_2, x \,|\, 1)} \\
&= \frac{1}{2} \sum_{y_1, y_2, x} \sqrt{W_1(y_1 \,|\, x \oplus 0) W_2(y_2 \,|\, 0) W_1(y_1 \,|\, x \oplus 1) W_2(y_2 \,|\, 1)} \\
&= \frac{1}{2} \sum_{y_2, x} \sqrt{W_2(y_2 \,|\, 0) W_2(y_2 \,|\, 1)} \sum_{y_1} \sqrt{W_1(y_1 \,|\, x \oplus 0) W_1(y_1 \,|\, x \oplus 1)} \\
&= \frac{1}{2} \sum_{y_2, x} \sqrt{W_2(y_2 \,|\, 0) W_2(y_2 \,|\, 1)} Z_1 \\
&= Z_1 Z_2.
\end{aligned}$$

□

3 Source Coding

In this chapter, we construct polar codes that achieve Shannon's rate-distortion bound for a large class of sources using an SC encoding algorithm of complexity $O(N \log(N))$. This is arguably the central result of this thesis.

We start with a discussion of the symmetric rate-distortion function (Section 3.1). It is the equivalent of the symmetric mutual information for the case of channel coding. We then introduce the successive cancellation encoder in Section 3.2. The main result is given in Section 3.3. In Section 3.4 we show that for some class of sources all the codes in a given polar code ensemble have the same performance. This simplifies the code design problem considerably. We end in Section 3.5 by presenting some simulation results and discussing the relationship to channel coding. This chapter is based on the results obtained in [53].

3.1 Rate-Distortion

We model the source as a sequence of i.i.d. realizations of a random variable $Y \in \mathcal{Y}$. Let $\mathcal{X}$ denote the reconstruction space. Let $\mathtt{d} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$ denote the distortion function with maximum value $\mathtt{d}_{\max}$. The distortion function naturally extends to vectors as $\mathtt{d}(\bar{y}, \bar{x}) = \sum_{i=0}^{N-1} \mathtt{d}(y_i, x_i)$.

Let $\mathtt{C}_N \subseteq \mathcal{X}^N$ be the code used for compression. The encoder maps a source sequence $\bar{Y}$ to an index $f_N(\bar{Y})$. The index refers to a codeword $\bar{X} \in \mathtt{C}_N$. Therefore, the number of bits required for the encoder to convey the index is at most $\log |\mathtt{C}_N|$, The decoder reconstructs the codeword $\bar{X}$ from the index. Let g_N denote the reconstruction function. The rate of such a scheme is given by $\frac{\log |\mathtt{C}_N|}{N}$ and the resulting average distortion is given by $D_N = \frac{1}{N}\mathbb{E}[\mathtt{d}(\bar{Y}, g_N(f_N(\bar{Y})))]$.

Shannon's rate-distortion theorem [2] characterizes the minimum rate required to achieve a given distortion.

Theorem 3.1 (Shannon's Rate-Distortion Theorem [2])**.** *Consider an i.i.d. source Y with probability distribution $P_Y(y)$. To achieve an average distortion D the required rate is at least*

$$R(D) = \min_{p(y,x):\mathbb{E}_p[\mathtt{d}(y,x)]\leq D, p(y)=P_Y(y)} I(Y;X). \tag{3.1}$$

Moreover, for any $R > R(D)$ there exist a sequence of codes $\mathtt{C}_N$ and functions f_N and g_N such that $|\mathtt{C}_N| \leq 2^{NR}$ and the average distortion D_N approaches D.

Let us define the symmetric rate-distortion function.

Definition 3.2 (Symmetric Rate-Distortion)**.** *For a random variable Y with probability distribution $P_Y(y)$ and a distortion function $\mathtt{d} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$, the symmetric rate distortion function $R_s(D)$ is defined as*

$$R_s(D) = \min_{p(y,x):\mathbb{E}_p[\mathtt{d}(y,x)]\leq D, p(y)=P_Y(y), p(x)=\frac{1}{|\mathcal{X}|}} I(Y;X). \tag{3.2}$$

In the definition of $R_s(D)$ we have the additional constraint that the induced probability distribution over $\mathcal{X}$ must be uniform. Clearly, $R_s(D) \geq R(D)$. The quantity $R_s(D)$ is similar in spirit to the symmetric capacity of a channel which is defined as the mutual information between the input and output of the channel when the input distribution is uniform. The significance of $R_s(D)$ is that, for any $R > R_s(D)$, there exist *affine* codes with rate at most R that achieve an average distortion D.

We will restrict our discussion to the case of a binary reconstruction alphabet, i.e., $\mathcal{X} = \{0, 1\}$. Let $p^*(y, x)$ be a probability distribution that minimizes (3.2). From the definition of $R_s(D)$, the distribution $p^*(y, x)$ is such that its marginal over $\mathcal{Y}$ is P_Y and its marginal over $\mathcal{X}$ is uniform. The conditional distribution $p^*(y \,|\, x)$ plays the role of a channel. It is therefore customary to refer to the distribution $p^*(y \,|\, x)$ as *test channel* and denote it by $W(y \,|\, x)$.

Example 3.3 (Binary Symmetric Source)**.** *Consider a Ber($\frac{1}{2}$) source Y. The test channel that achieves the rate-distortion bound is the BSC(D). The distribution induced on X by the test channel is Ber($\frac{1}{2}$), which is uniform. Therefore, $R_s(D) = R(D)$, which is equal to $1 - h_2(D)$.*

In the following, the length of the polar code is always denoted by N. For the sake of exposition, we use the notation $\bar{u}$ to denote the row vector $\bar{u} = (u_0, \ldots, u_{N-1})$ and $\bar{U}$ to denote the row vector $(U_0, \ldots, U_{N-1})$.

3.2 Successive Cancellation Encoder

Let the source be a sequence of i.i.d. realizations of a random variable $Y \in \mathcal{Y}$ with probability distribution P_Y. Let W be the test channel that achieves

$R_s(D)$ for design distortion D. From the Definition 3.2 we conclude the following.

$$R_s(D) = I(W), \tag{3.3}$$

$$\sum_{x,y} \frac{1}{2} W(y \,|\, x) \mathtt{d}(y, x) = D, \tag{3.4}$$

$$\sum_{x} \frac{1}{2} W(y \,|\, x) = P_Y(y). \tag{3.5}$$

Let us consider using a polar code $\mathtt{C}_N(F, u_F)$ (c.f. Definition 2.7) for source coding. The frozen set F as well as the vector u_F are known both to the encoder and the decoder. A source word $\bar{Y}$ is mapped to a codeword $\bar{X} \in \mathtt{C}_N(F, u_F)$. This codeword can be described by the index $u_{F^c} = (\bar{X}(G_2^{\otimes n})^{-1})_{F^c}$. Therefore, the required rate is $\frac{|F^c|}{N}$. Since the decoder knows u_F in advance, it recovers the codeword $\bar{X}$ by performing the operation $\bar{u}G_2^{\otimes n}$.

Let us recall some notation from the previous chapter. Consider the probability distribution $P_{\bar{U},\bar{X},\bar{Y}}$ over the space $\mathcal{X}^N \times \mathcal{X}^N \times \mathcal{Y}^N$ as defined in (2.5),

$$P_{\bar{U},\bar{X},\bar{Y}}(\bar{u}, \bar{x}, \bar{y}) = \underbrace{\frac{1}{2^N}}_{P_{\bar{U}}(\bar{u})} \underbrace{\mathbb{1}_{\{\bar{x}=\bar{u}G_2^{\otimes n}\}}}_{P_{\bar{X}\,|\,\bar{U}}(\bar{x}\,|\,\bar{u})} \underbrace{\prod_{i=0}^{N-1} W(y_i \,|\, x_i)}_{P_{\bar{Y}\,|\,\bar{X}}(\bar{y}\,|\,\bar{x})}. \tag{3.6}$$

Figure 3.1 provides a pictorial description of the relationship between $\bar{U}, \bar{X}$ and $\bar{Y}$. Let us verify that it is a valid probability distribution. Since $G_2^{\otimes n}$ is an invertible matrix, the uniform distribution of $\bar{U}$ over $\mathcal{X}^N$ induces a uniform distribution for $\bar{X}$. From property (3.5) it follows that the marginal induced by the above distribution over the space $\mathcal{Y}^N$ is indeed $\prod_{i=0}^{N-1} P_Y(y_i)$, where P_Y is the distribution of the source. Recall that $W_N^{(i)} : \mathcal{X} \to \mathcal{Y}^N \times \mathcal{X}^{i-1}$ denotes the channel with input u_i, output $(\bar{y}, u_0^{i-1})$, and transition probabilities given by

$$W_N^{(i)}(\bar{y}, u_0^{i-1} \,|\, u_i) = \frac{1}{2^{N-1}} \sum_{u_{i+1}^{N-1}} W_N(\bar{y} \,|\, \bar{u}). \tag{3.7}$$

Also recall from (2.12) that

$$L_N^{(i)}(y_0^{N-1}, u_0^{i-1}) = \frac{W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, 0)}{W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, 1)}.$$

Let $\bar{y}$ denote N i.i.d. realizations of the source Y. Let $\hat{U}(\bar{y}, u_F)$ denote the result of the following *encoding* operation using the code $\mathtt{C}_N(F, u_F)$. Given $\bar{y}$, for each i in the range 0 till $N-1$:

(i) If $i \in F$, then set $\hat{u}_i = u_i$.

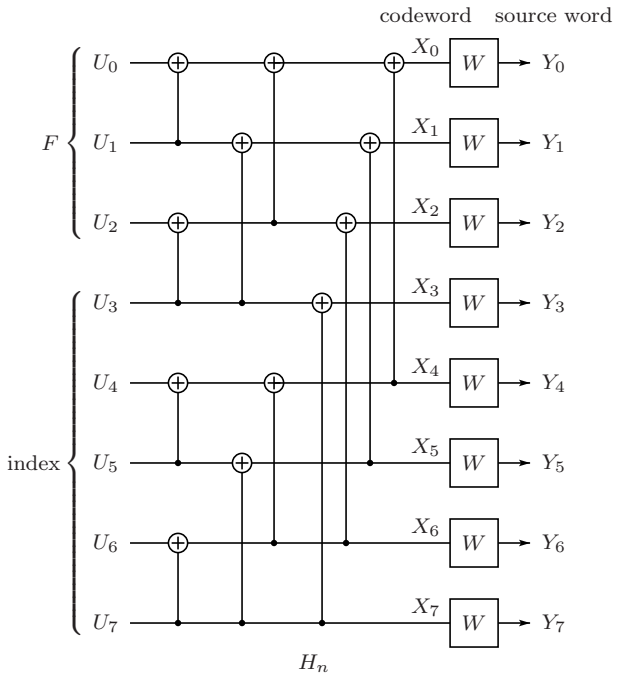

Figure 3.1: The probabilistic model used for source coding. The source word $\bar{Y}$ is treated as the output of a channel. The code is defined through its frozen set $F = \{0, 1, 2\}$. The codewords are indexed using U_3^7.

(ii) If $i \in F^c$, then compute $L_N^{(i)}(\bar{y}, \hat{u}_0^{i-1})$ and set

$$\hat{u}_i = \begin{cases} 0 & \text{w.p. } \frac{L_N^{(i)}}{1+L_N^{(i)}}, \\ 1 & \text{w.p. } \frac{1}{1+L_N^{(i)}}. \end{cases}$$

Note that unlike in channel coding, where we made a hard decision based on $L_N^{(i)}$ (MAP rule), $\hat{u}_i$ is chosen randomly with a probability based on $L_N^{(i)}$. We refer to this decision rule as *randomized rounding*. Randomized rounding as a decimation rule is not new. In [54] it was applied in the context of finding solutions of a random k-SAT problem.

Why do we use randomized rounding? In simulations, randomized rounding and the MAP rule perform similarly with a slight performance edge for the MAP rule. But for the purpose of analysis the randomized rounding rule is much more convenient. In fact, it is currently not clear if and how the MAP rule can be analyzed. Note that all of the existing source coding schemes use the MAP rule. This is most likely the main obstacles to their analysis. We believe that by combining randomized rounding with existing schemes like BID it might be possible to analyze the performance of LDGM codes for source coding.

Note that the probability $\frac{L_N^{(i)}(\bar{y},\hat{u}_0^{i-1})}{1+L_N^{(i)}(\bar{y},\hat{u}_0^{i-1})}$ can be rewritten as

$$\begin{aligned}\frac{L_N^{(i)}(\bar{y},\hat{u}_0^{i-1})}{1+L_N^{(i)}(\bar{y},\hat{u}_0^{i-1})} &\overset{(3.7)}{=} \frac{\sum_{u_{i+1}^{N-1}} W_N(\bar{y}\,|\,\hat{u}_0^{i-1},0,u_{i+1}^{N-1})}{\sum_{u_i^{N-1}} W_N(\bar{y}\,|\,\hat{u}_0^{i-1},u_i^{N-1})}\\ &= \frac{\sum_{u_{i+1}^{N-1}} P_{\bar{U}\,|\,\bar{Y}}(\hat{u}_0^{i-1},0,u_{i+1}^{N-1}\,|\,\bar{y})}{\sum_{u_i^{N-1}} P_{\bar{U}\,|\,\bar{Y}}(\hat{u}_0^{i-1},u_i^{N-1}\,|\,\bar{y})}\\ &= P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0\,|\,\hat{u}_0^{i-1},\bar{y}).\end{aligned}$$

In words, it is equal to the posterior of $U_i = 0$ given $(\bar{Y} = \bar{y}, U_0^{i-1} = \hat{u}_0^{i-1})$ under the distribution P.

The decoding, or the reconstruction operation, is given by $\bar{x} = \hat{u}G_2^{\otimes n}$. The decoder has knowledge of $\hat{u}_F$ (since $\hat{u}_F = u_F$) and hence the encoder needs to convey only the vector $(\hat{U}(\bar{y},u_F))_{F^c}$ to the decoder. This requires $|F^c|$ bits.

The average distortion incurred by this scheme is given by $\frac{1}{N}\mathbb{E}[\mathtt{d}(\bar{Y},\bar{X})]$, where the expectation is over the source randomness and the randomness involved in the randomized rounding at the encoder.

For lossy source compression, the SC operation is employed at the encoder side to map the source vector to a codeword. Therefore, from now onwards we refer to this operation as SC *encoding*. The encoding (decoding) task for source coding is the same as the decoding (encoding) task for channel coding. Therefore, as shown in the previous chapter, both the operations can be implemented with complexity $O(N\log N)$.

3.3 Polar Codes Achieve the Rate-Distortion Bound

Let us now show that properly designed polar codes with SC encoding are optimal.

Theorem 3.4 (Polar Codes Achieve the Symmetric Rate-Distortion Bound)**.** *Consider an i.i.d. source Y and a distortion function $\mathtt{d}: \mathcal{Y}\times\mathcal{X}\to\mathbb{R}$. Fix a distortion D and $0<\beta<\frac{1}{2}$. For any rate $R > R_s(D)$, there exists a sequence of polar codes of length N and rate $R_N \le R$, so that under SC encoding using randomized rounding they achieve expected distortion D_N satisfying*

$$D_N \le D + O(2^{-(N^\beta)}).$$

The encoding as well as decoding complexity of these codes is $O(N\log(N))$.

Note that the encoding function $\hat{U}(\bar{y},u_F)$ is random, i.e., the encoding function may result in different outputs for the same input. More precisely, in step i of the encoding process, $i\in F^c$, $\hat{U}_i = 0$ with probability proportional to

the posterior (randomized rounding) $P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, \hat{u}_0^{i-1}, \bar{y})$. This implies that the probability of picking a vector $\hat{u}_0^{N-1}$ given $\bar{y}$ is equal to

$$\begin{cases} 0, & \text{if } \hat{u}_F \neq u_F, \\ \prod_{i \in F^c} P_{U_i \,|\, U_0^{i-1},\bar{Y}}(\hat{u}_i \,|\, \hat{u}_0^{i-1}, \bar{y}), & \text{if } \hat{u}_F = u_F. \end{cases}$$

Therefore, the average (over $\bar{y}$ and the randomness of the encoder) distortion for the code $\mathtt{C}_N(F, u_F)$ is given by

$$D_N(F, u_F) = \sum_{\bar{y}} P_{\bar{Y}}(\bar{y}) \sum_{\hat{u}_{F^c}} \left(\prod_{i \in F^c} P(\hat{u}_i \,|\, \hat{u}_0^{i-1}, \bar{y}) \right) \mathtt{d}(\bar{y}, \hat{u}_0^{N-1} G_2^{\otimes n}), \tag{3.8}$$

where $\hat{u}_F = u_F$.

We want to show that there exists a set F of cardinality roughly $N(1 - R_s(D))$ and a vector u_F such that $D_N(F, u_F) \approx ND$. This will show that polar codes achieve the rate-distortion bound. For the proof it is convenient not to determine the distortion for a fixed choice of u_F but to compute the average distortion over the ensemble $\mathtt{C}_N(F)$ (c.f. Definition 2.8) with a uniform distribution over the codes in the ensemble, i.e., a uniform distribution over the choices of u_F. We will show that this average distortion over the ensemble is roughly ND. This implies that for at least one choice of u_F the distortion is as low as ND, leading to the desired result. Later, in Section 3.4, we will see that for some cases the distortion *does not depend* on the choice of u_F. In this case a convenient choice is to set u_F to zero.

Let us therefore start by computing the *average distortion*, call it $D_N(F)$. We will show that $D_N(F)$ is close to D.

The distortion $D_N(F)$ can be written as

$$\begin{aligned} D_N(F) &= \sum_{u_F \in \mathcal{X}^{|F|}} \frac{1}{2^{|F|}} D_N(F, u_F) \\ &= \sum_{u_F} \frac{1}{2^{|F|}} \sum_{\bar{y}} P_{\bar{Y}}(\bar{y}) \sum_{\bar{u}_{F^c}} \left(\prod_{i \in F^c} P(u_i \,|\, u_0^{i-1}, \bar{y}) \right) \mathtt{d}(\bar{y}, \bar{u} G_2^{\otimes n}) \\ &= \sum_{\bar{y}} P_{\bar{Y}}(\bar{y}) \sum_{\bar{u}} \frac{1}{2^{|F|}} \left(\prod_{i \in F^c} P(u_i \,|\, u_0^{i-1}, \bar{y}) \right) \mathtt{d}(\bar{y}, \bar{u} G_2^{\otimes n}). \end{aligned} \tag{3.9}$$

Let $Q_{\bar{U},\bar{Y}}$ denote the distribution defined by $Q_{\bar{Y}}(\bar{y}) = P_{\bar{Y}}(\bar{y})$ and $Q_{\bar{U} \,|\, \bar{Y}}$ defined by

$$Q(u_i \,|\, u_0^{i-1}, \bar{y}) = \begin{cases} \frac{1}{2}, & \text{if } i \in F, \\ P_{U_i \,|\, U_0^{i-1},\bar{Y}}(u_i \,|\, u_0^{i-1}, \bar{y}), & \text{if } i \in F^c. \end{cases} \tag{3.10}$$

Then (3.9) is equivalent to

$$D_N(F) = \mathbb{E}_Q[\mathtt{d}(\bar{Y}, \bar{U} G_2^{\otimes n})],$$

where $\mathbb{E}_Q[\cdot]$ denotes the expectation with respect to the distribution $Q_{\bar{U},\bar{Y}}$. Similarly, let $\mathbb{E}_P[\cdot]$ denote the expectation with respect to the distribution $P_{\bar{U},\bar{Y}}$. We can write $P_{\bar{U}\,|\,\bar{Y}}$ in the form

$$P_{\bar{U}\,|\,\bar{Y}}(\bar{u}\,|\,\bar{y}) = \prod_{i=0}^{N-1} P_{U_i\,|\,U_0^{i-1},\bar{Y}}(u_i\,|\,u_0^{i-1},\bar{y}).$$

If we compare Q to P we see that they have the same structure except for the components $i \in F$. Indeed, in the following lemma we show that the total variation distance between Q and P can be bounded in terms of how much the posteriors $Q_{U_i\,|\,U_0^{i-1},\bar{Y}}$ and $P_{U_i\,|\,U_0^{i-1},\bar{Y}}$ differ for $i \in F$.

Lemma 3.5 (Bound on the Total Variation Distance)**.** *Let F denote the set of frozen indices and let the probability distributions Q and P be as defined above. Then*

$$\sum_{\bar{u},\bar{y}} |Q(\bar{u},\bar{y}) - P(\bar{u},\bar{y})| \le 2 \sum_{i\in F} \mathbb{E}_P\left[\Big|\frac{1}{2} - P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0\,|\,U_0^{i-1},\bar{Y})\Big|\right].$$

Proof.

$$\begin{aligned}
&\sum_{\bar{u}} |Q(\bar{u}\,|\,\bar{y}) - P(\bar{u}\,|\,\bar{y})| \\
&= \sum_{\bar{u}} \Big| \prod_{i=0}^{N-1} Q(u_i\,|\,u_0^{i-1},\bar{y}) - \prod_{i=0}^{N-1} P(u_i\,|\,u_0^{i-1},\bar{y})\Big| \\
&= \sum_{\bar{u}} \Big| \sum_{i=0}^{N-1} \Big[\big(Q(u_i\,|\,u_0^{i-1},\bar{y}) - P(u_i\,|\,u_0^{i-1},\bar{y})\big) \cdot \\
&\quad \Big(\prod_{j=0}^{i-1} P(u_j\,|\,u_0^{j-1},\bar{y})\Big)\Big(\prod_{j=i+1}^{N-1} Q(u_j\,|\,u_0^{j-1},\bar{y})\Big)\Big]\Big|.
\end{aligned}$$

In the last step we have used the following telescoping expansion:

$$A_0^{N-1} - B_0^{N-1} = \sum_{i=0}^{N-1} \Big(A_i^{N-1}B_0^{i-1} - A_{i+1}^{N-1}B_0^{i}\Big) = \sum_{i=0}^{N-1} (A_i - B_i) A_{i+1}^{N-1} B_0^{i-1},$$

where A_k^j denotes the product $\prod_{i=k}^{j} A_i$ and $A_k^j = 1$ if $j < k$. For example, if $N = 3$, the above expansion is equivalent to

$$\begin{aligned}
A_0A_1A_2 - B_0B_1B_2 =& (A_0A_1A_2 - A_1A_2B_0) + (A_1A_2B_0 - A_2B_0B_1) \\
&+ (A_2B_0B_1 - B_0B_1B_2).
\end{aligned}$$

Now note that if $i \in F^c$ then $Q(u_i\,|\,u_0^{i-1},\bar{y}) = P(u_i\,|\,u_0^{i-1},\bar{y})$, so that these terms vanish. The above sum therefore reduces to

$$\sum_{\bar{u}} \Big| \sum_{i\in F} \Big[\underbrace{\big(Q(u_i\,|\,u_0^{i-1},\bar{y}) - P(u_i\,|\,u_0^{i-1},\bar{y})\big)}_{\le |\,\frac{1}{2} - P(u_i\,|\,u_0^{i-1},\bar{y})\,|} \cdot$$

$$\Big(\prod_{j=0}^{i-1} P(u_j \,|\, u_0^{j-1}, \bar{y})\Big)\Big(\prod_{j=i+1}^{N-1} Q(u_j \,|\, u_0^{j-1}, \bar{y})\Big)\Big]\Big|$$

$$\leq \sum_{i\in F}\sum_{\bar{u}_0^i} \Big|\frac{1}{2} - P(u_i \,|\, u_0^{i-1}, \bar{y})\Big| \prod_{j=0}^{i-1} P(u_j \,|\, u_0^{j-1}, \bar{y})$$

$$\leq 2\sum_{i\in F} \mathbb{E}_{P_{\bar{U}\,|\,\bar{Y}=\bar{y}}}\left[\Big|\frac{1}{2} - P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1}, \bar{y})\Big|\right].$$

In the last step the summation over u_i gives rise to the factor 2, whereas the summation over u_0^{i-1} gives rise to the expectation.

Note that $Q_{\bar{Y}}(\bar{y}) = P_{\bar{Y}}(\bar{y})$ by definition. The claim follows by taking the expectation over $\bar{Y}$. □

Lemma 3.6 (Distortion under Q versus Distortion under P)**.** *Let F be chosen such that for $i \in F$*

$$\mathbb{E}_P\left[\Big|\frac{1}{2} - P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1}, \bar{Y})\Big|\right] \leq \delta_N. \tag{3.11}$$

The average distortion is then bounded by

$$\frac{1}{N}\mathbb{E}_Q[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] \leq \frac{1}{N}\mathbb{E}_P[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] + |F|2\mathtt{d}_{\max}\delta_N,$$

where $\mathtt{d}_{\max} = \max_{\{y\in\mathcal{Y}, x\in\mathcal{X}\}} \mathtt{d}(y, x)$.

Proof.

$$\begin{aligned}
&\mathbb{E}_Q[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] - \mathbb{E}_P[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] \\
&\quad= \sum_{\bar{u},\bar{y}} \Big(Q(\bar{u}, \bar{y}) - P(\bar{u}, \bar{y})\Big)\mathtt{d}(\bar{y}, \bar{u}G_2^{\otimes n}) \\
&\quad\leq N\mathtt{d}_{\max} \sum_{\bar{u},\bar{y}} \Big|Q(\bar{u}, \bar{y}) - P(\bar{u}, \bar{y})\Big| \\
&\quad\overset{\text{Lem. 3.5}}{\leq} 2N\mathtt{d}_{\max} \sum_{i\in F} \mathbb{E}_P\left[\Big|\frac{1}{2} - P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1}, \bar{Y})\Big|\right] \\
&\quad\leq |F|2N\mathtt{d}_{\max}\delta_N.
\end{aligned}$$

□

From Lemma 3.6 we see that the average distortion $D_N(F)$ is upper bounded by the average distortion with respect to P plus a term which bounds the "distance" between Q and P.

Lemma 3.7 (Distortion under P)**.**

$$\mathbb{E}_P[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] = ND.$$

Proof. Let $\bar{X} = \bar{U}G_2^{\otimes n}$ and write

$$\begin{aligned}
\mathbb{E}_P[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] &= \sum_{\bar{u},\bar{y}} P_{\bar{U},\bar{Y}}(\bar{u},\bar{y})\ \mathtt{d}(\bar{y}, \bar{u}G_2^{\otimes n}) \\
&= \sum_{\bar{y},\bar{u},\bar{x}} P_{\bar{U},\bar{X},\bar{Y}}(\bar{u},\bar{x},\bar{y})\ \mathtt{d}(\bar{y}, \bar{u}G_2^{\otimes n}) \\
&= \sum_{\bar{y},\bar{u},\bar{x}} P_{\bar{X},\bar{Y}}(\bar{x},\bar{y}) \underbrace{P_{\bar{U}\,|\,\bar{X},\bar{Y}}(\bar{u}\,|\,\bar{x},\bar{y})}_{\mathbb{1}_{\{\bar{x}=\bar{u}G_2^{\otimes n}\}}}\ \mathtt{d}(\bar{y},\bar{x}) \\
&= \sum_{\bar{y},\bar{x}} P_{\bar{X},\bar{Y}}(\bar{x},\bar{y})\ \mathtt{d}(\bar{y},\bar{x}).
\end{aligned}$$

Note that the unconditional distribution of $\bar{X}$ is the uniform one and that the channel between $\bar{X}$ and $\bar{Y}$ is memoryless and identical for each component. Therefore, we can write this expectation as

$$\begin{aligned}
\mathbb{E}_P[\mathtt{d}(\bar{Y}, \bar{U}G_2^{\otimes n})] &= N \sum_{x_0,y_0} P_{X_0,Y_0}(x_0,y_0)\ \mathtt{d}(y_0,x_0) \\
&\overset{(a)}{=} N \sum_{x_0,y_0} \frac{1}{2} W(y_0\,|\,x_0)\ \mathtt{d}(y_0,x_0) \\
&\overset{(3.4)}{=} ND,
\end{aligned}$$

where (a) follows from the fact that $P_{Y\,|\,X}(y\,|\,x) = W(y\,|\,x)$. □

This implies that if we use all the variables $\{U_i\}$ to represent the source word, i.e., if F is empty, then the algorithm results in an average distortion D, as desired. But the rate of such a code would be 1. Fortunately, the rate problem is easily fixed. If we choose F to consist of those variables which are "essentially random," (i.e., satisfying (3.11)) then there is only a small distortion penalty (namely, $|F|2\delta_N$) to pay with respect to the previous case and the rate has been decreased to $1 - |F|/N$.

Lemma 3.6 shows that the guiding principle for choosing the set F is to include the indices with small δ_N in (3.11). In the following lemma we find a sufficient condition, for an index to satisfy (3.11), which is easier to handle.

Lemma 3.8 ($Z(W_N^{(i)})$ Close to 1 is Good)**.** *If* $Z(W_N^{(i)}) \geq 1 - 2\delta_N^2$, *then*

$$\mathbb{E}_P\left[\left|\frac{1}{2} - P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0\,|\,U_0^{i-1},\bar{Y})\right|\right] \leq \delta_N.$$

Proof.

$$\begin{aligned}
&\mathbb{E}_P\left[\sqrt{P_{U_i\,|\,U_0^{i-1},\bar{Y}}(0\,|\,U_0^{i-1},\bar{Y}) P_{U_i\,|\,U_0^{i-1},\bar{Y}}(1\,|\,U_0^{i-1},\bar{Y})}\right] \\
&\quad = \sum_{u_0^{i-1},\bar{y}} P_{U_0^{i-1},\bar{Y}}(u_0^{i-1},\bar{y})
\end{aligned}$$

$$
\begin{aligned}
&\sqrt{P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, u_0^{i-1},\bar{y}) P_{U_i \,|\, U_0^{i-1},\bar{Y}}(1 \,|\, u_0^{i-1},\bar{y})} \\
&= \sum_{u_0^{i-1},\bar{y}} \sqrt{P_{U_0^{i-1},U_i,\bar{Y}}(u_0^{i-1},0,\bar{y}) P_{U_0^{i-1},U_i,\bar{Y}}(u_0^{i-1},1,\bar{y})} \\
&= \sum_{u_0^{i-1},\bar{y}} \sqrt{\sum_{u_{i+1}^{N-1}} P_{\bar{U},\bar{Y}}((u_0^{i-1},0,u_{i+1}^{N-1}),\bar{y})} \\
&\qquad\qquad \sqrt{\sum_{u_{i+1}^{N-1}} P_{\bar{U},\bar{Y}}((u_0^{i-1},1,u_{i+1}^{N-1}),\bar{y})} \\
&\overset{(a)}{=} \frac{1}{2^N} \sum_{u_0^{i-1},\bar{y}} \sqrt{\sum_{u_{i+1}^{N-1}} P_{\bar{Y}\,|\,\bar{U}}(\bar{y} \,|\, u_0^{i-1},0,u_{i+1}^{N-1})} \\
&\qquad\qquad \sqrt{\sum_{u_{i+1}^{N-1}} P_{\bar{Y}\,|\,\bar{U}}(\bar{y} \,|\, u_0^{i-1},1,u_{i+1}^{N-1})} \\
&= \frac{1}{2} Z(W_N^{(i)}).
\end{aligned}
$$

The equality (a) follows from the fact that $P_{\bar{U}}(\bar{u}) = \frac{1}{2^N}$ for all $\bar{u} \in \{0,1\}^N$.

Assume now that $Z(W_N^{(i)}) \geq 1 - 2\delta_N^2$. Then

$$
\mathbb{E}_P \left[\frac{1}{2} - \sqrt{P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1},\bar{Y}) P_{U_i \,|\, U_0^{i-1},\bar{Y}}(1 \,|\, U_0^{i-1},\bar{Y})} \right] \leq \delta_N^2.
$$

Multiplying and dividing the term inside the expectation with

$$
\frac{1}{2} + \sqrt{P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, u_0^{i-1},\bar{y}) P_{U_i \,|\, U_0^{i-1},\bar{Y}}(1 \,|\, u_0^{i-1},\bar{y})},
$$

and upper bounding this term in the denominator with 1, we get

$$
\mathbb{E}_P \left[\frac{1}{4} - P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1},\bar{Y}) P_{U_i \,|\, U_0^{i-1},\bar{Y}}(1 \,|\, U_0^{i-1},\bar{Y}) \right].
$$

Now, using the equality $\frac{1}{4} - p\bar{p} = (\frac{1}{2} - p)^2$, we get

$$
\mathbb{E}_P \left[\left(\frac{1}{2} - P_{U_i \,|\, U_0^{i-1},\bar{Y}}(0 \,|\, U_0^{i-1},\bar{Y}) \right)^2 \right] \leq \delta_N^2.
$$

The result now follows by applying the Cauchy-Schwartz inequality. □

We are now ready to prove Theorem 3.4. In order to show that there exists a polar code which achieves the rate-distortion trade-off, we show that the size of the set F can be made arbitrarily close to $N(1 - R_s(D))$ while keeping the penalty term $|F|2\delta_N$ arbitrarily small.

Proof of Theorem 3.4:

Let $\beta < \frac{1}{2}$ be a constant and let $\delta_N = \frac{1}{2N\mathtt{d}_{\max}} 2^{-N^\beta}$. Consider a polar code with frozen set F_N,

$$F_N = \{i \in \{0, \ldots, N-1\} : Z(W_N^{(i)}) \geq 1 - 2\delta_N^2\}.$$

For N sufficiently large there exists a $\beta' < \frac{1}{2}$ such that $2\delta_N^2 > 2^{-N^{\beta'}}$. Theorem 3.15 and (3.21) imply that

$$\lim_{N=2^n, n\to\infty} \frac{|F_N|}{N} = 1 - I(W) \overset{(3.3)}{=} 1 - R_s(D). \tag{3.12}$$

The above equation implies that for any $\epsilon > 0$ and for N sufficiently large there exists a set F_N such that

$$\frac{|F_N|}{N} \geq 1 - R_s(D) - \epsilon.$$

In other words

$$R_N = 1 - \frac{|F_N|}{N} \leq R_s(D) + \epsilon.$$

Finally, from Lemma 3.6 we know that

$$D_N(F_N) \leq D + 2|F_N|\mathtt{d}_{\max}\delta_N \leq D + O(2^{-(N^\beta)}) \tag{3.13}$$

for any $0 < \beta < \frac{1}{2}$.

Recall that $D_N(F_N)$ is the average of the distortion over all choices of u_{F_N}. Since the average distortion fulfills (3.13) it follows that there must be at least one choice of u_{F_N} for which

$$D_N(F_N, u_{F_N}) \leq D + O(2^{-(N^\beta)})$$

for any $0 < \beta < \frac{1}{2}$.

The complexity of the encoding and decoding algorithms is $O(N \log(N))$ as shown in Chapter 2. □

3.4 Value of Frozen Bits Does Not Matter

In the previous sections we have considered $D_N(F)$, the distortion if we average over the code ensemble $\mathtt{C}_N(F)$. For problems with symmetric test channels and distortion functions, we will now show a stronger result, namely that *all* choices for u_F lead to the same distortion, i.e., $D_N(F, u_F)$ is independent of u_F. This implies that the components belonging to the frozen set F can be set to any value. A convenient choice is to set them to 0. In the following let F be a fixed set. The results here are independent of the set F.

Since W is symmetric, Definition 1.1 implies that there exists a permutation $\pi_1 : \mathcal{Y} \to \mathcal{Y}$ such that $\pi_1 = \pi_1^{-1}$ and $W(y \,|\, 0) = W(\pi_1(y) \,|\, 1)$. Let $\pi_0 : \mathcal{Y} \to \mathcal{Y}$ denote the identity permutation.

Definition 3.9 (Symmetric Distortion Function)**.** *We say that the distortion function is symmetric with respect to the permutation $\pi : \mathcal{Y} \to \mathcal{Y}$ if $\mathtt{d}(y, 0) = \mathtt{d}(\pi(y), 1)$.*

The permutations $\{\pi_0, \pi_1\}$ form an Abelian group under function composition. In the following we treat the set $\mathcal{X}$ as a group with the usual XOR " $\oplus$ " as the group operation. Then for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the operation $x \cdot y$ defined as $x \cdot y = \pi_x(y)$ forms a group action. Let the distortion function be symmetric with respect to the permutation π_1, i.e., $\mathtt{d}(y, x) = \mathtt{d}(z \cdot y, x \oplus z)$ for $z \in \mathcal{X}$.

The group action can be easily extended to vectors as follows. The group $\mathcal{X}^N$ consists of $\{0,1\}^N$ and the group operation $\bar{z}_1 \oplus \bar{z}_2$ is defined as $(z_{1,0} \oplus z_{2,0}, \ldots, z_{1,N-1} \oplus z_{2,N-1})$. For $\bar{z} \in \mathcal{X}^N$ and $\bar{y} \in \mathcal{Y}^N$, the group action $\bar{z} \cdot \bar{y}$ is defined as $\bar{z} \cdot \bar{y} = (z_0 \cdot y_0, \ldots, z_{N-1} \cdot y_{N-1})$. The orbit of an element $\bar{y} \in \mathcal{Y}^N$ is the set $\{\bar{z} \cdot \bar{y} : \bar{z} \in \mathcal{X}^N\}$. It is well known that the set of orbits form a partition. Let $\mathcal{Y}_1^N, \ldots, \mathcal{Y}_K^N$ denote the orbits of $\mathcal{Y}^N$.

Example 3.10 (Orbits of the Binary Symmetric Source)**.** *If W is a BSC, then the set $\mathcal{Y}$ is equal to $\{0,1\}$ and the permutation π_z is defined as $\pi_z(y) = z \oplus y$. The set of orbits consists of only the set $\mathcal{Y}$. For the N dimensional case, the set $\mathcal{Y}^N$ again has only one orbit $\mathcal{Y}^N$.*

Example 3.11 (Orbits of the Binary Erasure Source)**.** *If W is a BEC, then then set $\mathcal{Y}$ consists of three elements $\{0, 1, *\}$. The permutation π_0 is the identity permutation. The permutation π_1 is defined as $\pi_1(0) = 1, \pi_1(1) = 0, \pi_1(*) = *$. In this case, there are 2 orbits, namely $\{0,1\}$ and $\{*\}$. For the N dimensional case, the set $\mathcal{Y}^N$ has 2^N orbits. For every subset $A \subseteq \{0, \ldots, N-1\}$, the orbit $\mathcal{Y}_A^N$ is defined as*

$$\{\bar{y} : y_i = * \; \forall i \in A, y_i \in \{0,1\} \; \forall i \in A^c\}.$$

Let W^N denote the channel $W^N(\bar{y} \,|\, \bar{x}) = \prod_{i=0}^{N-1} W(y_i \,|\, x_i)$. The symmetry of the channel W implies the following in the language of group action.

$$W^N(\bar{y} \,|\, \bar{x}) = W^N(\bar{z} \cdot \bar{y} \,|\, \bar{x} \oplus \bar{z}), \tag{3.14}$$

$$W_N(\bar{y} \,|\, \bar{u}) = W_N(\bar{z} \cdot \bar{y} \,|\, \bar{u} \oplus (\bar{z}(G_2^{\otimes n})^{-1})). \tag{3.15}$$

Lemma 3.12 (Gauge Transformation)**.** *Let $\bar{y}, \bar{y}' \in \mathcal{Y}_k^N$ and let $\bar{y}' = \bar{z} \cdot \bar{y}$ for some $\bar{z} \in \mathcal{X}^N$. Let $\bar{u}, \bar{u}' \in \mathcal{X}^N$ be such that $u'^{i-1}_0 = u_0^{i-1} \oplus (\bar{z}(G_2^{\otimes n})^{-1})_0^{i-1}$. Then*

$$L_N^{(i)}(\bar{y}, u_0^{i-1}) = \begin{cases} L_N^{(i)}(\bar{y}', u'^{i-1}_0), & \text{if } (\bar{z}(G_2^{\otimes n})^{-1})_i = 0, \\ 1/L_N^{(i)}(\bar{y}', u'^{i-1}_0), & \text{if } (\bar{z}(G_2^{\otimes n})^{-1})_i = 1. \end{cases}$$

Proof.

$$L_N^{(i)}(\bar{y}, u_0^{i-1}) = \frac{W_N^{(i)}(\bar{y}, u_0^{i-1} \,|\, 0)}{W_N^{(i)}(\bar{y}, u_0^{i-1} \,|\, 1)}$$

$$
\begin{aligned}
&= \frac{\sum_{u_{i+1}^{N-1}} W_N(\bar{y} \,|\, u_0^{i-1}, 0, u_{i+1}^{N-1})}{\sum_{u_{i+1}^{N-1}} W_N(\bar{y} \,|\, u_0^{i-1}, 1, u_{i+1}^{N-1})} \\
&\overset{(3.15)}{=} \frac{\sum_{u_{i+1}^{N-1}} W_N(\bar{y}' \,|\, (u_0^{i-1}, 0, u_{i+1}^{N-1}) \oplus \bar{z}(G_2^{\otimes n})^{-1})}{\sum_{u_{i+1}^{N-1}} W_N(\bar{y}' \,|\, (u_0^{i-1}, 1, u_{i+1}^{N-1}) \oplus \bar{z}(G_2^{\otimes n})^{-1})} \\
&= \frac{\sum_{u_{i+1}^{N-1}} W_N(\bar{y}' \,|\, (u'^{i-1}_0, 0 \oplus (\bar{z}(G_2^{\otimes n})^{-1})_i, u_{i+1}^{N-1})}{\sum_{u_{i+1}^{N-1}} W_N(\bar{y}' \,|\, (u'^{i-1}_0, 1 \oplus (\bar{z}(G_2^{\otimes n})^{-1})_i, u_{i+1}^{N-1})} \\
&= \frac{W_N^{(i)}(\bar{y}', u'^{i-1}_0 \,|\, 0 \oplus (\bar{z}(G_2^{\otimes n})^{-1})_i)}{W_N^{(i)}(\bar{y}', u'^{i-1}_0 \,|\, 1 \oplus (\bar{z}(G_2^{\otimes n})^{-1})_i)}.
\end{aligned}
$$

The claim follows by considering the two possible values of $(\bar{z}(G_2^{\otimes n})^{-1})_i$. □

Recall that the decision process involves randomized rounding on the basis of $L_N^{(i)}$. Consider at first two tuples $(\bar{y}, u_0^{i-1})$ and $(\bar{y}', u'^{i-1}_0)$ so that their associated $L_N^{(i)}$ values are equal; we have seen in the previous lemma that many such tuples exist. In this case, if both tuples have access to the same source of randomness, we can couple the two instances so that they make the same decision on U_i. An equivalent statement is true in the case when the two tuples have the same reliability $|\log(L_N^{(i)}(\bar{y}, u_0^{i-1}))|$ but different signs. In this case there is a simple coupling that ensures that if for the first tuple the decision is lets say $U_i = 0$ then for the second tuple it is $U_i = 1$ and vice versa. Hence, if in the sequel we compare two instances of "compatible" tuples which have access to the same source of randomness, then we assume exactly this coupling.

Lemma 3.13 (Symmetry and Distortion)**.** *Let $\bar{y}, \bar{y}' \in \mathcal{Y}_k^N$ and $\bar{y}' = \bar{z} \cdot \bar{y}$ for some $\bar{z} \in \mathcal{X}^N$. Let $F \subseteq \{0, \ldots, N-1\}$, and $u_F, u'_F \in \mathcal{X}^{|F|}$ be such that $u_F = u'_F \oplus (\bar{z}(G_2^{\otimes n})^{-1})_F$, then under the coupling through a common source of randomness*

$$\hat{U}(\bar{y}, u_F) = \hat{U}(\bar{y}, u'_F) \oplus (\bar{z}(G_2^{\otimes n})^{-1}).$$

Proof. Let $\hat{u}_0^{N-1}, \hat{u}_0'^{N-1}$ be the outputs of $\hat{U}(\bar{y}, u_F)$ and $\hat{U}(\bar{y}', u'_F)$. We use induction. Fix $0 \le i \le N-1$. We assume that for $j < i$, $\hat{u}_j = \hat{u}'_j \oplus (\bar{z}(G_2^{\otimes n})^{-1})_j$. This is in particular correct if $i = 0$, which serves as our anchor.

By Lemma 3.12 we conclude that under our coupling the respective decisions are related as $\hat{u}_i = \hat{u}'_i \oplus (\bar{z}(G_2^{\otimes n})^{-1})_i$ if $i \in F^c$. On the other hand, if $i \in F$, then the claim is true by assumption. □

The symmetry property of the distortion function implies

$$\mathtt{d}(\bar{y}, \bar{x}) = \mathtt{d}(\bar{z} \cdot \bar{y}, \bar{x} \oplus \bar{z}). \tag{3.16}$$

Moreover, from (3.14) we have

$$P_{\bar{Y}}(\bar{y}) = \sum_{\bar{x}} \frac{1}{2^N} W^N(\bar{y} \,|\, \bar{x}) = \sum_{\bar{x}} \frac{1}{2^N} W^N(\bar{z} \cdot \bar{y} \,|\, \bar{x} \oplus \bar{z}) = P_{\bar{Y}}(\bar{z} \cdot \bar{y}),$$

which implies that

$$P_{\bar{Y}}(\bar{y}) = P_{\bar{Y}}(\bar{y}') \ \ \forall \bar{y}, \bar{y}' \in \mathcal{Y}_k^N. \tag{3.17}$$

it follows that all $\bar{y}$ belonging to an orbit have the same probability, i.e.,

To every set $\mathcal{Y}_k^N$, associate a vector $\bar{y}_k \in \mathcal{Y}_k^N$ as its leader. Let $B(\bar{y}) = \{\bar{z} : \bar{y} = \bar{z} \cdot \bar{y}_k\}$. The set $B(\bar{y})$ may contain more than one element. One can check that $\{B(\bar{y}) : \forall \bar{y} \in \mathcal{Y}_k^N\}$ form a partition of $\mathcal{X}^N$. Moreover,

$$|B(\bar{y})| = |B(\bar{y}')| \ \forall \ \bar{y}, \bar{y}' \in \mathcal{Y}_k^N. \tag{3.18}$$

Let $\bar{y}_k(\bar{z}) = \bar{z} \cdot \bar{y}_k$. Combining (3.17) with (3.18) implies

$$\sum_{\bar{y} \in \mathcal{Y}_k^N} P(\bar{y}) \mathtt{d}(\bar{y}, \hat{U}(\bar{y}, u_F)) = \sum_{\bar{z} \in \mathcal{X}^N} \frac{P(\mathcal{Y}_k^N)}{2^N} \mathtt{d}(\bar{y}_k(\bar{z}), \hat{U}(\bar{y}_k(\bar{z}), u_F)). \tag{3.19}$$

Let $\bar{v} \in \mathcal{X}^{|F|}$ and let $A(\bar{v}) \subset \mathcal{X}^N$ denote the coset

$$A(\bar{v}) = \{\bar{z} : (\bar{z}(G_2^{\otimes n})^{-1})_F = \bar{v}\}.$$

The set $\mathcal{X}^N$ can be partitioned as

$$\mathcal{X}^N = \cup_{\bar{v} \in \mathcal{X}^{|F|}} A(\bar{v}).$$

Now we are ready to prove the main result of this section.

Lemma 3.14 (Independence of Average Distortion w.r.t. u_F)**.** *The average distortion $D_N(F, u_F)$ is independent of the choice of $u_F \in \{0,1\}^{|F|}$.*

Proof. Let $u_F, u_F' \in \{0,1\}^{|F|}$ be two fixed vectors. We will now show that $D_N(F, u_F) = D_N(F, u_F')$. Let $\bar{z}, \bar{z}' \in \mathcal{X}^N$ be such that $\bar{z} \in A(\bar{v})$ and $\bar{z}' \in A(\bar{v} \oplus u_F \oplus u_F')$. This implies that for any orbit $\mathcal{Y}_k^N$, $\bar{y}_k(\bar{z}) = (\bar{z} \oplus \bar{z}') \cdot \bar{y}_k(\bar{z}')$ and $u_F' = u_F \oplus ((\bar{z} \oplus \bar{z}')(G_2^{\otimes n})^{-1})_F$. Applying Lemma 3.13, we get

$$\hat{U}(\bar{y}_k(\bar{z}'), u_F') = \hat{U}(\bar{y}_k(\bar{z}), u_F) \oplus ((\bar{z} \oplus \bar{z}')(G_2^{\otimes n})^{-1}).$$

Therefore,

$$\begin{aligned} \mathtt{d}(\bar{y}_k(\bar{z}), \hat{U}(\bar{y}_k(\bar{z}), u_F) G_2^{\otimes n}) &\overset{(3.16)}{=} \mathtt{d}((\bar{z} \oplus \bar{z}') \cdot \bar{y}_k(\bar{z}), \hat{U}(\bar{y}(\bar{z}), u_F) G_2^{\otimes n} \oplus (\bar{z} \oplus \bar{z}')) \\ &= \mathtt{d}(\bar{y}_k(\bar{z}'), \hat{U}(\bar{y}_k(\bar{z}'), u_F') G_2^{\otimes n}), \end{aligned}$$

which further implies

$$\sum_{\bar{z} \in A(\bar{v})} \mathtt{d}(\bar{y}_k(\bar{z}), \hat{U}(\bar{y}_k(\bar{z}), u_F) G_2^{\otimes n}) = \sum_{\bar{z}' \in A(\bar{v} \oplus u_F \oplus u_F')} \mathtt{d}(\bar{y}_k(\bar{z}'), \hat{U}(\bar{y}_k(\bar{z}'), u_F') G_2^{\otimes n}). \tag{3.20}$$

Hence, the average distortions over the set $\mathcal{Y}_k^N$ satisfy

$$\begin{aligned}
&\sum_{\bar{y}\in\mathcal{Y}_k^N} P(\bar{y}) \quad \mathtt{d}(\bar{y}, \hat{U}(\bar{y}, u_F)G_2^{\otimes n}) \\
&\overset{(3.19)}{=} \sum_{\bar{v}\in\{0,1\}^{|F|}} \frac{P(\mathcal{Y}_k^N)}{2^N} \sum_{\bar{z}\in A(\bar{v})} \mathtt{d}(\bar{y}_k(\bar{z}), \hat{U}(\bar{y}_k(\bar{z}), u_F)G_2^{\otimes n}) \\
&\overset{(3.20)}{=} \sum_{\bar{v}\in\{0,1\}^{|F|}} \frac{P(\mathcal{Y}_k^N)}{2^N} \sum_{\bar{z}'\in A(\bar{v}\oplus u_F\oplus u'_F)} \mathtt{d}(\bar{y}_k(\bar{z}'), \hat{U}(\bar{y}_k(\bar{z}'), u'_F)G_2^{\otimes n}) \\
&= \sum_{\bar{v}\in\{0,1\}^{|F|}} \frac{P(\mathcal{Y}_k^N)}{2^N} \sum_{\bar{z}'\in A(\bar{v})} \mathtt{d}(\bar{y}_k(\bar{z}'), \hat{U}(\bar{y}_k(\bar{z}'), u'_F)G_2^{\otimes n}) \\
&= \sum_{\bar{y}\in\mathcal{Y}_k^N} P(\bar{y})\mathtt{d}(\bar{y}, \hat{U}(\bar{y}, u'_F)G_2^{\otimes n}).
\end{aligned}$$

As mentioned before, the encoding function $\hat{U}(\cdot,\cdot)$ is not deterministic and the above equality is valid under the assumption of coupling with a common source of randomness. Averaging over this common randomness and summing over all the orbits, we get $D_N(F, u_F) = D_N(F, u'_F)$.

□

3.5 Simulation Results and Discussion

Let us consider how polar codes behave in practice. Consider again the set-up of Example 3.3. I.e., we consider the BSS and the Hamming distortion function $\mathtt{d}(0,1) = \mathtt{d}(1,0) = 1$ and $\mathtt{d}(0,0) = \mathtt{d}(1,1) = 0$, the rate distortion function is given by $R(D) = 1 - h_2(D)$. The test channel is the BSC(D). As mentioned in Example 3.3, polar codes achieve the rate-distortion trade-off for this source. Moreover, since both the channel and the distortion function satisfy the symmetry criterion, the frozen indices can be fixed to any value.

Recall that the length N of the code is always a power of 2, i.e., $N = 2^n$. Figure 3.2 shows the performance of the SC encoding algorithm combined with randomized rounding. As asserted by Theorem 3.4, the points approach the rate-distortion bound as the blocklength increases.

Let us now discuss the similarities and differences between the polar codes designed for channel coding and source coding. Consider channel coding for a B-DMC W and source coding with the same channel W as a test channel. In both cases, in order to fully specify the code we need to specify the set of frozen indices. Let F_c denote the frozen set for the channel code and F_s denote the frozen set for the source code.

The code construction for both problems requires knowledge of $\{Z(W_N^{(i)}) : \forall i \in \{0,\ldots,N-1\}\}$. Therefore, the complexity issues discussed for the channel code construction in Section 2.5.2 apply for source coding as well. The set F_c is chosen according to $\{i : Z(W_N^{(i)}) \geq 2^{-(N)^\beta}\}$ and the set F_s is chosen according

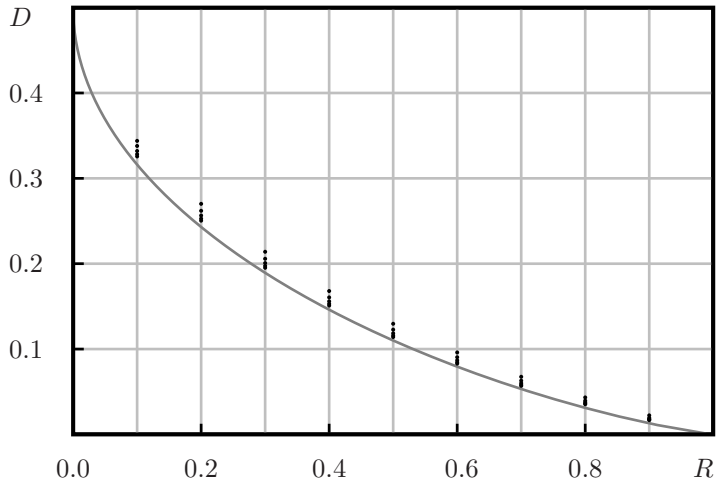

Figure 3.2: The rate-distortion performance for the SC encoding algorithm with randomized rounding for $n = 9, 11, 13, 15$, and 17. The rates that we consider are $0.1, 0.2, \ldots, 0.9$. As the blocklength increases the points move closer to the rate-distortion bound.

to $\{i : Z(W_N^{(i)}) \geq 1 - 2^{-(N)^\beta}\}$. Therefore, we always have $F_s \subseteq F_c$. The polarization of the channel implies that for channel coding we always have $R < I(W)$ and to source coding we always have $R > I(W)$. In other words, by adding a few generator vectors we have converted a good channel code into a good source code.

In practice we fix a threshold δ close to 0 and chose F_c as $\{i : Z(W_N^{(i)}) \geq \delta\}$ and similarly F_s is defined as $\{i : Z(W_N^{(i)}) \geq 1 - \delta\}$. By varying the value of δ we get the trade-off between rate and probability of error (distortion) for channel (source) coding.

The role of the encoding and the decoding algorithms for the two problems are reversed. For channel coding the encoding operation is a simple matrix multiplication which is in fact the decoding operation for source coding. On the other hand, the decoding operation for channel coding involves computing the likelihoods of an SC decoder; this in turn is the same procedure that is required in the encoding operation of source coding. When making the decision on bit U_i using SC encoding, it is natural to choose that value for U_i which maximizes the posterior. This is what we do for channel decoding and such a scheme works well in practice for source encoding as well. For the analysis however it is more convenient to use randomized rounding in source coding.

Recall from our discussion in Section 1.2.2 that in source coding there are typically many codewords that, if chosen, result in similar distortion. As a result, the standard message-passing algorithms fail to perform well. We also mentioned that one way to overcome this problem is to combine message-passing with decimation. Note that in the SC encoding algorithm we decide the bits U_i in the order 0 to $N - 1$. Therefore the decimation step is already incorporated in the algorithm. This is in fact a crucial reason why SC algorithm

succeeds for source coding. In fact, the SC encoder can be interpreted as a particular instance of BID where the order of the decimation is fixed in advance $(0, \ldots, N-1)$. The decimation based on randomized rounding can be interpreted as choosing one of the candidate codewords at random.

On the other hand, for channel coding there is typically one codeword (namely the transmitted one) which has a posterior that is significantly larger than all other codewords. This makes it possible for a greedy message-passing algorithm to successfully move towards this codeword in small steps, using at any given moment "local" information provided by the decoder. Therefore in the case of channel coding, instead of SC decoder, we can run the standard BP decoder and decide the bits at the end. As we will see in Chapter 6, in fact BP performs better than SC decoder. However, the analysis of BP decoder is much more difficult.

3.A Appendix

Consider the random process $\{W_n; n \geq 0\}$ as defined in Section 2.3. As mentioned there, the relevance of this process is that

$$\Pr(Z_n \in (a,b)) = \frac{\left|\left\{i \in \{0,\ldots,2^n-1\} : Z(W_N^{(i)}) \in (a,b)\right\}\right|}{2^n}. \tag{3.21}$$

For lossy source coding, the quantity of interest is the rate at which the random variable Z_n approaches 1 (as compared to 0 for channel coding). Let us now show the result mirroring Theorem 2.11 for this case.

Theorem 3.15 (Rate of Z_n Approaching 1)**.** *Given a B-DMC W, and a $0 \leq \beta < \frac{1}{2}$,*

$$\lim_{n\to\infty} \Pr(Z_n \geq 1 - 2^{-2^{n\beta}}) = 1 - I(W).$$

Proof. The random process $\{Z_n\}$ satisfies

$$\begin{array}{ll} Z_{n+1} \overset{\text{Lem. 3.16}}{\geq} \sqrt{2Z_n^2 - Z_n^4} & \text{w.p. } \frac{1}{2}, \\ Z_{n+1} \overset{\text{Lem. 2.16}}{=} Z_n^2 & \text{w.p. } \frac{1}{2}. \end{array}$$

The above statements can be rewritten as

$$\begin{array}{ll} 1 - Z_{n+1}^2 \leq (1 - Z_n^2)^2 & \text{w.p. } \frac{1}{2}, \\ 1 - Z_{n+1}^2 = 1 - Z_n^4 \leq 2(1 - Z_n^2) & \text{w.p. } \frac{1}{2}. \end{array}$$

Let X_n denote $X_n = 1 - Z_n^2$. Then $\{X_n : n \geq 0\}$ satisfies

$$X_{n+1} \leq X_n^2 \text{ w.p. } \frac{1}{2},$$

$$X_{n+1} \leq 2X_n \text{ w.p. } \frac{1}{2}.$$

In Chapter 2 we have seen that the process $\{Z_n\}$ converges almost surely to a random variable Z_∞ with $\Pr(Z_\infty = 0) = I(W)$ and $\Pr(Z_\infty = 1) = 1 - I(W)$. This implies that the process $\{X_n\}$ converges almost surely to a random variable X_∞ such that $\Pr(X_\infty = 0) = 1 - I(W)$ and $\Pr(X_\infty = 1) = I(W)$. Therefore, the process $\{X_n\}$ satisfies the conditions of Theorem 2.10 with $q = 2$ and $P_\infty = 1 - I(W)$. This implies that for any $\beta < \frac{1}{2}$, the process $\{X_n\}$ satisfies,

$$\lim_{n\to\infty} \Pr(X_n \leq 2^{-2^{n\beta}}) = 1 - I(W).$$

Using the relation $X_n = 1 - Z_n^2 \geq 1 - Z_n$, we get

$$\lim_{n\to\infty} \Pr(1 - Z_n \leq 2^{-2^{n\beta}}) = 1 - I(W).$$

□

Lemma 3.16 (Lower Bound on $Z(W_1 \circledast W_2)$)**.** *Let W_1 and W_2 be two B-DMCs. Then*

$$Z(W_1 \circledast W_2) \geq \sqrt{Z(W_1)^2 + Z(W_2)^2 - Z(W_1)^2 Z(W_2)^2}.$$

Proof. Let $Z = Z(W_1 \circledast W_2)$ and $Z_i = Z(W_i)$. Z. As shown in Lemma 2.15, Z can be expressed as

$$Z = \frac{Z_1 Z_2}{2} \mathbb{E}_{1,2} \left[\sqrt{\left(A_1(Y_1)\right)^2 + \left(A_2(Y_2)\right)^2 - 4} \right].$$

where $\mathbb{E}_i$ denotes the expectation with respect to the probability distribution P_i over $\mathcal{Y}_i$

$$P_i(y_i) = \frac{\sqrt{W_i(y_i \,|\, 0) W_i(y_i \,|\, 1)}}{Z_i},$$

and

$$A_i(y) \triangleq \sqrt{\frac{W_i(y \,|\, 0)}{W_i(y \,|\, 1)}} + \sqrt{\frac{W_i(y \,|\, 1)}{W_i(y \,|\, 0)}}.$$

The arithmetic-mean geometric-mean inequality implies that $A_i(y) \geq 2$. Therefore, for any $y_i \in \mathcal{Y}_i$, $A_i(y_i)^2 - 4 \geq 0$. Note that the function $f(x) = \sqrt{x^2 + a}$ is convex for $a \geq 0$. Applying Jensen's inequality first with respect to the expectation $\mathbb{E}_1$ and then with respect to $\mathbb{E}_2$, we get

$$\begin{aligned} Z &\geq \frac{Z_1 Z_2}{2} \mathbb{E}_2 \left[\sqrt{\left(\mathbb{E}_1 \left[A_1(Y_1)\right]\right)^2 + \left(A_2(Y_2)\right)^2 - 4} \right] \\ &\geq \frac{Z_1 Z_2}{2} \sqrt{\left(\mathbb{E}_1 \left[A_1(Y_1)\right]\right)^2 + \left(\mathbb{E}_2 \left[A_2(Y_2)\right]\right)^2 - 4}. \end{aligned}$$

The claim follows by substituting $\mathbb{E}_i[A_i(Y_i)] = \frac{2}{Z_i}$. □

4 Multi-Terminal Scenarios

In Chapter 2 we have seen that polar codes provide a low-complexity scheme to achieve the symmetric capacity of B-DMCs. In Chapter 3 we have seen that polar codes achieve the symmetric rate-distortion bound for lossy source compression. The natural question to ask next is whether these codes are also suitable for problems that involve both quantization as well as error correction.

Perhaps the two most prominent examples in this area are the source coding problem with side information (Wyner-Ziv problem [55]) and the channel coding problem with side information (Gelfand-Pinsker problem [56]). As discussed in [57], these problems can be tackled using nested linear codes. Polar codes are equipped with such a nested structure and are, hence, natural candidates for these problems. We will show that, by taking advantage of this nested structure, one can construct polar codes that are optimal in both settings (for the binary versions of these problems). Hence, polar codes provide the first provably optimal low-complexity solution.

Let us quickly review the state-of-the-art schemes for these problems. In [38] the authors constructed MN codes [21] which have the required nested structure for both the Wyner-Ziv as well as the Gelfand-Pinsker problem. They show that these codes achieve the optimum performance under MAP decoding. How these codes perform under low-complexity message-passing algorithms is still an open problem. Trellis and turbo-based codes were considered in [58, 59, 60, 61] for the Wyner-Ziv problem. It was empirically shown that they achieve good performance with low-complexity message-passing algorithms. A similar combination was considered in [62, 63, 64] for the Gelfand-Pinsker problem. Again, empirical results close to the optimum performance were obtained.

We also consider applications to multi-terminal setups including the Slepian-Wolf [65] problem, the one-helper problem [66], the multiple access chan-

nel (MAC), and the degraded broadcast channel (DBC). For all the above problems we show that polar codes achieve optimal performance using low-complexity encoding and decoding algorithms.

For the sake of clarity we restrict ourselves to the binary versions of the various problems mentioned above. More precisely, the channel we consider is the BSC(p), and the source is the BSS. The reconstruction alphabet is also binary. Let the distortion function be the Hamming distortion function, i.e., $\mathtt{d}(0,1) = \mathtt{d}(1,0) = 1$ and $\mathtt{d}(0,0) = \mathtt{d}(1,1) = 0$. The results of this chapter can be extended to general B-DMCs.

We start by proving the optimality of polar codes for the Wyner-Ziv problem (Section 4.1). We then show the optimality for the Gelfand-Pinsker problem (Section 4.2), the Slepian-Wolf problem (Section 4.3) and the one-helper problem (Section 4.4). In Section 4.5 we discuss applications to asymmetric channels, multiple access channels and degraded broadcast channels.

4.1 Wyner-Ziv Problem

Let Y be a BSS and let the decoder have access to a random variable Y'. This random variable is usually called the *side information*. We assume that Y' is correlated to Y in the following fashion: $Y' = Y + Z$, where Z is a Ber(p) random variable. The task of the encoder is to compress the source Y, call the result X, such that a decoder with access to (Y', X) can reconstruct the source to within a distortion D.

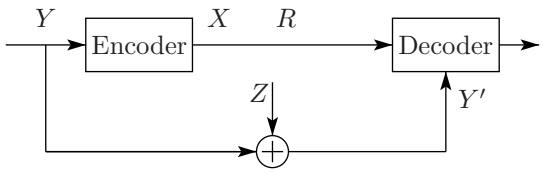

Figure 4.1: The Wyner-Ziv problem. The task of the decoder is to reconstruct the source Y to within a distortion D given (Y', X).

Wyner and Ziv [55] have shown that the rate-distortion curve for this problem is given by

$$\text{l.c.e.}\Big\{(R_{\text{WZ}}(D), D), (0, p)\Big\},$$

where $R_{\text{WZ}}(D) = h_2(D * p) - h_2(D)$, l.c.e. denotes the *lower convex envelope*, and $D * p = D(1-p) + p(1-D)$. Here we focus on achieving the rates of the form $R_{\text{WZ}}(D)$. The remaining rates can be achieved by appropriate time-sharing with the pair $(0, p)$.

The proof that the Wyner-Ziv rate distortion can be achieved by polar codes is based on the following nested code construction. Let $\mathtt{C}_s$ denote the

polar code $\mathtt{C}_N(F_s, \bar{0})$. Let $F_c \supseteq F_s$ denote another frozen set and let $\bar{v} \in \mathcal{X}^{|F_c \backslash F_s|}$. Let $\mathtt{C}_c(\bar{v})$ denote the polar code $\mathtt{C}_N(F_c, u_{F_c}(\bar{v}))$ where $u_{F_c}(\bar{v})$ is defined as $u_{F_s} = \bar{0}$ and $u_{F_c \backslash F_s} = \bar{v}$. This implies that the code $\mathtt{C}_s$ can be partitioned as $\mathtt{C}_s = \cup_{\bar{v}} \mathtt{C}_c(\bar{v})$. The code $\mathtt{C}_s$ (the set F_s) is designed to be a good source code for distortion D. Further, for each $\bar{v}$ the code $\mathtt{C}_c(\bar{v})$ (the set F_c) is designed to be a good channel code for the BSC$(D * p)$.

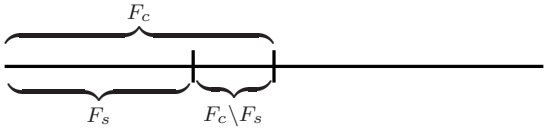

Figure 4.2: A pictorial representation of the subset structure of the frozen sets F_s and F_c. Let the line represent the set of indices. The set F_s is fixed to 0. The bits belonging to the set $F_c \backslash F_s$ are transmitted to the decoder.

The encoder maps the source vector $\bar{Y}$ to a vector $\hat{U}(\bar{Y}, u_{F_s} = \bar{0})$. By the definition of F_s, the reconstruction word $\bar{X}$, given by $\bar{X} = \hat{U}(\bar{Y}, u_{F_s} = \bar{0}) G_2^{\otimes n}$, represents $\bar{Y}$ with average distortion roughly D.

Rather than sending the whole vector $\hat{U}_{F_s^c}$ to the decoder, the encoder takes advantage of the fact that the decoder knows $\bar{Y}'$ and declares only the sub-vector $\bar{V} = \hat{U}_{F_c \backslash F_s}$ to the decoder. Therefore, the rate of such a scheme is $\frac{|F_c \backslash F_s|}{N}$. The decoder hence knows $\hat{U}_{F_c}$ and wants to find $\hat{U}$ (or equivalently $\bar{X}$) given $\bar{Y}'$. The problem is therefore equivalent to decoding the codeword $\bar{X}$ belonging to $\mathtt{C}_N(F_c, \hat{U}_{F_c})$ given the observation $\bar{Y}'$. Assume at first that $\bar{Y}'$ is the result of transmitting $\bar{X}$ over the BSC$(D * p)$. In this case, it is clear that the decoder can recover $\bar{X}$ with high probability since F_c is chosen to be the frozen set of a good channel code for the BSC$(D * p)$.

The key observation is that $\bar{Y}'$ is statistically close to the output of $\bar{X}$ sent through the BSC$(D * p)$. This claim is made precise in Lemma 4.6, where it is shown that the distribution of the quantization error $\bar{Y} \oplus \bar{X}$ is close to a Ber(D) vector with i.i.d. components.

Putting the above ideas together, we get the following result.

Theorem 4.1 (Optimality for the Wyner-Ziv Problem)**.** *Let Y be a BSS and Y' be a Bernoulli random variable correlated to Y as $Y' = Y \oplus Z$, where $Z \sim Ber(p)$. Fix the* design *distortion D, $0 < D < \frac{1}{2}$. For any rate $R > h_2(D * p) - h_2(D)$ and any $0 < \beta < \frac{1}{2}$, there exists a sequence of nested polar codes of length N with rates $R_N < R$ so that under SC encoding using randomized rounding at the encoder and SC decoding at the decoder, they achieve expected distortion D_N satisfying*

$$D_N \leq D + O(2^{-(N^\beta)}).$$

Further the block error probability satisfies

$$P_N \leq O(2^{-(N^\beta)}).$$

The encoding as well as decoding complexity of this scheme is $O(N \log(N))$.

Proof. Let $\epsilon > 0$ and $0 < \beta < \frac{1}{2}$ be some constants. Let $Z_N^{(i)}(q)$ denote $Z(W_N^{(i)})$, with W set to a BSC(q). Let $\delta_N = \frac{1}{N}2^{-(N^\beta)}$. Let F_s and F_c denote the sets

$$F_s = \{i : Z_N^{(i)}(D) \geq 1 - \delta_N^2\},$$
$$F_c = \{i : Z_N^{(i)}(D * p) \geq \delta_N\}.$$

Theorem 3.15 implies that for N sufficiently large

$$\frac{|F_s|}{N} \geq h_2(D) - \frac{\epsilon}{2}.$$

Similarly, Theorem 2.11 implies that for N sufficiently large

$$\frac{|F_c|}{N} \leq h_2(D * p) + \frac{\epsilon}{2}.$$

It is clear that the BSC($D*p$) is degraded with respect to the BSC(D). Indeed, concatenating a BSC(D) with a BSC(p) results in a BSC($D * p$). For small δ_N, we have $\delta_N \leq 1 - \delta_N^2$. Therefore if $i \in F_s$, then

$$\delta_N \leq 1 - \delta_N^2 \leq Z_N^{(i)}(D) \overset{\text{Lem. } 4.7}{\leq} Z_N^{(i)}(D * p).$$

This implies that if $i \in F_s$ then $i \in F_c$, i.e., $F_s \subseteq F_c$. The bits u_{F_s} are fixed to 0 and it is known both to the encoder and the decoder.

Consider now the encoding process. We encode the source vector $\bar{y}$ using the operation $\hat{U}(\bar{y}, u_{F_s} = \bar{0})$. We have seen in Chapter 3 that the average distortion D_N of such a scheme is bounded by

$$D_N \leq D + 2|F_s|\delta_N \leq D + O(2^{-(N^\beta)}).$$

The encoder transmits the vector $\hat{u}_{F_c \setminus F_s}$ to the decoder. The required rate is

$$R_N = \frac{|F_c| - |F_s|}{N} \leq h_2(D * p) - h_2(p) + \epsilon.$$

It remains to show that at the decoder the block error probability incurred in decoding $\bar{X}$ given $\bar{Y}'$ is $O(2^{-(N^\beta)})$.

Let $\bar{E}$ denote the quantization error, $\bar{E} = \bar{Y} \oplus \bar{X}$. The observation $\bar{Y}'$ available at the decoder can be expressed as

$$\bar{Y}' = \bar{X} \oplus \bar{E} \oplus \bar{Z}.$$

Consider the code $\mathtt{C}_c(\bar{v})$ for a given $\bar{v}$ and transmission over the BSC($D*p$). The channel symmetry implies that for every codeword belonging to the code $\mathtt{C}_c(\bar{v})$ the set of noise vectors that result in an error under SC decoding is the same. The channel symmetry also implies that this set is independent of $\bar{v}$.

Let us denote this set by $\mathcal{E} \in \mathcal{X}^N$. The block error probability of our scheme can then be expressed as

$$P_N = \mathbb{E}[\mathbb{1}_{\{\bar{E}\oplus\bar{Z}\in\mathcal{E}\}}],$$

where the expectation is over the randomness of the Bernoulli noise $\bar{Z}$ and also the quantization error $\bar{E}$.

The exact distribution of the quantization error is not known, but Lemma 4.6 provides a bound on the total variation distance between this distribution and an i.i.d. Ber(D) distribution. Let $\bar{B}$ denote an i.i.d. Ber(D) vector. Let $P_{\bar{E}}$ and $P_{\bar{B}}$ denote the distribution of $\bar{E}$ and $\bar{B}$ respectively. Then from Lemma 4.6, we have

$$\sum_{\bar{e}} |P_{\bar{E}}(\bar{e}) - P_{\bar{B}}(\bar{e})| \leq 2|F_s|\delta_N \leq O(2^{-(N^\beta)}). \tag{4.1}$$

Let $\Pr(\bar{B}, \bar{E})$ denote the so-called *optimal coupling* between $\bar{E}$ and $\bar{B}$. I.e., a joint distribution of $\bar{E}$ and $\bar{B}$ with marginals equal to $P_{\bar{E}}$ and $P_{\bar{B}}$, and satisfying

$$\Pr(\bar{E} \neq \bar{B}) = \sum_{\bar{e}} |P_{\bar{E}}(\bar{e}) - P_{\bar{B}}(\bar{e})|. \tag{4.2}$$

It is known [67] that such a coupling exists. Let $\bar{E}$ and $\bar{B}$ be generated according to $\Pr(\cdot,\cdot)$. Then, the block error probability can be expanded as

$$\begin{aligned} P_N &= \mathbb{E}[\mathbb{1}_{\{\bar{E}\oplus\bar{Z}\in\mathcal{E}\}}\mathbb{1}_{\{\bar{E}=\bar{B}\}}] + \mathbb{E}[\mathbb{1}_{\{\bar{E}\oplus\bar{Z}\in\mathcal{E}\}}\mathbb{1}_{\{\bar{E}\neq\bar{B}\}}] \\ &\leq \mathbb{E}[\mathbb{1}_{\{\bar{B}\oplus\bar{Z}\in\mathcal{E}\}}] + \mathbb{E}[\mathbb{1}_{\{\bar{E}\neq\bar{B}\}}]. \end{aligned}$$

The first term in the sum refers to the block error probability for the BSC($D * p$), which can be bounded as

$$\mathbb{E}[\mathbb{1}_{\{\bar{B}\oplus\bar{Z}\in\mathcal{E}\}}] \leq \sum_{i\in F_c^c} Z_N^{(i)}(D * p) \leq O(2^{-(N^\beta)}). \tag{4.3}$$

Using (4.1), (4.2) and (4.3) we get

$$P_N \leq O(2^{-(N^\beta)}).$$

□

Figure 4.3 shows simulation results for the Wyner-Ziv problem. In these simulations, we assume that the channel corresponding to the side-information is the BSC(0.3). As mentioned before, we consider here the performance only for rate-distortion pairs of the form $(R_{\mathrm{WZ}}(D), D)$.

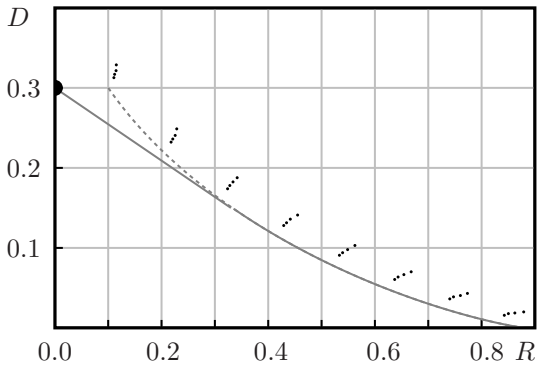

Figure 4.3: The Wyner-Ziv function $R_{WZ}(D)$ (dashed line). The solid line is the lower convex envelope of $R_{WZ}(D)$ and $(0, p = 0.3)$. The performance curves are shown for $n = 9, 11, 13$, and 15.

4.2 Gelfand-Pinsker Problem

The Gelfand-Pinsker problem is a channel coding problem over a channel with state. The state is known a-causally to the encoder but not known to the decoder. The problem was studied and its capacity was determined by Gelfand and Pinsker in [56].

Let us consider the binary version of this problem; it is also referred to as the information embedding problem. Let S denote a Ber($\frac{1}{2}$) random variable. The random variable S is the state of a B-DMC. The output of the B-DMC is given by

$$Y = X \oplus S \oplus Z,$$

where X is the input and Z is a Ber(p) random variable. The state S is known to the encoder a-causally but not known to the decoder. The channel input at the encoder, denoted by X, is constrained to satisfy $\mathbb{E}[X] \leq D$. In other words, the encoder operates under the constraint of using at most a fraction D of ones on average. Without this constraint, the problem is trivial. In this case the encoder can completely cancel the effect of the state by adding $\bar{S}$ to $\bar{X}$ ($\bar{S} \oplus \bar{X}$). The resulting channel between the encoder and the decoder would in this case be the BSC(p). The problem becomes non-trivial however if we impose the input constraint. The input constraint is similar to the power constraint for the continuous input case.

The Gelfand-Pinsker rate region for this channel is determined in [68]. The achievable rate-weight pairs are given by

$$\text{u.c.e.}\Big\{(R_{\text{GP}}(D), D), (0, 0)\Big\},$$

where $R_{\text{GP}}(D) = h_2(D) - h_2(p)$, and u.c.e denotes the upper convex envelope. Here we focus on achieving the rates of the form $R_{\text{GP}}(D)$. The remaining rates can be achieved by appropriate time-sharing with the pair $(0, 0)$.

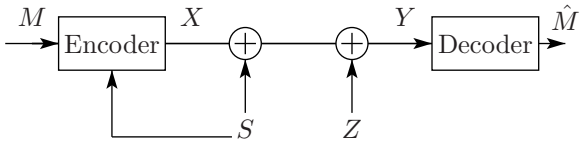

Figure 4.4: The Gelfand-Pinsker problem. The state S is known to the encoder a-causally but not known to the decoder. The transmission at the encoder is constrained to have only a fraction D of ones on average, i.e., $\mathbb{E}[X] \leq D$.

Similar to the Wyner-Ziv problem, we need a nested code for this problem. But the roles of the channel and source codes are reversed. Let $\mathtt{C}_c$ denote the polar code $\mathtt{C}_N(F_c, \bar{0})$. Let $F_s \supseteq F_c$ denote another frozen set and let $\bar{v} \in \mathcal{X}^{|F_s \backslash F_c|}$. Let $\mathtt{C}_s(\bar{v})$ denote the polar code $\mathtt{C}_N(F_s, u_{F_s}(\bar{v}))$ with $u_{F_s}(\bar{v})$ defined as $u_{F_c} = \bar{0}$ and $u_{F_s \backslash F_c} = \bar{v}$. This implies that the code $\mathtt{C}_c$ can be partitioned into $\mathtt{C}_s(\bar{v})$ for $\bar{v} \in \mathcal{X}^{|F_s \backslash F_c|}$, i.e., $\mathtt{C}_c = \cup_{\bar{v}} \mathtt{C}_s(\bar{v})$. The code $\mathtt{C}_c$ (the set F_c) is designed to be a good channel code for the BSC(p). Further, for each $\bar{v}$, the code $\mathtt{C}_s(\bar{v})$ (the set F_s) is designed to be a good source code for distortion D.

The encoder uses the bits belonging to the set $F_s \backslash F_c$ for transmitting information. Therefore, the rate of transmission is $\frac{|F_s \backslash F_c|}{N}$. Let $\bar{V} = \bar{U}_{F_s \backslash F_c}$. The encoder maps the state vector $\bar{S}$ to $\hat{U}(\bar{S}, u_{F_s}(\bar{V}))$. Let $\bar{S}'$ be the reconstruction vector $\bar{S}' = \hat{U}(\bar{S}, u_{F_s}(\bar{V}))G_2^{\otimes n}$. The encoder transmits the vector $\bar{X} = \bar{S} \oplus \bar{S}'$ through the channel. Since the codes $\mathtt{C}_s(\bar{V})$ are good source codes, the expected distortion $\frac{1}{N}\mathbb{E}[\mathtt{d}(\bar{S}, \bar{S}')]$ is close to D (see Lemma 3.14). Therefore, the average weight of the encoder output $\bar{X}$ is also close to D and hence it satisfies the input constraint.

The output at the decoder is the result of sending $\bar{X} \oplus \bar{S}$ through the BSC(p). By design, $\bar{X} \oplus \bar{S} = \bar{S}'$ and $\bar{S}' \in \mathtt{C}_N(F_c, \bar{0})$. Since the frozen set F_c is designed to be a good channel code for the BSC(p), the decoder will successfully decode $\bar{S}'$ with high probability. The decoder then recovers the information bits $\bar{V}$ from $\bar{S}'$ as $\bar{V} = (\bar{S}' H_n^{-1})_{F_s \backslash F_c}$.

Here we have slightly simplified the discussion because, as we will see, it is not possible to find the sets F_c and F_s satisfying all the required properties. More precisely, we cannot show that $F_c \subseteq F_s$. Figure 4.5 provides a clear picture of the problem. However, what we can show is that the set $F_c \backslash F_s$ is very small. We therefore modify our transmission scheme to account for this discrepancy. However, the gist of the proof is the same as explained above.

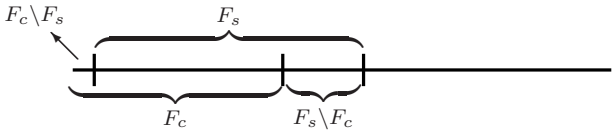

Figure 4.5: A pictorial representation of the subset structure of the frozen sets F_s and F_c. Let the line represent the set of indices. The indices in $F_s \backslash F_c$ are used for transmitting the message. The set $F_c \backslash F_s$ is not empty.

Theorem 4.2 (Optimality for the Gelfand-Pinsker Problem)**.** *Let S be a symmetric Bernoulli random variable. Fix D^*, $p < D^* < \frac{1}{2}$. For any rate $R < h_2(D^*) - h_2(p)$, expected weight $D > D^*$ and any $0 < \beta < \frac{1}{2}$, there exists a sequence of polar codes of length N so that under SC encoding using randomized rounding at the encoder and SC decoding at the decoder, the achievable rate satisfies*

$$R_N > R,$$

with the expected weight of X, D_N, satisfying

$$D_N \leq D.$$

Further, the block error probability satisfies

$$P_N \leq O(2^{-(N^\beta)}).$$

The encoding as well as decoding complexity of this scheme is $O(N \log(N))$.

Proof. Let $\epsilon > 0$ and $0 < \beta < \frac{1}{2}$ be some constants. Let $Z_N^{(i)}(q)$ denote $Z(W_N^{(i)})$, with W set to a BSC(q). Let $\delta_N = \frac{1}{N} 2^{-(N^\beta)}$. Let F_s and F_c denote the sets

$$F_s = \{i : Z_N^{(i)}(D^*) \geq 1 - \delta_N^2\}, \tag{4.4}$$

$$F_c = \{i : Z_N^{(i)}(p) \geq \delta_N\}. \tag{4.5}$$

Theorem 3.15 implies that for N sufficiently large

$$\frac{|F_s|}{N} \geq h_2(D^*) - \frac{\epsilon}{2}.$$

Similarly, Theorem 2.11 implies that for N sufficiently large

$$\frac{|F_c|}{N} \leq h_2(p) + \frac{\epsilon}{2}.$$

For the moment, assume that $F_c \subseteq F_s$. The vector $u_{F_s \setminus F_c}$ is defined by the message that is transmitted. Therefore, the rate of transmission is

$$\frac{|F_s| - |F_c|}{N} \geq h_2(D^*) - h_2(p) - \epsilon.$$

The vector $\bar{S}$ is compressed using the source code with frozen set F_s. The frozen vector u_{F_s} is defined in two stages. The subvector u_{F_c} is fixed to 0 and is known to both the transmitter and the receiver. The subvector $u_{F_s \setminus F_c}$ is defined by the message being transmitted.

Let $\bar{S}'$ be the reconstruction vector that $\bar{S}$ is mapped to. Lemma 3.14 implies that the average distortion is independent of the value of the frozen bits. This implies

$$\frac{1}{N} \mathbb{E}[\mathtt{d}(\bar{S}, \bar{S}')] \leq D^* + 2|F_s|\delta_N \leq D^* + O(2^{-(N^\beta)}).$$

Therefore, a transmitter which sends $\bar{X} = \bar{S} \oplus \bar{S}'$ will on average be using $D^* + O(2^{-(N^\beta)})$ fraction of 1s. The received vector is given by

$$\bar{Y} = \bar{X} \oplus \bar{S} \oplus \bar{Z} = \bar{S}' \oplus \bar{Z}.$$

The vector $\bar{S}'$ is a codeword of $\mathtt{C}_c$, the code designed for the BSC(p) (see (4.5)). Therefore, the block error probability of the SC decoder in decoding $\bar{S}'$ (and hence $\bar{V}$) is bounded as

$$P_N \leq \sum_{i \in F_c^c} Z_N^{(i)}(p) \leq O(2^{-(N^\beta)}).$$

The above discussion assumes $F_c \subseteq F_s$, which may not be true. Let us now consider the case when $F_c \nsubseteq F_s$. The indices in $F_c \cap F_s^c$ should be fixed for the channel decoding to succeed and they should be set free for the source encoding to succeed.

We therefore proceed with the following two phase approach. In the first phase, the bits belonging to the set $F_c \cap F_s^c$ are set free. Therefore, the source encoding will be successful. We accumulate the bits belonging to the set $F_c \cap F_s^c$ of many first phase transmissions and send them in the second phase.

In the second phase the transmitter uses a BSC(p) code, i.e., a polar code with frozen set $\{i : Z_N^{(i)}(p) \geq \delta_N\}$. Moreover, the decoder cancels the state noise $\bar{S}$ completely. Therefore, the block error probability of the second phase is $O(2^{-(N)^\beta})$. After the decoding the second phase, the decoder has the knowledge of the bits belonging to $F_c \cap F_s^c$ for various first phase transmissions. Now it proceeds with the decoding of the first phase. Therefore for the first phase also the block error probability is $O(2^{-(N)^\beta})$.

The second phase does not involve any transmission of information. Therefore it decreases the rate of transmission. Moreover, the second phase cancels the state completely and hence uses a fraction $\frac{1}{2}$ ones on average. Therefore, for the scheme to approach optimal performance the second phase must be much smaller than the first phase. Using channel polarization we show that this is indeed possible. For that purpose, consider the following modified set

$$\tilde{F}_c = \{i : Z_N^{(i)}(p) \geq 1 - \delta_N^2\}. \tag{4.6}$$

Since $D^* > p$, BSC(D^*) $\preceq$ BSC(p) which in turn implies $\tilde{F}_c \subseteq F_s$.

Theorem 2.11 and Theorem 3.15 together imply that for any $\eta > 0$ there exists N_0 such that for $N > N_0$, $\frac{1}{N}|F_c \backslash \tilde{F}_c| \leq \eta$. Therefore,

$$\frac{|F_c \cap F_s^c|}{N} = \frac{|F_c \backslash \tilde{F}_c \cap F_s^c|}{N} + \underbrace{\frac{|\tilde{F}_c \cap F_s^c|}{N}}_{=0} \leq \frac{|F_c \backslash \tilde{F}_c|}{N} \leq \eta.$$

Therefore, we require one second phase for every $O(\frac{1}{\eta})$ first phase transmissions. Hence the resulting penalty for both the rate and weight is $O(\eta)$, which can be made as small as desired. □

4.3 Lossless Compression and Slepian-Wolf Problem

Let us first consider the lossless compression of a Ber(p) source. The lossless compression problem can be mapped to the channel coding problem over the BSC(p). The mapping results in the following optimality result.

Theorem 4.3 (Optimality for Lossless Compression)**.** *Let X be a Ber(p) random variable. For any rate $R > h_2(p)$ there exists a sequence of polar codes of length N and rate $R_N < R$ so that under SC decoding at the decoder the source can be compressed losslessly with rate R_N. The encoding as well as the decoding complexity of this scheme is $O(N \log(N))$.*

Proof. Let $\epsilon > 0$, $R = h_2(p) + \epsilon$, and $0 < \beta < \frac{1}{2}$. Let $\bar{x} = (x_0, \ldots, x_{N-1})$ be a sequence of N i.i.d. realizations of the source. Let $\delta_N = \frac{1}{N} 2^{-(N)^\beta}$. Let F_N denote the set

$$F_N = \{i : Z_N^{(i)}(p) \geq \delta_N\}.$$

Theorem 2.11 implies that for N sufficiently large

$$\frac{|F_N|}{N} \leq h_2(p) + \frac{\epsilon}{2}.$$

The compression is done by mapping the vector $\bar{x}$ to its syndrome computed as $\bar{s} = (\bar{x}(G_2^{\otimes n})^{-1})_{F_N}$. The aim of the decoder is to reconstruct $\bar{x}$ with only the knowledge of $\bar{s}$.

The task of recovering $\bar{x}$ from $\bar{s}$ can be formulated as a channel decoding problem for the BSC(p). Let $\bar{x}$ be the input of a channel whose output is always $\bar{0}$. The noise during the transmission, denoted by $\bar{z}$, is given by $\bar{z} = \bar{x}$ ($\bar{0} = \bar{x} \oplus \bar{z}$). Since $\bar{x}$ is an i.i.d. Ber(p) source, it implies that the noise $\bar{z}$ is also i.i.d. Ber(p) vector. Consider a decoder which knows $\bar{s}$. By construction, $\bar{x} \in \mathtt{C}_N(F_N, \bar{s})$. Therefore, reconstructing $\bar{x}$ from $\bar{y} = \bar{0}$ is equivalent to decoding at the output of the BSC(p).

The above discussion not only maps the source compression problem to channel decoding, it also provides a decoding algorithm for the former problem. From the definition of F_N, we know that the SC decoder can perform this decoding with failure probability at most

$$P_N = \sum_{i \in F_N^c} Z_N^{(i)}(p) \leq O(2^{-(N)^\beta}). \tag{4.7}$$

However, the encoder knows the information seen by the decoder (unlike channel coding there is no noise involved here). Therefore, the encoder can replicate the decoding operation and check whether it is successful or not. In case of failure, the encoder will transmit the source vector $\bar{x}$ as is without

compression. We need only one additional bit to specify whether compression is done or not. Therefore, the required rate satisfies

$$R_N = \frac{|F_N|+1}{N} + P_N \overset{(4.7)}{\leq} \frac{|F_N|+1}{N} + O(2^{-(N)^\beta}).$$

Therefore for N sufficiently large, we have $R_N < R$.

It now remains to show that the scheme can be implemented with $O(N \log N)$ complexity. The encoding operation can be expressed as

$$\bar{x}(G_2^{\otimes n})^{-1} \overset{\text{Lem. } 4.8}{=} \bar{x}(G_2^{-1})^{\otimes n} \overset{G_2^{-1}=G_2}{=} \bar{x}G_2^{\otimes n}.$$

The above operation is equivalent to the encoding operation in channel coding with two additional permutations. Therefore, the encoding can be accomplished with complexity $O(N \log N)$. The decoding operation is implemented using the SC decoder, which is also of complexity $O(N \log N)$. □

Zero compression error is possible because the encoder knows whether the decoder can reconstruct the source vector successfully or not. This ability of the encoder can also be used to decrease the failure probability of the decoder with only a small loss in rate as follows [69]. In case of failure, the encoder can retry the compression procedure by using a permutation of the source vector. This permutation is fixed a priori and is known both to the encoder as well as the decoder. In order to completely specify the system, the encoder must inform the decoder which permutation was finally used. This results in a small loss of rate but it brings down the probability of decoding failure. Note that the extra number of bits that need to be transmitted grows only logarithmically with the number of permutations used, but that the error probability decays exponentially as long as the various permuted source vectors look like independent source samples. With this trick one can make the curves essentially arbitrarily steep with a very small loss in rate.

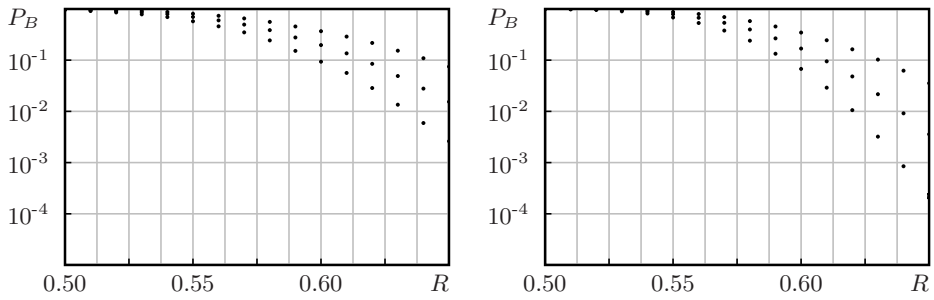

Figure 4.6: Comparison of SC decoding for a Ber(0.11) source with 0,1, and 2 bits for permutations. The performance curves are shown for $n = 10$ (left) and 11 (right). By increasing the number of permutations the curves can be made steeper and steeper.

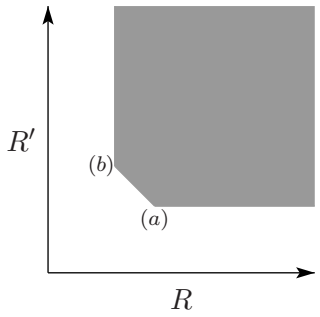

Figure 4.7: Typical rate region for the Slepian-Wolf problem. The corner point (a) corresponds to the rate pair $(1, h_2(p))$ and the corner point (b) corresponds to the rate pair $(h_2(p), 1)$.

Figure 4.6 shows the performance of polar codes for Ber(0.11) source. The entropy of this source is close to $\frac{1}{2}$. We also show the performance curves with 1 and 2 bits of permutations. We see that by increasing the number of permutations, the curves can be made steeper.

Let us now move to the Slepian-Wolf problem. Let X and X' be two BSSs and let $X = X' \oplus Z$, where $Z \sim \text{Ber}(p)$. The celebrated result by Slepian and Wolf [65] shows that the source (X, X') can be compressed to its entropy $H(X, X')$ by independently describing X and X'. Recall that the Slepian-Wolf rate region, denoted by R_{SW}, is the unbounded polytope

$$R_{\text{SW}} = \{(R, R') : R > 1, R' > 1, R + R' > 1 + h_2(p)\}.$$

The points $(R, R') = (1, h_2(p))$ and $(R, R') = (h_2(p), 1)$ are the *corner points.*

Theorem 4.4 (Optimality for Slepian-Wolf Problem)**.** *Let X and X' be two symmetric Bernoulli random variables such that $X = X' \oplus Z$ where $Z \sim Ber(p)$. For any rate $(R, R') \in R_{SW}$ and any $0 < \beta < \frac{1}{2}$, using polar codes combined with time-sharing there exists a sequence of polar codes of length N such that the rates used satisfy $R_N < R, R'_N < R'$ and probability of error satisfies*

$$P_N \leq O(2^{-(N^\beta)}),\ P'_N \leq O(2^{-(N^\beta)}).$$

The encoding as well as decoding complexity of this scheme is $O(N \log(N))$.

Proof. As shown in Figure 4.7, all points of the rate region (shaded) are dominated by at least one convex combination (time-sharing) of the two corner points. Therefore, achieving the corner points combined with time-sharing will imply the result. Because of symmetry it suffices to show how to achieve one such corner point, say $(1, h_2(p))$.

The source X is transmitted as is, i.e, $R_N = 1$. The compression of X' is treated as the lossless compression of a $\text{Ber}(p)$ source. Let F denote the frozen set of a code designed for communication over the $\text{BSC}(p)$. The encoder for X'

computes the syndrome $\bar{s} = (\bar{x}'(G_2^{\otimes n})^{-1})_F$ and transmits it to the receiver. At the receiver using $\bar{s}$ and $\bar{x}$, we compute $\bar{s}' = ((\bar{x} \oplus \bar{x}')(G_2^{\otimes n})^{-1})_F = (\bar{z}(G_2^{\otimes n})^{-1})_F$. The resulting problem of estimating $\bar{z}$ is equivalent to the decoding problem of lossless compression of a Ber(p) discussed in Theorem 4.3.

Unlike in lossless compression, we cannot achieve zero error probability because the encoder for the source X' does not know the information available at the decoder (the vector $\bar{x}$). □

4.4 One-Helper Problem

Let Y be a BSS. We want to reliably transmit Y to the decoder. Let Y' be another BSS which is correlated to Y in the following fashion: $Y' = Y \oplus Z$, where Z is a Ber(p) random variable. As shown in Figure 4.8, let Y' be available to another terminal which we refer to as helper. As the name suggests, the role of the helper is to assist the decoder in recovering Y.

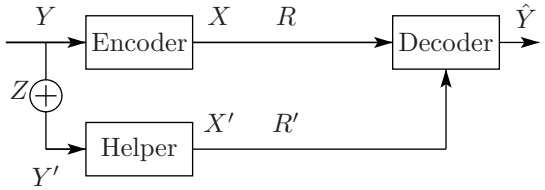

Figure 4.8: The helper transmits quantized version of Y'. The decoder uses the information from the helper to decode Y reliably.

Let the rates used by the encoder and the helper be R and R' respectively. Wyner [66] showed that the achievable rate-distortion tuples (R, R', D) satisfy

$$\{(R, R', D) : R > h_2(D * p), R' > 1 - h_2(D), D \in [0, 1/2]\}.$$

For this problem, we require a good channel code at the encoder and a good source code at the helper.

Theorem 4.5 (Optimality for the One Helper Problem)**.** *Let Y be a BSS and Y' be a Bernoulli random variable correlated to Y as $Y' = Y \oplus Z$, where $Z \sim Ber(p)$. Fix the* design *distortion D, $0 < D < \frac{1}{2}$. For any rate pair $R > h_2(D*p), R' > 1-h_2(D)$ and any $0 < \beta < \frac{1}{2}$, there exist sequences of polar codes of length N with rates $R_N < R$ and $R'_N < R'$ so that under syndrome computation at the encoder, SC encoding using randomized rounding at the helper and SC decoding at the decoder, they achieve the block error probability satisfying*

$$P_N \leq O(2^{-(N^\beta)}).$$

The encoding as well as decoding complexity of this scheme is $O(N \log(N))$.

Proof. Let $\epsilon > 0$ and $0 < \beta < \frac{1}{2}$ be some constants. Let $Z_N^{(i)}(q)$ denote $Z(W_N^{(i)})$, with W set to a BSC(q). Let $\delta_N = \frac{1}{N}2^{-(N^\beta)}$. The frozen sets of the helper and the encoder, namely F'_N and F_N, are defined as

$$F'_N = \{i : Z_N^{(i)}(D) \geq 1 - \delta_N^2\}, \tag{4.8}$$

$$F_N = \{i : Z_N^{(i)}(D * p) \geq \delta_N\}. \tag{4.9}$$

Theorem 3.15 and Theorem 2.11 imply that for N sufficiently large

$$\frac{|F'_N|}{N} \geq h_2(D) - \epsilon, \ \frac{|F_N|}{N} \leq h_2(D * p) + \epsilon.$$

The helper compresses the vector $\bar{Y}$ using the code $\mathtt{C}_N(F'_N, \bar{0})$. The helper transmits $(\hat{U}(\bar{Y}, u_{F'_N} = \bar{0}))_{F'^c_N}$ to the decoder. The rate of such a scheme is $R'_N \leq 1 - h_2(D) + \epsilon$.

The encoder transmits the syndrome $(\bar{Y}G_2^{\otimes n})_{F_N}$ to the decoder. The rate of this transmission satisfies $R_N \leq h_2(D * p) + \epsilon$. The decoder first constructs $\bar{Y}'' = \hat{U}G_2^{\otimes n}$ which is the reconstruction vector of $\bar{Y}$. The received vector can be expressed as

$$\bar{Y}'' = \bar{Y}' \oplus (\bar{Y}' \oplus \bar{Y}'') = \bar{Y} \oplus \bar{Z} \oplus \bar{E},$$

where $\bar{E} = \bar{Y}' \oplus \bar{Y}''$ quantization noise.

The remaining task is to decode the codeword $\bar{Y}$ belonging to the code $\mathtt{C}_N(F_N, (\bar{Y}G_2^{\otimes n})_{F_N})$ from the observation $\bar{Y}''$. As shown in the Wyner-Ziv setting the noise $\bar{E} \oplus \bar{Z}$ is very "close" to Ber($D * p$). Using the coupling argument of Theorem 4.1, we can show that the block error probability satisfies

$$P_N \leq O(2^{-(N^\beta)}).$$

□

4.5 Non-Binary Polar Codes

Polar codes for non-binary, say q-ary, channels were recently constructed by Arıkan and Telatar [70]. These codes achieve the symmetric capacity of the q-ary DMC. Let $\mathcal{X} = \{0, \ldots, q-1\}$ be the input of the channel and let $W : \mathcal{X} \to \mathcal{Y}$ be a q-DMC. Then using polar codes and SC decoding, one can achieve rates close to $I(W)$, where

$$I(W) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \frac{1}{q} W(y \,|\, x) \log \frac{W(y \,|\, x)}{\frac{1}{q}\sum_{x'=0}^{q-1} W(y \,|\, x')}. \tag{4.10}$$

In the rest of this section we discuss polar codes for asymmetric channels, broadcast channels as well as multiple access channels. The reason for grouping these three problems together with non-binary codes is that we can map all these problems to the construction of polar codes for appropriately defined q-ary channels.

4.5.1 Asymmetric Channels

Consider an asymmetric B-DMC, e.g., the Z-channel. Due to the asymmetry, the capacity-achieving input distribution is in general not the uniform one. To be concrete, assume that it is $(p(0) = \frac{1}{3}, p(1) = \frac{2}{3})$. This causes problems for any scheme which employs linear codes, since linear codes induce uniform marginals. To get around this problem, "augment" the channel to a q-ary input channel by duplicating some of the inputs. For our running example, Figure 4.9 shows the ternary channel which results when duplicating the input "1." Note that the capacity-achieving input distribution for this ternary-input

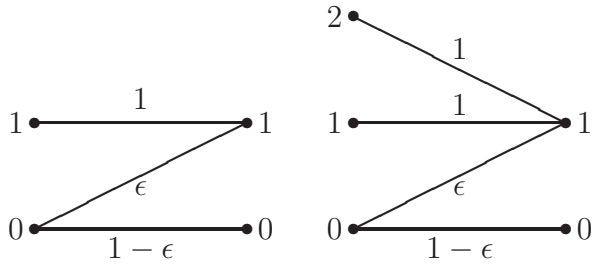

Figure 4.9: The Z-channel and its corresponding augmented channel with ternary input alphabet.

channel is the uniform one. Assume that we can construct a ternary polar code which achieves the symmetric mutual information of this new channel. Then this gives rise to a capacity-achieving coding scheme for the original binary Z-channel by mapping the ternary set $\{0, 1, 2\}$ into the binary set $\{0, 1\}$ in the following way; $\{1, 2\} \mapsto 1$ and $0 \mapsto 0$.

More generally, by augmenting the input alphabet and constructing a code for the extended alphabet, we can achieve rates arbitrarily close to the capacity of a q-ary DMC, assuming only that we know how to achieve the symmetric mutual information.

A similar remark applies to the setting of source coding. By extending the reconstruction alphabet if necessary and by using only test channels that induce a uniform distribution on this extended alphabet one can achieve a rate-distortion performance arbitrarily close to the Shannon bound, assuming only that for the uniform case we can get arbitrarily close.

4.5.2 Degraded Broadcast Channels

Consider a multi-terminal scenario with one transmitter and two receivers. Let the channel between the transmitter and the two receivers be $W_1(y_1 \,|\, x)$ and $W_2(y_2 \,|\, x)$. Let $W_2 \preceq W_1$, i.e., the channel W_2 is degraded with respect to W_1. Such a channel is known as the degraded broadcast channel.

The degradation implies that user 1 can decode whatever user 2 can decode. Therefore, the information received by user 2 is common for both the users.

We therefore refer to the rate received by the second user as common rate and denote it as R_c. Since user 1 has a better channel, it is possible to send additional information to user 1 which user 2 cannot decode. We refer to the rate of this additional information as private rate and denote it as R_p. Note that the polar code design for $R_p = 0$ is easy. This follows from the fact that $W_{2_N}^{(i)} \preceq W_{1_N}^{(i)}$ for all i and hence the code designed for W_2 (the worse channel) will work for W_1 (the better channel) as well.

The capacity region for this channel is given by [71, Section 15.6.3]

$$\{(R_p, R_c) : R_p \leq I(X; Y_1 \,|\, U),\, R_c \leq I(U; Y_2) \\ \forall p(u)p(x \,|\, u)W_1(y_1 \,|\, x)W_2(y_2 \,|\, x)\}. \tag{4.11}$$

The random variable U is an auxiliary variable whose cardinality is bounded by $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\}$. Since we restrict here to $\mathcal{X} = \{0, 1\}$, we have $|\mathcal{U}| \leq 2$. It is clear that there is a trade-off between the common rate and the private rate. Figure 4.10 shows the typical capacity region for degraded broadcast channels. By varying the distributions $p(u)$ and $p(x \,|\, u)$, we achieve different points in this region.

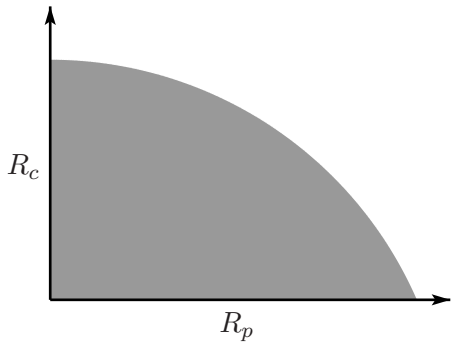

Figure 4.10: Typical capacity region for a two-user degraded broadcast channel. Every distribution $p(u)p(x \,|\, u)$, corresponds to a point in this region.

For simplicity, let us consider the case where W_1 and W_2 are both BSCs with flip probability $p_1 \leq p_2$. The capacity region for this channel is given by [71, Example 15.6.5]

$$\{(R_p, R_c) : R_p \leq h_2(\alpha * p_1) - h_2(p_1),\ R_c \leq 1 - h_2(\alpha * p_2)\ \ \forall \alpha \in [0, 1/2]\}, \tag{4.12}$$

where $\alpha * p = \alpha(1 - p) + (1 - \alpha)p$. Evaluating (4.11) with U fixed to $\mathrm{Ber}(\frac{1}{2})$ and $p(x \,|\, u)$ as the transition probability of the $\mathrm{BSC}(\alpha)$ $\forall \alpha \in [0, \frac{1}{2}]$, covers the whole capacity region (4.12).

Let $\bar{X}_c$ and $\bar{X}_p$ denote the codewords for the common and the private message respectively. The transmitter sends $\bar{X} = \bar{X}_c \oplus \bar{X}_p$. The outputs at the two receivers are

$$Y_1 = \bar{X}_c \oplus \bar{X}_p \oplus \bar{Z}_1,$$

$$Y_2 = \bar{X}_c \oplus \bar{X}_p \oplus \bar{Z}_2,$$

where $\bar{Z}_1$ and $\bar{Z}_2$ are i.i.d. Ber(p_1) and Ber(p_2) vectors respectively.

Fix an α. Let us first assume that the codewords $\bar{X}_p$ are statistically equivalent to an i.i.d. Ber(α) distribution. Then the channel between the input $\bar{X}_c$ and the output of user 2, i.e., Y_2, is the BSC($\alpha * p_2$). Using polar codes, we can achieve rates close to $1 - h_2(\alpha * p_2)$ for receiver 2. Since W_2 is degraded with respect to W_1, $\bar{X}_c$ can be decoded at receiver 1. Therefore a common rate of $1 - h_2(\alpha * p_2)$ is achievable.

Since $\bar{X}_c$ is known at receiver 1, its contribution can be removed and hence the resulting channel for $\bar{X}_p$ is the BSC(p_1). The capacity of such a channel with input weight restricted to α is given by $h_2(\alpha * p_1) - h_2(p_1)$. To construct polar codes with non-uniform marginals ($\alpha \neq \frac{1}{2}$), we use q-ary polar codes as shown in the previous section. For example, consider $\alpha = \frac{1}{3}$. Consider a channel with input alphabet $\{-1, 0, 1\}$, where -1 plays the role of 0. The transition probabilities of such a channel are shown in Figure 4.11.

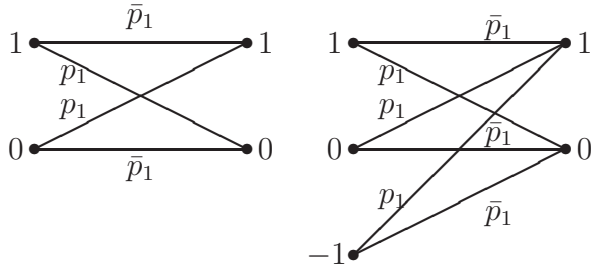

Figure 4.11: The BSC and its corresponding augmented channel with ternary input alphabet $\{-1, 0, 1\}$.

Under the mapping $\{-1, 0\} \mapsto 0, 1 \mapsto 1$, the ternary channel with uniform distribution over the input would create a Ber(1/3) distribution for the BSC. The symmetric mutual information for the ternary channel is indeed equal to $h_2(1/3 * p_1) - h_2(p_1)$.

The above scheme is optimal for any two-user degraded broadcast channel whose capacity region is achieved by evaluating (4.11) with $p(x \mid u)$ restricted to BSCs. Similar results can be obtained for more than two users.

4.5.3 Multiple Access Channels

Let us now consider a multi-terminal scenario with two transmitters and one receiver. Such a channel is called a multiple access channel. Let $W : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$ be the channel between the two transmitters and the receiver. The capacity region of this channel is given by [71, Theorem 15.3.1]

$$\{(R_1, R_2) : R_1 \leq I(X_1; Y \mid X_2),\ R_2 \leq I(X_2; Y \mid X_1),\\ R_1 + R_2 \leq I(X_1, X_2; Y)\ \ \forall p(x_1)p(x_2)\}. \tag{4.13}$$

The achievable rate region for any fixed input distribution $p(x_1)p(x_2)$ forms a pentagon. An example of such a region is shown in Figure 4.12. It is sufficient to provide a scheme to achieve the corner points (a) and (b) (see Figure 4.12). The remaining rate pairs can be achieved by appropriate time sharing.

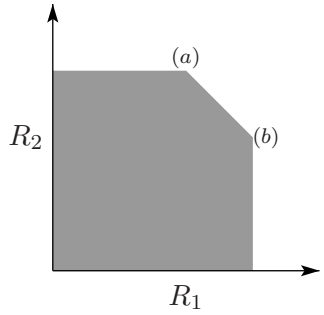

Figure 4.12: Typical rate region of a two-user multiple access channel for a fixed input distribution $p(x_1)p(x_2)$. The corner point (a) corresponds to the rate pair $(I(X_1;Y), I(X_2;Y \mid X_1))$ and the corner point (b) corresponds to the rate pair $(I(X_1;Y \mid X_2), I(X_2;Y))$.

For simplicity let the channel inputs be binary, i.e., $\mathcal{X}_1 = \mathcal{X}_2 = \{0,1\}$. Fix a probability distribution $p(x_1)p(x_2)$. Let $p(x_1) = \text{Ber}(\alpha_1)$ and let $p(x_2) = \text{Ber}(\alpha_2)$. Let us focus on achieving the corner point (a), i.e, $R_1 = I(X_1;Y)$ and $R_2 = I(X_2;Y \mid X_1)$. This rate pair is achieved by first decoding the codeword of user 1 $(\bar{X}_1)$, and then decoding the codeword of user 2 $(\bar{X}_2)$. The other corner point is achieved by reversing the order of decoding. If $\alpha_1 \neq \frac{1}{2}$ or $\alpha_2 \neq \frac{1}{2}$, then we use the augmentation technique, as shown in the previous sections, to create the required non-uniform distribution. Therefore, let us assume that the codewords $\bar{X}_1$ and $\bar{X}_2$ have marginals $\text{Ber}(\alpha_1)$ and $\text{Ber}(\alpha_2)$ respectively.

Let P_α denote the $\text{Ber}(\alpha)$ distribution. Since,$\bar{X}_1$ is decoded first, the effective channel between $\bar{X}_1$ and $\bar{Y}$ is given by

$$W_1(y \mid x_1) = \sum_{x_2} P_{\alpha_2}(x_2) W(y \mid x_1, x_2).$$

Therefore, the polar code for user 1 is designed for W_1. Note that the term $I(X_1;Y)$ corresponds to the mutual information of W_1.

After decoding $\bar{X}_1$, we proceed to decode $\bar{X}_2$. Therefore, $\bar{X}_1$ is now a part of the output at the decoder and hence the effective channel for user 2 is given by

$$W_2(y, x_1 \mid x_2) = P_{\alpha_1}(x_1) W(y \mid x_1, x_2).$$

Therefore, the polar code for user 2 is designed for W_2. Note that the term $I(X_2;Y \mid X_1)$ corresponds to the mutual information of the channel W_2.

The above mentioned scheme achieves the corner point (a). To achieve the corner point (b) we use a similar scheme with the roles of user 1 and user 2

swapped. Since this scheme works for any α_1 and α_2, we can achieve all the rates belonging to the capacity region (4.13).

For simplicity we have considered binary input channels in the above scenarios. However, the arguments can be extended to non-binary channels or even a combination of both asymmetric and non-binary channels. In theory we can achieve optimal rates as long as the input distribution that we need to approximate is a rational number. For irrational numbers we can approach as close as desired by considering larger and larger q. However, the decoding complexity of q-ary polar codes scales as $O(q^2 N \log N)$. In practice the alphabet size can therefore not be chosen too large.

4.A Appendix

The first part of this section deals with the distribution of the quantization error for source coding. The encoding function for the polar code $\mathtt{C}_N(F, u_F)$, denoted as $\hat{U}(\bar{y}, u_F)$, is defined in Section 3.2. Let $P_{\bar{E}}^{u_F}$ denote the distribution of the quantization noise, i.e.,

$$P_{\bar{E}}^{u_F}(\bar{x}) = \mathbb{E}[\mathbb{1}_{\{\bar{Y} \oplus (\hat{U}(\bar{Y}, u_F) G_2^{\otimes n}) = \bar{x}\}}],$$

where the expectation is over the randomness involved in the source and randomized rounding. For a BSS the test channel is the BSC(D), which is symmetric and the distortion function is also symmetric. Section 3.4 implies that the frozen indices can be set to any value. Using similar reasoning as in Lemma 3.14, we can show that the distribution $P_{\bar{E}}^{u_F}$ is independent of u_F. Combining this with Lemma 3.5, we can bound the distance between $P_{\bar{E}}^{u_F}$ and an i.i.d. Ber(D) noise (denoted as $P_{(D)}$) as follows.

Lemma 4.6 (Distribution of Quantization Error)**.** *Let the frozen set F be*

$$F = \{i : Z(W_N^{(i)}) \geq 1 - 2\delta_N^2\}.$$

Then for any $u_F \in \mathcal{X}^{|F|}$,

$$\sum_{\bar{x}} |P_{\bar{E}}^{u_F}(\bar{x}) - \prod_i P_{(D)}(x_i)| \leq 2|F|\delta_N.$$

Proof. Let $\bar{v} \in \mathcal{X}^{|F|}$ and let $A(\bar{v}) \subset \mathcal{X}^N$ denote the coset $A(\bar{v}) = \{\bar{z} : (\bar{z}(G_2^{\otimes n})^{-1})_F = \bar{v}\}$. The set $\mathcal{X}^N$ can be partitioned as

$$\mathcal{X}^N = \cup_{\bar{v} \in \mathcal{X}^{|F|}} A(\bar{v}).$$

Consider a vector $\bar{y} \in A(\bar{v})$ and set $\bar{y}' = \bar{0}$. Lemma 3.13 implies that

$$\bar{y} \oplus \hat{U}(\bar{y}, u_F) G_2^{\otimes n} = \bar{0} \oplus \hat{U}(\bar{0}, u_F \oplus \bar{v}) G_2^{\otimes n}.$$

This implies that all vectors belonging to $A(\bar{v})$ have the same quantization error and this error is equal to the error incurred by the all-zero word when the frozen bits are set to $u_F \oplus \bar{v}$.

Moreover, the uniform distribution of the source induces a uniform distribution on the cosets $\{A(\bar{v}) : \bar{v} \in \mathcal{X}^{|F|}\}$. Therefore, the distribution of the quantization error $P_{\bar{E}}^{u_F}$ is equivalent to first picking the coset uniformly at random (i.e., the bits in F are generated by a $\mathrm{Ber}(\frac{1}{2})$ distribution) and then generating the error $\bar{x}$ according to $\bar{x} = \hat{U}(\bar{0}, u_F)G_2^{\otimes n}$. In other words, $P_{\bar{E}}^{u_F}(\bar{x}) = \mathbb{1}_{\{\bar{x}=\bar{u}G_2^{\otimes n}\}}Q(\bar{u} \,|\, \bar{0})$ with Q as defined in (3.10). Consider the distribution $P_{\bar{U},\bar{X},\bar{Y}}$ defined in (3.6). Note that since W is the BSC(D), $P_{\bar{X} \,|\, \bar{Y}}(\bar{x} \,|\, \bar{0}) = \prod_{i=0}^{N-1} P_{(D)}(x_i)$ and hence the Bernoulli distribution can be expressed as

$$\prod_i P_{(D)}(x_i) = \mathbb{1}_{\{\bar{x}=\bar{u}G_2^{\otimes n}\}} P_{\bar{U} \,|\, \bar{Y}}(\bar{u} \,|\, \bar{0}).$$

Therefore

$$\begin{aligned}
&\sum_{\bar{x}} |P_{\bar{E}}^{u_F}(\bar{x}) - \prod_i P_{(D)}(x_i)| \\
&\quad = \sum_{\bar{u}} |Q(\bar{u} \,|\, \bar{0}) - P_{\bar{U} \,|\, \bar{Y}}(\bar{u} \,|\, \bar{0})| \\
&\quad \overset{(a)}{=} \sum_{\bar{u}} \frac{1}{2^N} \sum_{\bar{y}} |Q(\bar{u} \oplus (\bar{y}(G_2^{\otimes n})^{-1}) \,|\, \bar{y}) - P_{\bar{U} \,|\, \bar{Y}}(\bar{u} \oplus (\bar{y}(G_2^{\otimes n})^{-1}) \,|\, \bar{y})| \\
&\quad = \frac{1}{2^N} \sum_{\bar{y}} \sum_{\bar{u}} |Q(\bar{u} \,|\, \bar{y}) - P_{\bar{U} \,|\, \bar{Y}}(\bar{u} \,|\, \bar{y})| \\
&\quad \overset{(b)}{\leq} 2|F|\delta_N.
\end{aligned}$$

The equality (a) follows from the symmetry (arising from the channel symmetry of W) of the distributions Q and P. The inequality (b) follows from Lemma 3.5 and Lemma 3.8. □

Lemma 4.7 (Degradation of $W_N^{(i)}$)**.** *Let $W : \mathcal{X} \to \mathcal{Y}$ and $W' : \mathcal{X} \to \mathcal{Y}'$ be two B-DMCs such that $W \preceq W'$ then for all i, $W_N^{(i)} \preceq W'^{(i)}_N$ and hence $Z(W_N^{(i)}) \geq Z(W'^{(i)}_N)$.*

Proof. Let the n-bit ($N = 2^n$) binary expansion of i be $b_1, \ldots, b_n$. Recall from Section 2.3 that $W_N^{(i)} = W_{(b_1,\ldots,b_n)}$ where

$$W_{(b_1,\ldots,b_{k-1},b_k)} = \begin{cases} W_{(b_1,\ldots,b_{k-1})} \circledast W_{(b_1,\ldots,b_{k-1})}, & \text{if } b_k = 0, \\ W_{(b_1,\ldots,b_{k-1})} \circledcirc W_{(b_1,\ldots,b_{k-1})}, & \text{if } b_k = 1, \end{cases}$$

where

$$W_{(b_1)} = \begin{cases} W \circledast W, & \text{if } b_1 = 0, \\ W \circledcirc W, & \text{if } b_1 = 1. \end{cases}$$

Therefore to show that $W_N^{(i)} \preceq W'^{(i)}_N$, it is sufficient to show that both $\boxplus$ and $\circledast$ operations preserve degradation. Definition 1.7 implies that there exists a channel $W'' : \mathcal{Y}' \to \mathcal{Y}$ such that

$$W(y \,|\, x) = \sum_{y' \in \mathcal{Y}'} W'(y' \,|\, x) W''(y \,|\, y'). \tag{4.14}$$

Consider the channel $W \boxplus W$.

$$\begin{aligned}
&(W \boxplus W)(y_1, y_2 \,|\, x) \\
&= \frac{1}{2} \sum_u W(y_1 \,|\, u \oplus x) W(y_2 \,|\, x) \\
&\overset{(4.14)}{=} \frac{1}{2} \sum_u \sum_{y'_1 \in \mathcal{Y}'} W'(y'_1 \,|\, u \oplus x) W''(y_1 \,|\, y'_1) \sum_{y'_2 \in \mathcal{Y}'} W'(y'_2 \,|\, u \oplus x) W''(y_2 \,|\, y'_2) \\
&= \frac{1}{2} \sum_u \sum_{(y'_1, y'_2) \in \mathcal{Y}'^2} W'(y'_1 \,|\, u \oplus x) W'(y'_2 \,|\, u \oplus x) W''(y_1 \,|\, y'_1) W''(y_2 \,|\, y'_2).
\end{aligned}$$

From Definition 1.7 it follows that $W \boxplus W \preceq W' \boxplus W'$. Using similar arguments we can show that $W \circledast W \preceq W' \circledast W'$. The statement about the Bhattacharyya parameters follows from Lemma 1.8. □

Lemma 4.8 (Inverse of Kronecker Product [72])**.** *Let G be a square matrix and let $G^{\otimes n}$ denote the n-th Kronecker product of G. Then $(G^{\otimes n})^{-1} = (G^{-1})^{\otimes n}$.*

The proof follows from Corollary 4.2.11 of [72].

5 Exponent of Polar Codes

In this chapter we consider a generalization of polar codes. The original polar code construction is based on combining two channels at a time using the matrix G_2,

$$G_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{5.1}$$

It was conjectured in [32] that polarization is a general phenomenon, and that is not restricted to the particular transformation $G_2^{\otimes n}$. In this chapter we prove this conjecture. In particular, we consider transformations of the form $G^{\otimes n}$ where G is an $\ell \times \ell$ matrix for $\ell \geq 3$ and provide necessary and sufficient conditions for such Gs to polarize symmetric B-DMCs.

In Chapter 2 we have seen that the matrix G_2 results in a block error probability of $O(2^{-(N)^\beta})$ for any fixed $\beta < \frac{1}{2}$. We say that G_2 has *channel coding exponent* $\frac{1}{2}$. Similarly, in Chapter 3 we have seen that the distortion approaches the design distortion D as $D + O(2^{-(N)^\beta})$ for any fixed $\beta < \frac{1}{2}$. Therefore, we say that G_2 has *source coding exponent* $\frac{1}{2}$. We show that this exponent can be improved by considering larger matrices instead of G_2. In fact, the exponent can be made arbitrarily close to 1.

The exponent for a given matrix G in the context of channel coding was analyzed in joint work with Şaşoğlu and Urbanke in [48]. Here, we concentrate mainly on the exponent for source coding and only state the results for channel coding.

The chapter is organized as follows. We first discuss the recursive channel transform using $\ell \times \ell$ matrices (Section 5.1). In Section 5.2, we provide necessary conditions for a matrix to be suitable for source coding. In Section 5.3 we characterize the source coding exponent of a given matrix. We provide similar results for channel coding in Section 5.4. In Section 5.5 we discuss a duality

relationship between the two exponents. In Section 5.6 we provide bounds on the best possible exponent for any $\ell \times \ell$ matrix. Finally in Section 5.7 we give an explicit construction of a family of matrices, derived from BCH codes, with exponent approaching 1 for large ℓ.

In the following, we will use $\mathtt{C}$ to denote a linear code and $\mathtt{dmin}(\mathtt{C})$ to denote its minimum distance. We let $\langle g_1, \ldots, g_k \rangle$ denote the linear code generated by the vectors $g_1, \ldots, g_k$. We let $d_H(a, b)$ denote the Hamming distance between binary vectors a and b. We also let $d_H(a, \mathtt{C})$ denote the minimum distance between a vector a and a code $\mathtt{C}$, i.e., $d_H(a, \mathtt{C}) = \min_{c \in \mathtt{C}} d_H(a, c)$.

5.1 Channel Transform using an $\ell \times \ell$ Matrix

Let G be an $\ell \times \ell$ invertible matrix with entries in $\{0, 1\}$. Consider a random ℓ-vector $U_0^{\ell-1}$ that is uniformly distributed over $\{0, 1\}^\ell$. Let $X_0^{\ell-1} = U_0^{\ell-1} G$, where the multiplication is performed over GF(2). Also, let $Y_0^{\ell-1}$ be the output of ℓ uses of W when the input is $X_0^{\ell-1}$. The channel between $U_0^{\ell-1}$ and $Y_0^{\ell-1}$ is defined by the transition probabilities

$$W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1}) \triangleq \prod_{i=0}^{\ell-1} W(y_i \,|\, x_i) = \prod_{i=0}^{\ell-1} W(y_i \,|\, (u_0^{\ell-1} G)_i).$$

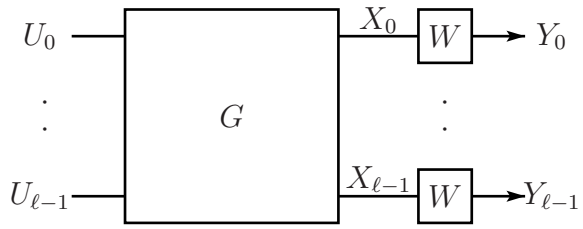

Figure 5.1: One step of the channel combining operation using the matrix G.

Using the chain rule, the mutual information between $U_0^{\ell-1}$ and $Y_0^{\ell-1}$ can be expressed as

$$I(U_0^{\ell-1}; Y_0^{\ell-1}) = \sum_{i=0}^{\ell-1} I(U_i; Y_0^{\ell-1}, U_0^{i-1}).$$

Let $W_\ell^{(i)} : \mathcal{X} \to \mathcal{Y}^\ell \times \mathcal{X}^{i-1}$ denote the channel with input u_i, output $(y_0^{\ell-1}, u_0^{i-1})$ and transition probabilities

$$W_\ell^{(i)}(y_0^{\ell-1}, u_0^{i-1} \,|\, u_i) = \frac{1}{2^{\ell-1}} \sum_{u_{i+1}^{\ell-1}} W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1}). \tag{5.2}$$

It is easy to check that $I(U_i; Y_0^{\ell-1}, U_0^{i-1}) = I(W_\ell^{(i)})$. Moreover, since G is a bijection, $I(U_0^{\ell-1}; Y_0^{\ell-1}) = I(X_0^{\ell-1}; Y_0^{\ell-1}) = \ell I(W)$. Therefore,

$$\sum_{i=0}^{\ell-1} I(W_\ell^{(i)}) = \ell I(W). \tag{5.3}$$

Let us now consider the recursive channel combining operation using G. We proceed similar to Section 2.2. Applying n recursions of this construction combines $N = \ell^n$ channels. Let v_0^{N-1} be defined as

$$v_{i\ell}^{(i+1)\ell-1} = u_{i\ell}^{(i+1)\ell-1} G, \tag{5.4}$$

for all $0 \leq i \leq \ell^{n-1} - 1$. For $0 \leq k \leq \ell - 1$, let $v_{0,k}^j$ denote the subvector $(v_k, v_{\ell+k}, \ldots, v_{m\ell+k})$ where $m\ell \leq j < (m+1)\ell$. The channel W_N is defined as

$$W_N(y_0^{N-1} \,|\, u_0^{N-1}) = \prod_{k=0}^{\ell-1} W_{N/\ell}(y_{kN/\ell}^{(k+1)N/\ell-1} \,|\, v_{0,k}^{N-1}). \tag{5.5}$$

Figure 5.2 provides a pictorial representation for the recursion (5.5). The interleaver R_N^ℓ maps the vector $v_{0,k}^{N-1}$ to the input of the k-th $W_{N/\ell}$ channel. It

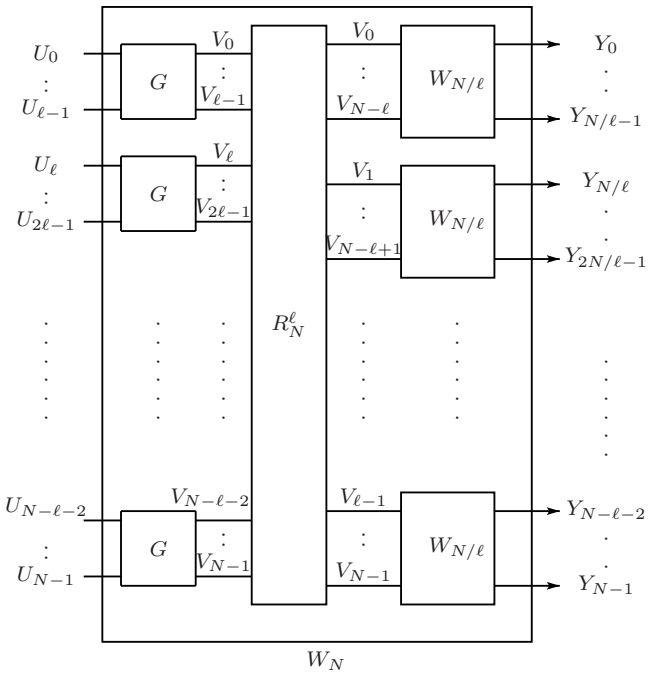

Figure 5.2: Recursive channel combining operation using the matrix G.

can be shown that the recursive combining operation is equivalent to applying the linear transform $G^{\otimes n}$, i.e., $X_0^{N-1} = U_0^{N-1} G^{\otimes n}$. Using the chain rule, we expand the mutual information between U_0^{N-1} and Y_0^{N-1} as

$$I(U_0^{N-1}; Y_0^{N-1}) = \sum_{i=0}^{N-1} I(U_i; Y_0^{N-1}, U_0^{i-1}).$$

The term $I(U_i; Y_0^{N-1}, U_0^{i-1})$ corresponds to the channel between U_i and the output (Y_0^{N-1}, U_0^{i-1}). Let us denote this channel by $W_N^{(i)} : \mathcal{X} \to \mathcal{Y}^N \times \mathcal{X}^{i-1}$. The transition probabilities are given by

$$W_N^{(i)}(y_0^{N-1}, u_0^{i-1} \,|\, u_i) = \frac{1}{2^{N-1}} \sum_{u_{i+1}^{N-1}} W_N(y_0^{N-1} \,|\, u_0^{N-1}).$$

The following lemma shows the relationship between the channels $\{W_N^{(i)}\}$ and the channels $\{W_{N/\ell}^{(i)}\}$ which is crucial for the analysis later. The proof is similar to the proof of Lemma 2.3. Recall that v_0^{N-1} is given by (5.4).

Lemma 5.1 (Recursive Relation between $W_N^{(i)}$ and $W_{N/\ell}^{(i)}$)**.** *For* $0 \le i \le N/\ell - 1$ *and* $0 \le j \le \ell - 1$,

$$\begin{aligned} &W_N^{(\ell i+j)}(y_0^{N-1}, u_0^{\ell i+j-1} \,|\, u_{\ell i+j}) \\ &\qquad = \frac{1}{2^{\ell-1}} \sum_{u_{\ell i+j+1}}^{u_{(\ell+1)i-1}} \prod_{k=0}^{\ell-1} W_{N/\ell}^{(i)}(y_{kN/\ell}^{(k+1)N/\ell-1}, v_{0,k}^{\ell i-1} \,|\, (u_{\ell i}^{(\ell+1)i-1} G)_k). \end{aligned} \tag{5.6}$$

For any B-DMC $W' : \mathcal{X} \to \mathcal{Y}'$ let $W'^{[i]} : \mathcal{X} \to \mathcal{Y}'^{\ell} \times \mathcal{X}^{i-1}$ denote a B-DMC with transition probabilities

$$W'^{[i]}(y_0^{\ell-1}, u_0^{i-1} \,|\, u_i) = \frac{1}{2^{\ell-1}} \sum_{u_{i+1}^{\ell}} \prod_{i=0}^{\ell-1} W'(y_i \,|\, ((u_0^{\ell-1})G)_i).$$

From the definition it is clear that $W_\ell^{(i)}$ of (5.2) is equal to $W^{[i]}$. Applying the relabeling argument of Section 2.2 we can show that the channel law of $W_N^{(\ell i+j)}$ is equivalent to

$$W_N^{(\ell i+j)} \equiv (W_{N/\ell}^{(i)})^{[j]}.$$

Using this relation recursively we can express the channel $W_N^{(i)}$ as follows. Let $l_1 \ldots l_n$ denote the n-digit ℓ-ary expansion of i and let $W_{(l_1,\ldots,l_n)} \triangleq W_N^{(i)}$. Then $W_N^{(i)}$ can be obtained by repeating the following operation,

$$W_{(l_1,\ldots,l_{k-1},l_k)} = W_{(l_1,\ldots,l_{k-1})}^{[l_k]} \tag{5.7}$$

where $W_{(l_1)} = W^{[l_1]}$.

Recall that in Section 2.3 to analyze the channels $\{W_N^{(i)}\}$ we define a random variable W_n that is uniformly distributed over the set $\{W_{\ell^n}^{(i)}\}_{i=0}^{\ell^n-1}$ (where $\ell = 2$ for the case $G = G_2$). The random variable W_n for our purpose is defined through a tree process $\{W_n : n \ge 0\}$ with $W_0 = W$ and

$$W_{n+1} = W_n^{[B_{n+1}]},$$

where $\{B_n; n \geq 1\}$ is a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, \mu)$, and where B_n is uniformly distributed over the set $\{0, \ldots, \ell - 1\}$. Defining $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(B_1, \ldots, B_n)$ for $n \geq 1$, we augment the above process by the processes $\{I_n : n \geq 0\} := \{I(W_n) : n \geq 0\}$ and $\{Z_n : n \geq 0\} := \{Z(W_n) : n \geq 0\}$.

Using the chain rule (5.3), we can show that the process $\{I_n\}$ satisfies the following lemma.

Lemma 5.2 ($\{I_n\}$ is a Martingale)**.** *$\{(I_n, \mathcal{F}_n)\}$ is a bounded martingale and therefore converges w.p. 1 and in $\mathcal{L}^1$ to a random variable I_∞.*

The proof is similar to the proof of Lemma 2.5. However, we cannot claim any such property for the process $\{Z_n\}$.

5.2 Polarizing Matrices

In this section we show that most matrices result in channel polarization by applying the recursive operation of the previous section.

Note that for a matrix G the performance is not effected if we permute the columns since this is equivalent to just relabeling the output. We claim that any invertible $\{0,1\}$ matrix G can be written as a real sum $G = P + P'$, where P is a permutation matrix, and P' is a $\{0,1\}$ matrix. In other words, there is a permutation matrix embedded in any invertible $\{0,1\}$ matrix. To see this, consider a bipartite graph whose biadjacency matrix is G. The graph consists of 2ℓ nodes. The ℓ left nodes correspond to the rows of the matrix and the ℓ right nodes correspond to the columns of the matrix. Connect left node i to right node j if $G_{ij} = 1$. The invertibility of G implies that for every subset of rows $\mathcal{R}$ the number of columns which contain non-zero elements in these rows is at least $|\mathcal{R}|$. Otherwise, there is are $|\mathcal{R}|$ rows with non-zero entries in at most $|\mathcal{R}| - 1$ indices. Therefore, these $|\mathcal{R}|$ rows are linearly dependent. By Hall's Theorem [73, Theorem 16.4.] this guarantees that there is a matching between the left and the right nodes of the graph and this matching represents a permutation. Therefore, for any invertible matrix G, there exists a column permutation so that all diagonal elements of the permuted matrix are 1. Therefore, from now on, and without loss of generality, we assume that that G has 1s on its diagonal.

For $k \geq 1$, let $W^{\boxtimes k} : \mathcal{X} \to \mathcal{Y}^k$ denote the B-DMC with transition probabilities

$$W^{\boxtimes k}(y_1^k \,|\, x) = \frac{1}{2^{k-1}} \sum_{x_1 \oplus \cdots \oplus x_k = x} \prod_{j=1}^{k} W(y_j \,|\, x_j). \tag{5.8}$$

In the following lemma we show that for any one to one transform G either all the channels $W^{[i]}$ remain the same or at least one channel transforms as $W^{\boxtimes k}$, for some $k \geq 2$.

Lemma 5.3 (Polarizing Matrices)**.** *Let W be a B-DMC.*

(i) If G is not upper triangular, then there exists an i for which $W^{[i]} \equiv W^{\boxplus k}$ for some $k \geq 2$.

(ii) If G is upper triangular, then $W^{[i]} \equiv W$ for all $0 \leq i \leq \ell - 1$.

Proof. Let $H = G^{-1}$. The inverse of an upper triangular matrix is also an upper triangular matrix. Therefore, it is sufficient to prove the above statements for H being upper triangular instead of G. Let the number of 1s in the first column of H be k. Clearly $W^{[0]} \equiv W^{\boxplus k}$. If $k \geq 2$ then H is not upper triangular and the first claim of the lemma holds. If $k = 1$ then $x_0 = u_0$. One can then write

$$\begin{aligned} W^{[i]}(y_0^{\ell-1}, u_0^{i-1} \,|\, u_i) &= \frac{1}{2^{\ell-1}} \sum_{u_{i+1}^{\ell-1}} W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1}) \\ &= \frac{1}{2^{\ell-1}} \sum_{x_0^{\ell-1}:(x_0^{\ell-1}H)_0^i = u_0^i} W^\ell(y_0^{\ell-1} \,|\, x_0^{\ell-1}) \\ &\overset{x_0=u_0}{=} \frac{1}{2^{\ell-1}} W(y_0 \,|\, u_0) \sum_{x_1^{\ell-1}:(x_0^{\ell-1}H)_1^i = u_1^i} W^{\ell-1}(y_1^{\ell-1} \,|\, x_1^{\ell-1}). \end{aligned}$$

Therefore, Y_0 is independent of the inputs to the channels $W^{[i]}$ for $i = 1, \ldots, \ell - 1$. If $H_{0i} = 0$ for $i \geq 1$, then u_0 does not influence the channels $W^{[i]}$ for $i \geq 1$. This is equivalent to saying that channels $W^{[1]}, \ldots, W^{[\ell-1]}$ are defined by the matrix $H^{(\ell-1)}$, where $H^{(\ell-i)}$ denotes the $(\ell - i) \times (\ell - i)$ matrix obtained from H by removing its first i rows and columns. Now consider the case of $H_{0i} \neq 0$ for some $i \geq 1$. Let $H_{01} = 1$. Then

$$W^{[1]}(y_0^{\ell-1}, u_0 \,|\, u_1) = \frac{W(y_0 \,|\, u_0)}{2^{\ell-1}} \sum_{x_1^{\ell-1}:(x_1^{\ell-1}H^{(\ell-1)})_1 = u_0 \oplus u_1} W^{\ell-1}(y_1^{\ell-1} \,|\, x_1^{\ell-1}).$$

If $u_0 = 0$, then the above channel is equal to the case of $H_{11} = 0$. If $u_0 = 1$, then channel laws for $u_1 = 0$ and $u_1 = 1$ are simply exchanged. Since both the inputs are equally likely, we are effectively using the same channel. Therefore, we can say that channels $W^{[1]}, \ldots, W^{[\ell-1]}$ are defined by the matrix $H^{(\ell-1)}$.

Applying the same argument to $H^{(\ell-1)}$ and repeating, we see that if H is upper triangular, then we have $W^{[i]} \equiv W$ for all i. On the other hand, if H is not upper triangular, then there exists an i for which $H^{(\ell-i)}$ has at least two 1s in the first column. This in turn implies that $W^{[i]} \equiv W^{\boxplus k}$ for some $k \geq 2$. □

Lemma 5.4 (I_∞ and Z_∞)**.** *If G is not upper triangular, then $\{I_n\}$ converges almost surely to a random variable I_∞ and $\{Z_n\}$ converges almost surely to a random variable Z_∞ such that*

$$I_\infty = \begin{cases} 1 & w.p.\ I(W), \\ 0 & w.p.\ 1 - I(W), \end{cases} \qquad Z_\infty = \begin{cases} 0 & w.p.\ I(W), \\ 1 & w.p.\ 1 - I(W). \end{cases}$$

Proof. The almost sure convergence of $\{I_n\}$ to a random variable I_∞ follows from Lemma 5.2. Lemma 5.3 implies that there exists an $i \in \{0, \ldots, \ell - 1\}$ and $k \geq 2$ such that for the tree process, we have

$$I(W_{n+1}) = I(W_n^{\boxplus k}) \text{ with probability at least } \frac{1}{\ell},$$

for some $k \geq 2$. Moreover by the convergence in $\mathcal{L}^1$ of I_n, we have $\mathbb{E}[|I_{n+1} - I_n|] \overset{n\to\infty}{\longrightarrow} 0$. This in turn implies

$$\mathbb{E}[|I_{n+1} - I_n|] \geq \frac{1}{\ell}\mathbb{E}[I(W_n) - (I(W_n^{\boxplus k})] \to 0. \tag{5.9}$$

It is shown in Lemma 5.37 in the Appendix that for any B-DMC W_n, if $I(W_n) \in (\delta, 1-\delta)$ for some $\delta > 0$, then there exists an $\eta(\delta) > 0$ such that $I(W_n) - I(W_n^{\boxplus k}) > \eta(\delta)$. Therefore, convergence in (5.9) implies $I_\infty \in \{0, 1\}$ w.p. 1. The claim on the probability distribution of I_∞ follows from the fact that $\{I_n\}$ is a martingale, i.e., $\Pr(I_\infty = 1) = \mathbb{E}[I_\infty] = \mathbb{E}[I_0] = I(W)$.

For any B-DMC W, Lemma 1.5 implies that when $I(W)$ takes on the value 0 or 1, $Z(W)$ takes on the value 1 or 0, respectively. This implies that $\{Z_n\}$ converges almost surely to a random variable Z_∞ with probabilities as stated. □

In Chapter 2 we show the convergence of $\{Z_n\}$ by proving that the process is a super martingale. Such a property is in general difficult to prove for arbitrary G. On the other hand, the process $\{I_n\}$ is a martingale for any invertible matrix G, which is sufficient to ensure convergence.

We have now shown that channel polarization happens whenever G is not upper triangular. For these matrices to be suitable for source coding, we still need to show that when the process $\{Z_n\}$ approaches 1, it should do so at a sufficiently fast rate. The following theorem provides a guaranty on the rate at which $\{Z_n\}$ approaches 1 for any polarizing matrix.

Theorem 5.5 (Universal Bound on Rate of Polarization)**.** *Consider a B-DMC W and an $\ell \times \ell$ matrix G. If G is not upper triangular, then for any $\beta < \frac{\log_\ell 2}{\ell}$,*

$$\lim_{n\to\infty} \Pr[Z_n \geq 1 - 2^{-\ell^{n\beta}}] = 1 - I(W).$$

Proof Idea: For any invertible matrix it can be shown that $Z_{n+1} \geq Z_n^\ell$ with probability 1. If G is not upper triangular, then Lemma 5.3 implies that there is at least one i such that $W^{[i]} \equiv W^{\boxplus k}$ for some $k \geq 2$. From (5.28), we claim that $1 - Z_{n+1} \leq (1 - Z_n^2)^2 \leq 4(1 - Z_n)^2$ with probability at least $1/\ell$. Let $X_n = 1 - Z_n$. Lemma 5.4 implies that $\{X_n\}$ almost surely converges to a random variable $\{X_\infty\}$ such that $\Pr(X_\infty = 0) = 1 - I(W)$. The proof then follows by applying Theorem 5.34 to the process $\{X_n\}$. □

Using a similar argument as in Theorem 3.4, we claim that using any $\ell \times \ell$ matrix that is not upper triangular we can achieve the symmetric rate-distortion trade-off (c.f. Theorem 3.4), with distortion D_N approaching the design distortion D as $D_N \leq D + O(2^{-(N)^\beta})$, for any $\beta < \frac{\log_\ell 2}{\ell}$.

5.3 Exponent of Source Coding

In this section we provide a characterization of the exponent of source coding. Let us first proceed with the definition.

Definition 5.6 (Exponent of Source Coding)**.** *For any B-DMC W with $0 < I(W) < 1$, we will say that an $\ell \times \ell$ matrix G has a source coding exponent $\mathtt{E}_s(G)$ if*

(i) For any fixed $\beta < \mathtt{E}_s(G)$,

$$\liminf_{n\to\infty} \Pr(Z_n \geq 1 - 2^{-\ell^{n\beta}}) = 1 - I(W).$$

(ii) For any fixed $\beta > \mathtt{E}_s(G)$,

$$\liminf_{n\to\infty} \Pr(Z_n \leq 1 - 2^{-\ell^{n\beta}}) = 1.$$

The above definition of the exponent provides a meaningful performance measure of polar codes for source coding using successive cancellation encoding. Let W be the test channel. For any rate $R > I(W)$, there exists a sequence of polar codes such that for sufficiently large N the expected distortion D_N satisfies,

$$D_N \leq D + 2^{-(N^\beta)}$$

for any $\beta < \mathtt{E}_s(G)$. Moreover, the part (ii) of the definition implies that this is the best possible exponent that we can achieve.

Theorem 5.5 implies that any matrix G that is not upper triangular is guaranteed to have an exponent of $\frac{\log_\ell 2}{\ell}$. We now proceed to find an exact characterization of $\mathtt{E}_s(G)$. The quantity $\mathtt{E}_s(G)$ refers to the rate at which $1 - Z_n$ approaches zero. It turns out that, as we will see later, it is sufficient to know how $1 - Z(W^{[i]})$ compares with $1 - Z(W)$. The following lemma provides such a relationship which is sufficient to characterize the exponent.

Lemma 5.7 (Bhattacharyya Parameter of $W^{[i]}$)**.** *Consider any B-DMC W and an $\ell \times \ell$ matrix G. Let $H = G^{-1}$ and let h_i denote the i-th column of H. Let*

$$\begin{aligned} D_0 &\triangleq d_H(h_0, 0), \\ D_i &= d_H(h_i, \langle h_0, \ldots, h_{i-1}\rangle), \qquad i = 1, \ldots, \ell - 1. \end{aligned}$$

If $D_i \leq D_{i-1}$, then

$$(1 - Z(W))^{D_i} \leq 1 - Z(W^{[i]}) \leq 2^{2i+1}(1 - Z(W))^{D_i}. \tag{5.10}$$

Proof. Let $\bar{y}, \bar{x}, \bar{u}$ denote the vectors $y_0^{\ell-1}, x_0^{\ell-1}, u_0^{\ell-1}$ respectively. Without loss of generality, we will assume that the weight of the column h_i is equal to D_i. Suppose for a matrix H this property is not satisfied. Then let $D_i = d_H(h_i, \sum_{j<i} \alpha_j h_j)$ and let $H' = [h_0, \ldots, h_{i-1}, h_i', h_{i+1}, \ldots, h_{\ell-1}]$ be a matrix such that $h_i' = h_i \oplus \sum_{j<i} \alpha_j h_j$. This implies that $H' = HA$ for an invertible upper triangular matrix A. From Lemma 5.39 we claim that $Z_{H'}(W^{[i]}) = Z_H(W^{[i]})$. Therefore, we can assume that for the matrix H the weight of the column h_i is equal to D_i.

We will first obtain an upper bound on $Z(W^{[i]})$, which will result in a lower bound for $1 - Z(W^{[i]})$. Let $W_d^{[i]}$ define the channel with input u_i and output $\bar{y}$ as follows

$$W_d^{[i]}(\bar{y} \,|\, u_i) = \frac{1}{2^{\ell-1}} \sum_{u_0^{i-1}, u_{i+1}^{\ell-1}} W_\ell(\bar{y} \,|\, \bar{u}).$$

The channel $W_d^{[i]}$ is degraded with respect to $W^{[i]} = W_\ell^{(i)}(\bar{y}, u_0^{i-1} \,|\, u_i)$, because the former channel is obtained by ignoring u_0^{i-1} from the output of the latter. Then,

$$Z(W^{[i]}) \overset{\text{Lem. } 1.8}{\leq} Z(W_d^{[i]}) = Z(W^{\boxplus D_i}) \overset{(5.29)}{\leq} 1 - (1 - Z(W))^{D_i}.$$

Therefore, $(1 - Z(W))^{D_i} \leq 1 - Z(W^{[i]})$.

Now, we proceed for the upper bound on $1 - Z(W^{[i]})$. We can express $Z(W^{[i]})$ as

$$Z(W^{[i]}) = \frac{1}{2^{\ell-1}} \sum_{\bar{y}, u_0^{i-1}} \sqrt{\sum_{(\bar{x}H)_0^i = (u_0^{i-1}, 0)} W(\bar{y} \,|\, \bar{x})} \sqrt{\sum_{(\bar{x}H)_0^i = (u_0^{i-1}, 1)} W(\bar{y} \,|\, \bar{x})}.$$

Consider a fixed u_0^{i-1}. Let $\bar{v}_0$, $\bar{v}_1$ be two vectors such that $(\bar{v}_0 H)_0^i = (u_0^{i-1}, 0)$ and $(\bar{v}_1 H)_0^i = (u_0^{i-1}, 1)$. Let $V(y \,|\, 0) = W(y \,|\, 0) + W(y \,|\, 1)$ and $V(y \,|\, 1) = W(y \,|\, 0) - W(y \,|\, 1)$ and let $\mathtt{C}$ denote the code $(\bar{x}H)_0^i = 0_0^i$. Using Lemma 5.35 with $z_i = W(y_i \,|\, 1)/W(y_i \,|\, 0)$, we get

$$\sum_{(\bar{x}H)_0^i = (u_0^{i-1}, 0)} W(\bar{y} \,|\, \bar{x}) = \frac{1}{|\mathtt{C}^\perp|} \sum_{\bar{x} \in \mathtt{C}^\perp} (-1)^{\bar{x} \cdot \bar{v}_0} \prod V(y_i \,|\, x_i)$$

where the dual code $\mathtt{C}^\perp$ is given by $\mathtt{C}^\perp = \langle h_0, \ldots, h_i \rangle$.

A similar expression with $\bar{v}_0$ replaced by $\bar{v}_1$ is obtained for the summation over $(\bar{x}H)_0^i = (u_0^{i-1}, 1)$. Note that for $0 \leq j < i$, $h_j \cdot \bar{v}_0 = h_j \cdot \bar{v}_1 = u_j$ and for $j = i$, $h_i \cdot \bar{v}_0 = 0$ and $h_i \cdot \bar{v}_1 = 1$. Therefore, for any $\bar{x} \in \mathtt{C}^\perp$, if $\bar{x} = \sum_{j=0}^i a_j h_j$ with $a_i = 0$ then $\bar{x} \cdot (\bar{v}_0 \oplus \bar{v}_1) = 0$, and if $a_i = 1$ then $\bar{x} \cdot (\bar{v}_0 \oplus \bar{v}_1) = 1$. Let T_0 and T_1 denote the terms

$$T_0(u_0^{i-1}) = \sum_{\bar{x} : \bar{x} = \sum_{j=0}^{i-1} a_j h_j} (-1)^{\bar{x} \cdot \bar{v}_0} \prod V(y_i \,|\, x_i),$$

$$T_1(u_0^{i-1}) = \sum_{\bar{x}:\bar{x}=\sum_{j=0}^{i-1} a_j h_j + h_i} (-1)^{\bar{x}\cdot\bar{v}_0} \prod V(y_i \,|\, x_i).$$

The vectors $\bar{v}_0, \bar{v}_1$ depend on u_0^{i-1}. Using this notation, we can write

$$\sum_{(\bar{x}H)_0^i=(u_0^{i-1},0)} W(\bar{y} \,|\, \bar{x}) = T_0(u_0^{i-1}) + T_1(u_0^{i-1}),$$
$$\sum_{(\bar{x}H)_0^i=(u_0^{i-1},1)} W(\bar{y} \,|\, \bar{x}) = T_0(u_0^{i-1}) - T_1(u_0^{i-1}).$$

$W(y_i \,|\, x_i) \geq 0$ implies $T_0 \geq T_1$ and $T_0 \geq 0$. The cardinality of the dual code $\mathtt{C}^\perp$ is given by $|\mathtt{C}^\perp| = 2^{i+1}$. Therefore $Z(W^{[i]})$ can be expressed as

$$\begin{aligned} Z(W^{[i]}) &= \frac{1}{2^{\ell+i}} \sum_{\bar{y},u_0^{i-1}} \sqrt{T_0(u_0^{i-1})^2 - T_1(u_0^{i-1})^2} \\ &\geq \frac{1}{2^{\ell+i}} \sum_{\bar{y},u_0^{i-1}} \prod_j V(y_j \,|\, 0) \Big(1 - \sum_{\bar{x}\in\mathtt{C}^\perp\setminus\bar{0}} \prod_j \frac{V(y_j \,|\, x_j)^2}{V(y_j \,|\, 0)^2}\Big). \end{aligned}$$

Therefore, we have

$$\begin{aligned} 1 - Z(W^{[i]}) &\leq \frac{1}{2^{\ell+i}} \sum_{\bar{y},u_0^{i-1}} \prod_j V(y_j \,|\, 0) \Big(\sum_{\bar{x}\in\mathtt{C}^\perp\setminus\bar{0}} \prod_j \frac{V(y_j \,|\, x_j)^2}{V(y_j \,|\, 0)^2}\Big) \\ &= \sum_{\bar{x}\in\mathtt{C}^\perp\setminus\bar{0}} \sum_{\bar{y}} \prod_j \frac{V(y_j \,|\, 0)}{2} \Big(\frac{V(y_j \,|\, x_j)}{V(y_j \,|\, 0)}\Big)^2. \end{aligned}$$

For any vector $\bar{x}$ of weight w we have

$$\begin{aligned} \sum_{\bar{y}} \prod_j \frac{V(y_j \,|\, 0)}{2} \Big(\frac{V(y_j \,|\, x_j)}{V(y_j \,|\, 0)}\Big)^2 &= \sum_{\bar{y}} \prod_{j:x_j=1} \frac{V(y_j \,|\, 0)}{2} \Big(1 - \frac{4W(y_j \,|\, 0)W(y_j \,|\, 1)}{V(y_j \,|\, 0)^2}\Big) \\ = \prod_{j:x_j=1} \sum_{y_j} \frac{V(y_j \,|\, 0)}{2} \Big(1 - \Big(\frac{\sqrt{W(y_j \,|\, 0)W(y_j \,|\, 1)}}{V(y_j \,|\, 0)/2}\Big)^2\Big) &\leq (1 - Z(W)^2)^w, \end{aligned}$$

where the last step follows from Jensen's inequality and $\sum_{y_i} V(y_i \,|\, 0) = 2$. The minimum distance of $\mathtt{C}^\perp$ is given by

$$\begin{aligned} \mathtt{dmin}(\langle h_0, \ldots, h_i\rangle) &= \min(\mathtt{dmin}(h_i, \langle h_0, \ldots, h_{i-1}\rangle), \mathtt{dmin}(0, \langle h_0, \ldots, h_{i-1}\rangle)) \\ &= \min(D_i, \mathtt{dmin}(\langle h_0, \ldots, h_{i-1}\rangle)) = \cdots = \min_{j\in\{0,\ldots,i\}} \{D_j\} = D_i. \end{aligned}$$

Therefore for $\bar{x} \in \mathtt{C}^\perp\setminus\bar{0}$ the weight of $\bar{x}$ is at least D_i, which implies

$$1 - Z(W^{[i]}) \leq 2^{i+1}(1 - Z(W)^2)^{D_i} \leq 2^{2i+1}(1 - Z(W))^{D_i}.$$

□

The condition $D_i \leq D_{i+1}$ is an artifact of our proof technique and we believe that the result must hold even without such a condition. We will later see that "good" matrices always satisfy such a property and hence the condition is not restrictive for finding matrices with large exponent. Combining this lemma with Theorem 5.34, we characterize the exponent as shown below.

Theorem 5.8 (Exponent of Source Coding)**.** *Consider any B-DMC W and an $\ell \times \ell$ matrix G. Let $H = G^{-1}$ and let h_i denote the i-th column of H. Let*

$$\begin{aligned} D_0 &\triangleq d_H(h_0, 0), \\ D_i &= d_H(h_i, \langle h_0, \ldots, h_{i-1} \rangle), \qquad i = 1, \ldots, \ell - 1. \end{aligned}$$

If $D_i \leq D_{i-1}$, then

$$\mathtt{E}_s(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} \log_\ell D_i. \tag{5.11}$$

Proof Idea: If G is upper triangular, from Lemma 5.3 we know that for such matrices polarization does not take place and the resulting channels are the same as the original channels W. Therefore, Definition 5.6 implies that $\mathtt{E}_s(G) = 0$. For such matrices we indeed have $D_i = 1$ for all i and (5.11) implies also that $\mathtt{E}_s(G) = 0$.

Let $X_n = 1 - Z_n$. If G is not upper triangular, the process $\{X_n\}$ almost surely converges to a random variable $\{X_\infty\}$ such that $\Pr(X_\infty = 0) = 1 - I(W)$. The proof of the exponent then follows from Lemma 5.7 and Theorem 5.34. □

5.4 Exponent of Channel Coding

In this section we discuss our results for the exponent of channel coding. These results are based on joint work in [48]. For the proofs of these results we refer the interested reader to [48]. The exponent in the context of channel coding is defined as follows.

Definition 5.9 (Exponent of Channel Coding)**.** *For any B-DMC W with $0 < I(W) < 1$, we will say that an $\ell \times \ell$ matrix G has a channel coding exponent $\mathtt{E}_c(G)$ if*

(i) For any fixed $\beta < \mathtt{E}_c(G)$,

$$\liminf_{n \to \infty} \Pr(Z_n \leq 2^{-\ell^{n\beta}}) = I(W).$$

(ii) For any fixed $\beta > \mathtt{E}_c(G)$,

$$\liminf_{n \to \infty} \Pr(Z_n \geq 2^{-\ell^{n\beta}}) = 1.$$

The definition of $\mathtt{E}_c(G)$ implies the following. Consider polar code construction using a matrix G with exponent $\mathtt{E}_c(G)$. For any B-DMC W, rate $0 < R < I(W)$ and $\beta < \mathtt{E}(G)$ using Lemma 2.9 the block error probability under successive cancellation decoding is bounded as

$$2^{-\ell^{n\beta_l}} \leq P_B \leq 2^{-\ell^{n\beta_u}}.$$

for any $\beta_l > \mathtt{E}_c(G)$ and $\beta_u \leq \mathtt{E}_c(G)$.

Similar to the source coding case, for any matrix G that is not upper triangular we can show that $\mathtt{E}_c(G) \geq \frac{\log_\ell 2}{\ell}$ [48]. We skip those details here and move on to the exact characterization of the exponent.

The channel coding exponent $\mathtt{E}_c(G)$ refers to the rate at which $\{Z_n\}$ approaches zero. Therefore, in this context it is relevant to know how $Z(W^{[i]})$ compares with $Z(W)$.

Lemma 5.10 (Bhattacharyya Parameter of $Z^{[i]}$)**.** *Consider a B-DMC W and any $\ell \times \ell$ matrix G. Let $G = \begin{bmatrix} g_0 \\ \vdots \\ g_{\ell-1} \end{bmatrix}$ and*

$$\begin{aligned} D_i &= d_H(g_i, \langle g_{i+1}, \ldots, g_{\ell-1} \rangle), \qquad i = 0, \ldots, \ell-2, \\ D_{\ell-1} &\triangleq d_H(g_{\ell-1}, 0). \end{aligned}$$

Then

$$Z(W)^{D_i} \leq Z^{(i)} \leq 2^{\ell-i} Z(W)^{D_i}. \tag{5.12}$$

Note that unlike in source coding, we do not require any ordering of the distances D_i. Combining Lemma 5.10 with Theorem 5.34 we get the following result.

Theorem 5.11 (Exponent of Channel Coding)**.** *Consider a B-DMC W and any $\ell \times \ell$ matrix G. Let $\{D_i\}$ be as defined in the previous lemma. Then*

$$\mathtt{E}_c(G) = \frac{1}{\ell} \sum_{i=0}^{\ell-1} \log_\ell D_i. \tag{5.13}$$

5.5 Duality of Exponents

In this section we show how to transform a matrix with good channel coding exponent into a matrix with good source coding exponent. We define the exponent of a matrix G, denoted as $\mathtt{E}(G)$, and show that finding good matrices for both source and channel coding problems is equivalent to finding matrices with large $\mathtt{E}(G)$. Let us first proceed with some definitions.

Definition 5.12 (Partial Distances)**.** *Given an $\ell \times \ell$ matrix $G = \begin{bmatrix} g_1 \\ \vdots \\ g_\ell \end{bmatrix}$, we define the partial distances D_i, for $i = 1, \ldots, \ell$ as*

$$D_i \triangleq d_H(g_i, \langle g_{i+1}, \ldots, g_\ell \rangle), \qquad i = 1, \ldots, \ell - 1,$$
$$D_\ell \triangleq d_H(g_\ell, 0).$$

Definition 5.13 (Exponent of a Matrix)**.** *Given an $\ell \times \ell$ matrix G with partial distances $\{D_i\}_{i=1}^{\ell}$, we define the exponent of G, denoted by $\mathtt{E}(G)$, as*

$$\mathtt{E}(G) = \frac{1}{\ell} \sum_{i=1}^{\ell} \log_\ell D_i.$$

Example 5.14 (Partial Distances and Exponent)**.** *The partial distances of the matrix*

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

are $D_1 = 1, D_2 = 1$, and $D_3 = 3$. The exponent is given by

$$\mathtt{E}(F) = \frac{1}{3}(\log_3 1 + \log_3 1 + \log_3 3) = \frac{1}{3}.$$

The definition of exponent implies that $\mathtt{E}_c(G) = \mathtt{E}(G)$. Let us now define the transformation $\tilde{G}$ of a matrix G, such that $\mathtt{E}_s(\tilde{G}) = \mathtt{E}(G)$.

Definition 5.15 (Matrix Transformation for Source Coding)**.** *For an $\ell \times \ell$ matrix $G = \begin{bmatrix} g_1 \\ \vdots \\ g_\ell \end{bmatrix}$, let $\tilde{G}$ denote the matrix $\tilde{G} = \left[g_\ell^\top, \cdots, g_1^\top\right]^{-1}$.*

It is easy to check that the source coding exponent of $\tilde{G}$ is indeed equal to $\mathtt{E}(G)$. Therefore, finding a matrix G with a good channel coding exponent immediately provides a matrix with the same exponent for source coding and vice-versa. Hence, finding good matrices for the two problems is equivalent to finding matrices with large $\mathtt{E}(G)$.

Example 5.16. *For our running example, the matrix $\tilde{F}$ is given by*

$$\tilde{F} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

If $\tilde{F}$ is used as a channel combining matrix for source coding, then Lemma 5.7 implies that

$$\mathtt{E}_s(\tilde{F}) = \frac{1}{3}(\log_3 3 + \log_3 1 + \log_3 1) = \frac{1}{3}.$$

Note that the restriction in Lemma 5.7 on the distances implies that the above relationship between the exponents is valid only when the partial distances satisfy $D_i \leq D_{i+1}$. As shown in Lemma 5.18 this restriction is not crucial because the best matrices always satisfy this property. In the following, at least for the BEC, we show that the restriction can be removed.

For the rest of this section let W be a BEC. Using a duality property for BEC we show that $\mathtt{E}_c(G) = \mathtt{E}_s(\tilde{G})$ for all G. Since the exponent of channel coding is characterized without any restriction on the ordering of the distances $\{D_i\}$, it implies that, for the BEC the restriction on $\{D_i\}$ in Lemma 5.7 can be removed. Moreover, note that in all our results the exponent is related only to the matrix G and is independent of the channel. This observation strongly suggests that even for other channels the restriction on the partial distances can be omitted.

Lemma 5.17 (Duality for the BEC)**.** *Let $Z_N^{(i)}(\epsilon) \triangleq Z(W_N^{(i)})$ when W is a BEC(ϵ) and the channel transform matrix is G. Similarly, let $\tilde{Z}_N^{(i)}(\epsilon) \triangleq Z(W_N^{(i)})$, when the channel transform matrix is $\tilde{G}$. Then*

$$Z_N^{(i)}(\epsilon) + \tilde{Z}_N^{(N-i-1)}(1-\epsilon) = 1.$$

Proof. For a matrix $M = [m_1, \ldots, m_k]$, let M^{F} denote the matrix obtained by flipping the columns i.e., $M^F = [m_k, \ldots, m_1]$. Let $\bar{Y}$ be the output of $\bar{X} = \bar{U}(G^{\otimes n})$ through the BEC(ϵ) and similarly let $\bar{Y}'$ be the output of $\bar{X} = \bar{U}\tilde{G}^{\otimes n}$ through the BEC($1-\epsilon$). Note that $Z_N^{(i)}(\epsilon)$ corresponds to the erasure probability of U_i when $(\bar{Y}, U_0^{i-1})$ is known at the decoder. Let g_j denote the j-th row of $G^{\otimes n}$. Therefore, $Z_N^{(i)}(\epsilon) = \Pr(U_i = * \,|\, \bar{Y})$ where $\bar{Y}$ is the output of $\bar{X} \in \langle g_i, g_{i+1}, \ldots, g_{N-1} \rangle$ through the BEC(ϵ).

Note that $\tilde{G}^{\otimes n} = (((G^\top)^{\mathrm{F}})^{-1})^{\otimes n} = (((G^{\otimes n})^\top)^{\mathrm{F}})^{-1}$. Therefore $\bar{X} = \bar{U}\tilde{G}^{\otimes n}$ is equivalent to $\bar{X}((G^{\otimes n})^\top)^{\mathrm{F}} = \bar{U}$. Hence

$$(\bar{X}((G^{\otimes n})^\top)^{\mathrm{F}})_j = \bar{X} g_{N-1-j}^\top = U_j.$$

Note that $\tilde{Z}_N^{(N-i-1)}(1-\epsilon)$ corresponds to the erasure probability of U_{N-i-1} when $(\bar{Y}', U_0^{N-i-2})$ is known to the decoder. Since BEC(ϵ) is a symmetric B-DMC, without loss of generality we can assume that $U_0^{N-i-2} = 0$. Therefore, $\tilde{Z}_N^{(N-i-1)}(1-\epsilon) = \Pr(\bar{X} g_i^\top = * \,|\, \bar{Y}')$ where $\bar{Y}'$ is the output of

$$\begin{aligned}\bar{X} &\in \{\bar{x} : \bar{x} g_{N-1-j}^\top = 0, \forall 0 \leq j \leq N-i-2\} \\ &= \{\bar{x} : \bar{x} g_j^\top = 0, \forall i+1 \leq j \leq N-1\}\end{aligned}$$

through the BEC($1-\epsilon$). The claim follows from Lemma 5.41. □

The above lemma implies that for the BEC

$$\Pr(Z_N^{(i)}(\epsilon) \leq 2^{-(N)^\beta}) = \Pr(1 - \tilde{Z}_N^{(i)}(1-\epsilon) \leq 2^{-(N)^\beta})$$

which further implies that $\mathtt{E}_c(G) = \mathtt{E}_s(\tilde{G})$.

5.6 Bounds on Exponent

For the matrix G_2, we have $\mathtt{E}(G_2) = \frac{1}{2}$. Note that for the case of 2×2 matrices, the only polarizing matrix is G_2. In order to address the question of whether the exponent can be improved by considering large matrices, we define

$$\mathtt{E}_\ell \triangleq \max_{G \in \{0,1\}^{\ell \times \ell}} \mathtt{E}(G). \tag{5.14}$$

The maximization problem in (5.14) is not feasible in practice even for moderate ℓ, say $\ell \geq 10$. The following lemma allows to restrict this maximization to a smaller set of matrices. Even though the maximization problem still remains intractable, by working on this restricted set, we obtain lower and upper bounds on $\mathtt{E}_\ell$.

Lemma 5.18 (Partial Distances Should Decrease)**.** *Let $G = [g_1^\top \dots g_\ell^\top]^\top$. Fix $k \in \{1, \dots, \ell\}$ and let $G' = [g_1^\top \dots g_{k+1}^\top g_k^\top \dots g_\ell^\top]^\top$ be the matrix obtained from G by swapping g_k and g_{k+1}. Let $\{D_i\}_{i=1}^\ell$ and $\{D'_i\}_{i=1}^\ell$ denote the partial distances of G and G' respectively. If $D_k > D_{k+1}$, then*

(i) $\mathtt{E}(G') \geq \mathtt{E}(G)$,

(ii) $D'_{k+1} > D'_k$.

Proof. Note first that $D_i = D'_i$ if $i \notin \{k, k+1\}$. Therefore, to prove the first claim, it suffices to show that $D'_k D'_{k+1} \geq D_k D_{k+1}$. To that end, write

$$\begin{aligned} D'_k &= d_H(g_{k+1}, \langle g_k, g_{k+2}, \dots, g_\ell \rangle), \\ D_k &= d_H(g_k, \langle g_{k+1}, \dots, g_\ell \rangle), \\ D'_{k+1} &= d_H(g_k, \langle g_{k+2}, \dots, g_\ell \rangle), \\ D_{k+1} &= d_H(g_{k+1}, \langle g_{k+2}, \dots, g_\ell \rangle), \end{aligned}$$

and observe that $D'_{k+1} \geq D_k$ since $\langle g_{k+2}, \dots, g_\ell \rangle$ is a sub-code of $\langle g_{k+1}, \dots, g_\ell \rangle$. D'_k can be computed as

$$\begin{aligned} &\min \left\{ \min_{c \in \langle g_{k+2}, \dots, g_\ell \rangle} d_H(g_{k+1}, c), \min_{c \in \langle g_{k+2}, \dots, g_\ell \rangle} d_H(g_{k+1}, c + g_k) \right\} \\ &= \min\{D_{k+1}, \min_{c \in \langle g_{k+2}, \dots, g_\ell \rangle} d_H(g_k, c + g_{k+1})\} \\ &= D_{k+1}, \end{aligned}$$

where the last equality follows from

$$\begin{aligned} \min_{c \in \langle g_{k+2}, \dots, g_\ell \rangle} d_H(g_k, c + g_{k+1}) &\geq \min_{c \in \langle g_{k+1}, g_{k+2}, \dots, g_\ell \rangle} d_H(g_k, c) \\ &= D_k > D_{k+1}. \end{aligned}$$

Therefore, $D'_k D'_{k+1} \geq D_k D_{k+1}$, which proves the first claim. The second claim follows from the inequality $D'_{k+1} \geq D_k > D_{k+1} = D'_k$. □

Corollary 5.19. *In the definition of $\mathtt{E}_\ell$ (5.14), the maximization can be restricted to the matrices G which satisfy $D_1 \leq D_2 \leq \dots \leq D_\ell$.*

5.6.1 Lower Bound

The following lemma provides a lower bound on $\mathtt{E}_\ell$ by using a Gilbert-Varshamov type construction.

Lemma 5.20 (Gilbert-Varshamov Bound)**.**

$$\mathtt{E}_\ell \geq \frac{1}{\ell} \sum_{i=1}^{\ell} \log_\ell \tilde{D}_i$$

where

$$\tilde{D}_i = \max \left\{ D : \sum_{j=0}^{D-1} \binom{\ell}{j} < 2^i \right\}. \tag{5.15}$$

Proof. We will construct a matrix $G = [g_1^\top, \ldots, g_\ell^\top]^\top$, with partial distances $D_i = \tilde{D}_i$. Let $S(c, d)$ denote the set of binary vectors with Hamming distance at most d from $c \in \{0,1\}^\ell$, i.e.,

$$S(c, d) = \{x \in \{0,1\}^\ell : d_H(x, c) \leq d\}.$$

To construct the i^{th} row of G with partial distance $\tilde{D}_i$, we find a $v \in \{0,1\}^\ell$ satisfying $d_H(v, \langle g_{i+1}, \ldots, g_\ell \rangle) = \tilde{D}_i$ and set $g_i = v$. Such a v satisfies $v \notin S(c, \tilde{D}_i - 1)$ for all $c \in \langle g_{i+1}, \ldots, g_\ell \rangle$ and exists if the sets $S(c, \tilde{D}_i - 1)$, $c \in \langle g_{i+1}, \ldots, g_\ell \rangle$ do not cover $\{0,1\}^\ell$. The latter condition is satisfied if

$$\begin{aligned} |\cup_{c \in \langle g_{i+1}, \ldots, g_\ell \rangle} S(c, \tilde{D}_i - 1)| &\leq \sum_{c \in \langle g_{i+1}, \ldots, g_\ell \rangle} |S(c, \tilde{D}_i - 1)| \\ &= 2^{\ell - i} \sum_{j=0}^{\tilde{D}_i - 1} \binom{\ell}{j} < 2^\ell. \end{aligned}$$

This is guaranteed by (5.15). □

The solid line in Figure 5.3 shows the lower bound of Lemma 5.20. The bound exceeds $\frac{1}{2}$ for $\ell = 85$, suggesting that the exponent can be improved by considering large matrices. In fact, the lower bound tends to 1 when ℓ tends to infinity.

Lemma 5.21 (Exponent 1 is Achievable)**.** $\lim_{\ell \to \infty} \mathtt{E}_\ell = 1$.

Proof. Fix $\alpha \in (0, \frac{1}{2})$. Let $\tilde{D}_{\lceil \alpha \ell \rceil}$ denote the $\lceil \alpha \ell \rceil$-th distance as defined in Lemma 5.20. Using Stirling's approximation for (5.15) it can be easily shown that

$$\lim_{\ell \to \infty} \tilde{D}_{\lceil \alpha \ell \rceil} \geq \ell h_2^{-1}(\alpha),$$

where $h_2(\cdot)$ is the binary entropy function. This is in fact the well-known asymptotic version of the Gilbert-Varshamov bound.

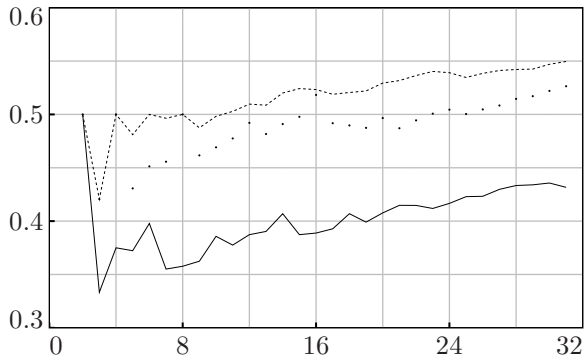

Figure 5.3: The solid curve shows the lower bound on $\mathtt{E}_\ell$ as described by Lemma 5.20. The dashed curve corresponds to the upper bound on $\mathtt{E}_\ell$ according to Lemma 5.26. The points show the performance of the best matrices obtained by the procedure described in Section 5.7.

Therefore, there exists an $\ell_0(\alpha) < \infty$ such that for all $\ell \geq \ell_0(\alpha)$ we have $\tilde{D}_{\lceil \alpha\ell \rceil} \geq \frac{1}{2}\ell h_2^{-1}(\alpha)$. Hence, for $\ell \geq \ell_0(\alpha)$ we can write

$$\begin{aligned}
\mathtt{E}_\ell &\geq \frac{1}{\ell} \sum_{i=\lceil \alpha\ell \rceil}^{\ell} \log_\ell \tilde{D}_i \\
&\geq \frac{1}{\ell}(1-\alpha)\ell \log_\ell \tilde{D}_{\lceil \alpha\ell \rceil} \\
&\geq \frac{1}{\ell}(1-\alpha)\ell \log_\ell \frac{\ell h_2^{-1}(\alpha)}{2} \\
&= 1 - \alpha + (1-\alpha)\log_\ell \frac{h_2^{-1}(\alpha)}{2},
\end{aligned}$$

where the first inequality follows from Lemma 5.20, and the second inequality follows from the fact that $\tilde{D}_i \leq \tilde{D}_{i+1}$ for all i. Therefore we obtain

$$\liminf_{\ell\to\infty} \mathtt{E}_\ell \geq 1 - \alpha \quad \forall \alpha \in (0, \frac{1}{2}). \tag{5.16}$$

Also, since $\tilde{D}_i \leq \ell$ for all i, we have $\mathtt{E}_\ell \leq 1$ for all ℓ. Hence,

$$\limsup_{\ell\to\infty} \mathtt{E}_\ell \leq 1. \tag{5.17}$$

Combining (5.16) and (5.17) concludes the proof. □

5.6.2 Upper Bound

Corollary 5.19 says that for any ℓ, there exists a matrix with $D_1 \leq \cdots \leq D_\ell$ that achieves the exponent $\mathtt{E}_\ell$. Therefore, to obtain upper bounds on $\mathtt{E}_\ell$, it suffices to bound the exponent achievable by this restricted class of matrices. The partial distances of these matrices can be bounded easily as shown in the following lemma.

Lemma 5.22 (Upper Bound on Exponent)**.** *Let $d(n,k)$ denote the largest possible minimum distance of a binary code of length n and dimension k. Then,*

$$\mathtt{E}_\ell \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \log_\ell d(\ell, \ell-i+1).$$

Proof. Let G be an $\ell \times \ell$ matrix with partial distances $\{D_i\}_{i=1}^{\ell}$ such that $\mathtt{E}(G) = \mathtt{E}_\ell$. Corollary 5.19 lets us assume without loss of generality that $D_i \leq D_{i+1}$ for all i. We therefore obtain

$$D_i = \min_{j \geq i} D_j = \mathtt{dmin}(\langle g_i, \ldots, g_\ell \rangle) \leq d(\ell, \ell-i+1).$$

□

Lemma 5.22 allows us to use existing bounds on the minimum distances of binary codes to bound $\mathtt{E}_\ell$:

Example 5.23 (Sphere Packing Bound)**.** *Applying the sphere packing bound for $d(\ell, \ell-i+1)$ in Lemma 5.22, we get*

$$\mathtt{E}_\ell \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \log_\ell \tilde{D}_i, \tag{5.18}$$

where

$$\tilde{D}_i = \max\left\{ D : \sum_{j=0}^{\lfloor \frac{D-1}{2} \rfloor} \binom{\ell}{j} \leq 2^{i-1} \right\}.$$

Note that for small values of n for which $d(n,k)$ is known for all $k \leq n$, the bound in Lemma 5.22 can be evaluated exactly.

5.6.3 Improved Upper Bound

Bounds given in Section 5.6.2 relate the partial distances $\{D_i\}$ to minimum distances of linear codes, but are loose since they do not exploit the dependence among the $\{D_i\}$. In order to improve the upper bound we use the following parametrization: Consider an $\ell \times \ell$ matrix $G = [g_1^\top, \ldots, g_\ell^\top]^\top$. Let

$$T_i = \{k : g_{ik} = 1, g_{jk} = 0 \text{ for all } j > i\},$$
$$S_i = \{k : \exists j > i \text{ s.t. } g_{jk} = 1\},$$

and let $t_i = |T_i|$.

Example 5.24. *For the matrix*

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

$T_2 = \{3\}$ and $S_2 = \{1, 2\}$.

Note that the T_i are disjoint and $S_i = \cup_{j=i+1}^{\ell} T_j$. Therefore, $|S_i| = \sum_{j=i+1}^{\ell} t_i$. Denoting the restriction of g_j to the indices in S_i by g_{jS_i}, we have

$$D_i = t_i + s_i, \tag{5.19}$$

where $s_i \triangleq d_H(g_{iS_i}, \langle g_{(i+1)S_i}, \ldots, g_{\ell S_i}\rangle)$. By a similar reasoning as in the proof of Lemma 5.18, it can be shown that there exists a matrix G with

$$s_i \leq d_H(g_{jS_i}, \langle g_{(j+1)S_i}, \ldots, g_{\ell S_i}\rangle) \quad \forall i < j,$$

and

$$\mathtt{E}(G) = \mathtt{E}_\ell.$$

Therefore, for such a matrix G, we have (cf. proof of Lemma 5.22)

$$s_i \leq d(|S_i|, \ell - i + 1). \tag{5.20}$$

Using the structure of the set S_i, we can bound s_i further.

Lemma 5.25 (Bound on Sub-Distances)**.** $s_i \leq \lfloor \frac{|S_i|}{2} \rfloor$.

Proof. We will find a linear combination of $\{g_{(i+1)S_i}, \ldots, g_{\ell S_i}\}$ whose Hamming distance to g_{iS_i} is at most $\lfloor \frac{|S_i|}{2} \rfloor$. To this end define $w = \sum_{j=i+1}^{\ell} \alpha_j g_{jS_i}$, where $\alpha_j \in \{0, 1\}$. Also define $w_k = \sum_{j=i+1}^{k} \alpha_j g_{jS_i}$. Noting that the sets T_js are disjoint with $\cup_{j=i+1}^{\ell} T_j = S_i$, we have $d_H(g_{iS_i}, w) = \sum_{j=i+1}^{\ell} d_H(g_{iT_j}, w_{T_j})$.

We now claim that choosing the α_js in the order $\alpha_{i+1}, \ldots, \alpha_\ell$ by

$$\mathrm{argmin}_{\alpha_j \in \{0,1\}} d_H(g_{iT_j}, w_{j-1T_j} + \alpha_j g_{jT_j}), \tag{5.21}$$

we obtain $d_H(g_{iS_i}, w) \leq \lfloor \frac{|S_i|}{2} \rfloor$. To see this, note that by definition of the sets T_j we have $w_{T_j} = w_{jT_j}$. Also observe that by the rule (5.21) for choosing α_j, we have $d_H(g_{iT_j}, w_{jT_j}) \leq \lfloor \frac{|T_j|}{2} \rfloor$. Thus,

$$\begin{aligned} d_H(g_{iS_i}, w) &= \sum_{j=i+1}^{\ell} d_H(g_{iT_j}, w_{T_j}) \\ &= \sum_{j=i+1}^{\ell} d_H(g_{iT_j}, w_{jT_j}) \\ &\leq \sum_{j=i+1}^{\ell} \left\lfloor \frac{|T_j|}{2} \right\rfloor \leq \left\lfloor \frac{|S_i|}{2} \right\rfloor. \end{aligned}$$

□

Combining (5.19), (5.20) and Lemma 5.25, and noting that the invertibility of G implies $\sum t_i = \ell$, we obtain the following:

Lemma 5.26 (Improved Upper Bound)**.**

$$\mathtt{E}_\ell \leq \max_{\sum_{i=1}^{\ell} t_i = \ell} \frac{1}{\ell} \sum_{i=1}^{\ell} \log_\ell(t_i + s_i)$$

where

$$s_i = \min\Big\{ \lfloor \frac{1}{2} \sum_{j=i+1}^{\ell} t_j \rfloor, d\big(\sum_{j=i+1}^{\ell} t_j, \ell - i + 1 \big) \Big\}.$$

The bound given in the above lemma is plotted in Figure 5.3. It is seen that no matrix with exponent greater than $\frac{1}{2}$ can be found for $\ell \leq 10$.

In addition to providing an upper bound to $\mathtt{E}_\ell$, Lemma 5.26 narrows down the search for matrices which achieve $\mathtt{E}_\ell$. In particular, it enables us to list all sets of possible partial distances with exponents greater than $\frac{1}{2}$. For $11 \leq \ell \leq 14$, an exhaustive search for matrices with a "good" set of partial distances bounded by Lemma 5.26 (of which there are 285) shows that no matrix with exponent greater than $\frac{1}{2}$ exists.

5.7 Construction Using BCH Codes

We will now show how to construct a matrix G of dimension $\ell = 16$ with exponent exceeding $\frac{1}{2}$. In fact, we will show how to construct the best such matrix. More generally, we will show how BCH codes give rise to "good matrices." Our construction of G consists of taking an $\ell \times \ell$ binary matrix whose k last rows form a generator matrix of a k-dimensional BCH code. The partial distance D_k is then at least as large as the minimum distance of this k-dimensional code.

To describe the partial distances explicitly we make use of the spectral view of BCH codes as sub-field sub-codes of Reed-Solomon codes as described in [74]. We restrict our discussion to BCH codes of length $\ell = 2^m - 1$, $m \in \mathbb{N}$.

Fix $m \in \mathbb{N}$. Partition the set of integers $\{0, 1, \ldots, 2^m - 2\}$ into a set $\mathcal{C}$ of *chords*,

$$\mathcal{C} = \cup_{i=0}^{2^m-2} \{2^k i \mod (2^m - 1) : k \in \mathbb{N}\}.$$

Example 5.27 (Chords for $m = 5$)**.** *For $m = 5$ the list of chords is given by*

$$\begin{aligned} \mathcal{C} = \{&\{0\}, \{1, 2, 4, 8, 16\}, \{3, 6, 12, 17, 24\}, \\ &\{5, 9, 10, 18, 20\}, \{7, 14, 19, 25, 28\}, \\ &\{11, 13, 21, 22, 26\}, \{15, 23, 27, 29, 30\}\}. \end{aligned}$$

□

Let C denote the number of chords and assume that the chords are ordered according to their smallest element as in Example 5.27. Let $\mu(i)$ denote the minimal element of chord i, $1 \leq i \leq C$ and let $l(i)$ denote the number of elements in chord i. Note that by this convention $\mu(i)$ is increasing. It is well known that $1 \leq l(i) \leq m$ and that $l(i)$ must divide m.

Example 5.28 (Chords for $m = 5$)**.** *In Example 5.27 we have $C = 7$, $l(1) = 1$, $l(2) = \cdots = l(7) = 5 = m$, $\mu(1) = 0$, $\mu(2) = 1$, $\mu(3) = 3$, $\mu(4) = 5$, $\mu(5) = 7$, $\mu(6) = 11$, $\mu(7) = 15$.* □

Consider a BCH code of length ℓ and dimension $\sum_{j=k}^{C} l(j)$ for some $k \in \{1, \ldots, C\}$. It is well-known that this code has minimum distance at least $\mu(k) + 1$. Further, the generator matrix of this code is obtained by concatenating the generator matrices of two BCH codes of respective dimensions $\sum_{j=k+1}^{C} l(j)$ and $l(k)$. This being true for all $k \in \{1, \ldots, C\}$, it is easy to see that the generator matrix of the ℓ dimensional (i.e., rate 1) BCH code, which will be the basis of our construction, has the property that its last $\sum_{j=k}^{C} l(j)$ rows form the generator matrix of a BCH code with minimum distance at least $\mu(k)+1$. This translates to the following lower bound on partial distances $\{D_i\}$: Clearly, D_i is at least as large as the minimum distance of the code generated by the last $\ell - i + 1$ rows of the matrix. Therefore, if $\sum_{j=k+1}^{C} l(j) \leq \ell - i + 1 \leq \sum_{j=k}^{C} l(j)$, then

$$D_i \geq \mu(k) + 1.$$

The exponent $\mathtt{E}$ associated with these partial design distances can then be bounded as

$$\mathtt{E} \geq \frac{1}{2^m - 1} \sum_{i=1}^{C} l(i) \log_{2^m - 1}(\mu(i) + 1). \tag{5.22}$$

Example 5.29 (BCH Construction for $\ell = 31$)**.** *From the list of chords computed in Example 5.27 we obtain*

$$\mathtt{E} \geq \frac{5}{31} \log_{31}(2 \cdot 4 \cdot 6 \cdot 8 \cdot 12 \cdot 16) \approx 0.526433.$$

An explicit check of the partial distances reveals that the above inequality is in fact an equality. □

For large m, the bound in (5.22) is not convenient to work with. The asymptotic behavior of the exponent is however easy to assess by considering the following bound. Note that no $\mu(i)$ (except for $i = 1$) can be an even number since otherwise $\mu(i)/2$, being an integer, would be contained in chord i, a contradiction. It follows that for the smallest exponent all chords (except chord 1) must be of length m and that $\mu(i) = 2i + 1$. This gives rise to the bound

$$\mathtt{E} \geq \frac{1}{(2^m - 1)\log(2^m - 1)} \tag{5.23}$$

$$\cdot \left(\sum_{k=1}^{a} m \log(2k) + (2^m - 2 - am) \log(2a+2) \right),$$

where $a = \lfloor \frac{2^m-2}{m} \rfloor$.

Lemma 5.30 (Exponent of BCH Matrices)**.** *Let $\ell = 2^m - 1$. Let $\mathtt{E}_{BCH_\ell}$ denote the exponent achievable using BCH codes. Then*

$$\lim_{\ell \to \infty} \mathtt{E}_{BCH_\ell} = 1.$$

The proof follows easily by using Stirling's approximation for (5.23). This is in fact the best exponent one can hope for (cf. Lemma 5.21). We have also seen in Example 5.29 that for $m = 5$ we achieve an exponent strictly above $\frac{1}{2}$.

Binary BCH codes exist for lengths of the form $2^m - 1$. To construct matrices of other lengths, we use *shortening*, a standard method to construct good codes of smaller lengths from an existing code, which we recall here: Given a code $\mathtt{C}$, fix a symbol, say the first one, and divide the codewords into two sets of equal size depending on whether the first symbol is a 1 or a 0. Choose the set having zero in the first symbol and delete this symbol. The resulting codewords form a linear code with both the length and dimension decreased by one. The minimum distance of the resulting code is at least as large as the initial distance. The generator matrix of the resulting code can be obtained from the original generator matrix by removing a generator vector having a one in the first symbol, adding this vector to all the remaining vectors starting with a one and removing the first column.

Now consider an $\ell \times \ell$ matrix G_ℓ. Find the column j with the longest run of zeros at the bottom, and let i be the last row with a 1 in this column. Then add the ith row to all the rows with a 1 in the jth column. Finally, remove the ith row and the jth column to obtain an $(\ell - 1) \times (\ell - 1)$ matrix $G_{\ell-1}$. The matrix $G_{\ell-1}$ satisfies the following property.

Lemma 5.31 (Partial Distances after Shortening)**.** *Let the partial distances of G_ℓ be given by $\{D_1 \leq \cdots \leq D_\ell\}$. Let $G_{\ell-1}$ be the resulting matrix obtained by applying the above shortening procedure with the ith row and the jth column. Let the partial distances of $G_{\ell-1}$ be $\{D'_1, \ldots, D'_{\ell-1}\}$. We have*

$$D'_k \geq D_k, \quad 1 \leq k \leq i-1 \tag{5.24}$$

$$D'_k = D_{k+1}, \quad i \leq k \leq \ell - 1. \tag{5.25}$$

Proof. Let $G_\ell = [g_1^\top, \ldots, g_\ell^\top]^\top$ and $G_{\ell-1} = [g_1'^\top, \ldots, g_{\ell-1}'^\top]^\top$ For $i \leq k$, g'_k is obtained by removing the jth column of g_{k+1}. Since all these rows have a zero in the jth position their partial distances do not change, which in turn implies (5.25).

For $k \leq i$, note that the minimum distance of the code $\mathtt{C}' = \langle g'_k, \ldots, g'_{\ell-1} \rangle$ is obtained by shortening $\mathtt{C} = \langle g_k, \ldots, g_\ell \rangle$. Therefore, $D'_k \geq \mathtt{dmin}(\mathtt{C}') \geq \mathtt{dmin}(\mathtt{C}) = D_k$. □

Example 5.32 (Shortening of Code)**.** *Consider the matrix*

$$\left[\begin{array}{ccccc} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{array}\right].$$

The partial distances of this matrix are $\{1, 2, 2, 2, 4\}$*. According to our procedure, we pick the* 3*rd column since it has a run of three zeros at the bottom (which is maximal). We then add the second row to the first row (since it also has a* 1 *in the third column). Finally, deleting column* 3 *and row* 2 *we obtain the matrix*

$$\left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array}\right].$$

The partial distances of this matrix are $\{1, 2, 2, 4\}$*.* □

Example 5.33 (Construction of Code with $\ell = 16$)**.** *Starting with the* 31×31 *BCH matrix and repeatedly applying the above procedure results in the exponents listed in Table 5.1.*

ℓ	exponent	ℓ	exponent	ℓ	exponent	ℓ	exponent
31	0.52643	27	0.50836	23	0.50071	19	0.48742
30	0.52205	26	0.50470	22	0.49445	18	0.48968
29	0.51710	25	0.50040	21	0.48705	17	0.49175
28	0.51457	24	0.50445	20	0.49659	16	0.51828

Table 5.1: The best exponents achieved by shortening the BCH matrix of length 31.

The 16×16 *matrix having an exponent* 0.51828 *is*

$$\begin{bmatrix}
1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}.$$

The partial distances of this matrix are $\{16, 8, 8, 8, 8, 6, 6, 4, 4, 4, 4, 2, 2, 2, 2, 1\}$. *Using Lemma 5.26 we observe that for the* 16×16 *case there are only* 11 *other possible sets of partial distances which have a better exponent than the above matrix. An exhaustive search for matrices with such sets of partial distances confirms that no such matrix exists. Hence, the above matrix achieves the best possible exponent among all* 16×16 *matrices.* □

5.A Appendix

Let us now restate a generalization of Theorem 2.10. This generalization can be proved using techniques similar to the proof of Theorem 2.10.

Theorem 5.34 (Rate of Convergence [47])**.** *Let* $\{X_n : n \geq 0\}$ *be a positive random process satisfying*

$$X_n^{s_i} \leq X_{n+1} \leq c_i X_n^{s_i} \quad w.p.\ \tfrac{1}{\ell},$$

for all $i = 0, \ldots, \ell-1$ *and some constants* $c_i > 0, s_i \geq 0$. *Let* $X_\infty := \lim_{n\to\infty} X_n$ *exist almost surely and* $\Pr(X_\infty = 0) = P_\infty$. *Then for any* $\beta < \frac{1}{\ell}\sum_{i=0}^{\ell-1} \log_\ell s_i$,

$$\lim_{n\to\infty} \Pr(X_n < 2^{-\ell^{n\beta}}) = P_\infty, \tag{5.26}$$

and for any $\beta > \frac{1}{\ell}\sum_{i=0}^{\ell-1} \log_\ell s_i$,

$$\lim_{n\to\infty} \Pr(X_n > 2^{-\ell^{n\beta}}) = 0. \tag{5.27}$$

Lemma 5.35 (Hartman-Rudolf [75])**.** *For any subset $S \subseteq \{0,1\}^k$ let $f_S(z_1, \ldots, z_k)$ denote the function*

$$f_S(z_1, \ldots, z_k) = \sum_{\bar{x} \in S} \prod_i z_i^{x_i}.$$

Let $\hat{f}$ denote the function

$$\hat{f}_{S,\bar{v}}(z_1, \ldots, z_k) = \sum_{\bar{x} \in S} \prod_i (-1)^{x_i v_i} z_i^{x_i}.$$

Let $\mathtt{C}$ denote a linear code and let $\mathtt{C}^\perp$ denote the dual code of $\mathtt{C}$. Let $\bar{v} \in \{0,1\}^k$ be a vector. Then

$$f_{\mathtt{C}+\bar{v}}(z_1, \ldots, z_k) = \frac{1}{|\mathtt{C}^\perp|} \hat{f}_{\mathtt{C}^\perp,\bar{v}} \left(\frac{1-z_1}{1+z_1}, \ldots, \frac{1-z_k}{1+z_k} \right) \prod_{i=1}^{k} (1+z_i).$$

Proof. The proof is similar to that of the Mac-Williams identities [75]. □

Lemma 5.36 (Bounds on $Z(W^{\boxtimes k})$)**.** *Let W be a B-DMC and let $W^{\boxtimes k}$ be as defined in* (5.8)*. Then*

$$1 - Z(W^{\boxtimes k}) \leq (1 - Z(W))^k, \tag{5.28}$$

$$1 - Z(W^{\boxtimes k})^2 \geq (1 - Z(W)^2)^k. \tag{5.29}$$

Proof. Let $W_1 = W^{\boxtimes k-1}$ and $W_2 = W$. Using Lemma 2.15, we get $1 - Z(W^{\boxtimes k}) \leq (1 - Z(W^{\boxtimes k-1}))(1 - Z(W))$. The proof of (5.28) follows by repeating this argument. Using similar arguments with Lemma 3.16 we get (5.29). □

Lemma 5.37 (Mutual Information of $W^{\boxtimes k}$)**.** *For any B-DMC W, if $I(W) \in (\delta, 1-\delta)$ for some $\delta > 0$, then there exists an $\eta(\delta) > 0$ such that $I(W) - I(W^{\boxtimes k}) > \eta(\delta)$.*

The proof of Lemma 5.37 is in turn based on the following theorem.

Theorem 5.38 (Extremes of Information Combining [76, 77])**.** *Let W be a symmetric B-DMC. Also let $W_{BSC}^{\boxtimes k}$ denote the channel with transition probabilities*

$$W_{BSC}^{\boxtimes k}(y_1^k \,|\, x) = \frac{1}{2^{k-1}} \sum_{x_1^k : x_1 \oplus \cdots \oplus x_k = x} \prod_{i=1}^{k} W_{BSC(\epsilon)}(y_i \,|\, x_i),$$

where $W_{BSC(\epsilon)}$ denotes the channel law of BSC(ϵ), with $\epsilon = h_2^{-1}(1 - I(W))$. Then, $I(W^{\boxtimes k}) \leq I(W_{BSC}^{\boxtimes k})$.

Proof of Lemma 5.37: Note that for $k \geq 2$,

$$I(W^{\boxtimes k}) \leq I(W^{\boxtimes 2}) \leq I(W).$$

By Theorem 5.38, for any symmetric B-DMC W, we have $I(W^{\boxtimes 2}) \leq I(W_{\mathrm{BSC}(\epsilon)}^{\boxtimes 2})$. A simple computation shows that

$$I(W_{\mathrm{BSC}(\epsilon)}^{\boxtimes 2}) = 1 - h(1 - 2\epsilon\bar{\epsilon}).$$

We can then write

$$\begin{aligned} I(W) - I(W^{\boxtimes 2}) &\geq I(W) - I(W_{\mathrm{BSC}(\epsilon)}^{\boxtimes 2}) \\ &= h_2(1 - 2\epsilon\bar{\epsilon}) - h_2(\epsilon). \end{aligned} \tag{5.30}$$

Note that $I(W) \in (\delta, 1-\delta)$ implies $\epsilon \in (\phi(\delta), \frac{1}{2} - \phi(\delta))$ where $\phi(\delta) > 0$, which in turn implies $h(1-2\epsilon\bar{\epsilon}) - h(\epsilon) > \eta(\delta)$ for some $\eta(\delta) > 0$. The result for general B-DMC W follows by symmetrizing the channel W as done in Chapter 1 (cf. Lemma 1.4 and Lemma 1.5). Consider $W_s : \mathcal{X} \to \mathcal{Y}_s$ where $\mathcal{Y}_s = \mathcal{Y} \times \mathcal{X}$ with transition probabilities

$$W_s(y, u \,|\, x) = \frac{1}{2} W(y \,|\, u \oplus x).$$

The result follows from the fact that $I(W_s^{\boxtimes k}) = I(W^{\boxtimes k})$ and W_s is a symmetric B-DMC. □

Lemma 5.39 (Equivalent Matrices)**.** *Let W be a B-DMC and let G and G' be two matrices, such that $G' = AG$, where A is an invertible upper triangular matrix. Let $Z_G(W^{[i]})$ and $Z_{G'}(W^{[i]})$ denote the resulting Bhattacharyya parameters using the matrices G and G' respectively. Then $Z_G(W^{[i]}) = Z_{G'}(W^{[i]})$ for all $i \in \{0, \ldots, \ell - 1\}$.*

Proof. Note that an invertible upper triangular matrix has 1 on all its diagonal entries. Let a_{ij} denote the entry in the i-th row and j-th column of the matrix A. Let S_α denote the set $S_\alpha = \{u_i^{\ell-1} : u_i = a, u_{i+1}^{\ell-1} \in \{0,1\}^{\ell-i-2}\}$ and let $v_i = \sum_{j<i} a_{ji} u_j$. Then,

$$\begin{aligned} Z_{G'}(W^{[i]}) &= \frac{1}{2^{\ell-1}} \sum_{u_0^{i-1}, y_0^\ell} \sqrt{\sum_{u_i^{\ell-1} \in S_0} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} G')} \sqrt{\sum_{u_i^{\ell-1} \in S_1} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} G')} \\ &= \frac{1}{2^{\ell-1}} \sum_{u_0^{i-1}, y_0^\ell} \sqrt{\sum_{u_i^{\ell-1} \in S_0} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} AG)} \sqrt{\sum_{u_i^{\ell-1} \in S_1} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} AG)} \\ &\overset{(a)}{=} \frac{1}{2^{\ell-1}} \sum_{u_0^{i-1}, y_0^\ell} \sqrt{\sum_{u_i^{\ell-1} \in S_{0 \oplus v_i}} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} G)} \sqrt{\sum_{u_i^{\ell-1} \in S_{1 \oplus v_i}} W^\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1} G)} \\ &= Z_G(W^{[i]}). \end{aligned}$$

The equality (a) follows from the fact that A is upper triangular and invertible, i.e., $a_{ij} = 0$ for $i > j$ and $a_{ii} = 1$. This implies that $(u_0^{\ell-1} A)_0^{i-1}$ depends only on u_0^{i-1} and takes all possible values in $\{0,1\}^i$. Moreover, if $u_i^{\ell-1}$ covers the set S_α then $(u_0^{\ell-1} A)_i^{\ell-1}$ covers the set $S_{a \oplus v_i}$. □

The following theorem is a restatement of the duality relationship of the EXIT functions for the BEC.

Theorem 5.40 (Duality of EXIT Function [78])**.** *Let* $\mathtt{C}$ *denote a linear code and* $\mathtt{C}^{\perp}$ *denote its dual code. Let* $X_1^{N+1} \in \mathtt{C}$ *be transmitted through the BEC*(ϵ) *and the output be* Y_1^{N+1}*. Similarly let* $X'^{N+1}_1 \in \mathtt{C}^{\perp}$ *be transmitted through the BEC*$(1-\epsilon)$ *and the output be* Y'^{N+1}_1*. Then*

$$\Pr(X_{N+1} = * \,|\, Y_1^N) + \Pr(X'_{N+1} = * \,|\, Y'^N_1) = 1.$$

The above theorem can be re-written as follows, which is useful for proving the duality of the exponents.

Lemma 5.41 (Duality Used by Lemma 5.17)**.** *Consider a set of* N*-dimensional vectors* $\{g_1, \ldots, g_k\}$*. Let* $X_1^N \in \langle g_1, \ldots, g_k \rangle$ *be transmitted through the BEC*(ϵ) *and the output be* Y_1^N*. Let* $X'^N_1 \in \langle g_1, \ldots, g_{k-1} \rangle^{\perp}$ *be transmitted through the BEC*$(1-\epsilon)$ *and the output be* Y'^N_1*. Then*

$$\Pr(U_k = * \,|\, Y_1^N) + \Pr(X'^N_1 g_k^{\top} = * \,|\, Y'^N_1) = 1,$$

where U_k *is the coefficient of* g_k *in the codeword* X_1^N*.*

Proof. Let $\mathtt{C}$ denote the $N+1$ length code

$$\mathtt{C} = \langle \tilde{g}_1, \ldots, \tilde{g}_k \rangle,$$

where $\tilde{g}_i = (g_i, 0)$ for $1 \le i \le k-1$ and $\tilde{g}_k = (g_k, 1)$. Therefore, $X_{N+1} = U_k$, which implies

$$\Pr(U_k = * \,|\, Y_1^N) = \Pr(X_{N+1} = * \,|\, Y_1^N).$$

Now consider the dual code

$$\mathtt{C}^{\perp} = \langle \tilde{g}_i, \ldots, \tilde{g}_k \rangle^{\perp}.$$

Any codeword $X'^{N+1}_1 \in \mathtt{C}^{\perp}$, satisfies $X'^{N+1}_1 \tilde{g}_k^{\top} = 0$, i.e., $X'^N_1 g_k^{\top} + X_{N+1} = 0$. Therefore,

$$\Pr(X'^N_1 g_k^{\top} = * \,|\, Y'^N_1) = \Pr(X'_{N+1} = * \,|\, Y'^N_1).$$

The claim now follows from Theorem 5.40. □

6 Extensions and Open Questions

Let us now discuss some generalizations as well point out what we consider important open problems. In each of the following cases we do not aim for completeness but rather we highlight what we consider the essence of the problem.

We start by discussing the relationship between polar codes and RM codes in Section 6.1. Using this relationship we characterize the minimum distance of polar codes and discuss the implication of the minimum distance on the performance. In Section 6.2 we propose some approaches based on message-passing algorithms to improve the finite length performance of polar codes. We then consider the compound capacity of polar codes in Section 6.3. Sections 6.4, 6.5 and 6.6 are dedicated to open problems. Most of the results of this chapter have appeared in [79].

6.1 RM Codes as Polar Codes and Some Consequences

Polar codes constructed using the matrix G_2 can be seen as a generalization of Reed-Muller (RM) codes. Let us briefly discuss the construction of RM codes. We follow the lead of [80] in which the Kronecker product is used. RM codes are specified by the two parameters n and r; the code is denoted by $\mathrm{RM}(n, r)$. An $\mathrm{RM}(n, r)$ code has block length 2^n and rate $\frac{1}{2^n} \sum_{i=0}^{r} \binom{n}{i}$. The code is defined through its generator matrix. Compute the Kronecker product $G_2^{\otimes n}$. This gives a $2^n \times 2^n$ matrix. Label the rows of this matrix as $0, \ldots, 2^n - 1$. Let $\mathtt{wt}(i)$ denote the number of ones in the binary expansion of i. One can check that the weight of the ith row of this matrix is equal to $2^{\mathtt{wt}(i)}$. The generator matrix of the code $\mathrm{RM}(n, r)$ consists of all the rows of $G_2^{\otimes n}$ which have weight

at least 2^{n-r}. There are exactly $\sum_{i=0}^{r} \binom{n}{i}$ such rows. An equivalent way of expressing this is to say that the codewords are of the form $\bar{x} = \bar{u}G_2^{\otimes n}$, where the components u_i of $\bar{u}$ corresponding to the rows of $G_2^{\otimes n}$ of weight less than 2^{n-r} are fixed to 0 and the remaining components contain the "information."

Recall from Definition 2.7 that a polar code $\mathtt{C}_N(F, \bar{0})$ consists of codewords of the form $\bar{u}G_2^{\otimes n}$ with $u_F = \bar{0}$. In other words, the generator matrix of $\mathtt{C}_N(F, \bar{0})$ is obtained by choosing the rows $\{i \in F^c\}$ of the matrix $G_2^{\otimes n}$. The following lemma follows easily from this definition.

Lemma 6.1 (RM Code as Polar Code)**.** *The RM(n, r) code is a polar code $\mathtt{C}_N(F, \bar{0})$ with $F = \{i : \mathtt{wt}(i) < n - r\}$.*

6.1.1 Minimum Distance of Polar Code

The following lemma characterizes the minimum distance of a polar code.

Lemma 6.2 (Minimum Distance of Polar Codes)**.** *The minimum distance of a polar code $\mathtt{C}_N(F, u_F)$ is given by*

$$d_{\min} = \min_{i \in F^c} 2^{\mathtt{wt}(i)}.$$

Proof. Let $w_{\min} = \min_{i \in I} \mathtt{wt}(i)$. Clearly, $d_{\min}$ cannot be larger than the minimum weight of the rows of the generator matrix. Therefore, $d_{\min} \leq 2^{w_{\min}}$. On the other hand, by adding some extra rows to the generator matrix we cannot increase the minimum distance. In particular, add all the rows of $G_2^{\otimes n}$ with weight at least $2^{w_{\min}}$. This results in the RM$(n, n - w_{\min})$ code. It is well known that $d_{\min}(\mathrm{RM}(n, r)) = 2^{n-r}$ [80]. Therefore,

$$d_{\min} \geq d_{\min}(\mathrm{RM}(n, n - w_{\min})) = 2^{w_{\min}}.$$

□

We conclude that for any given rate R, if the information bits are picked according to their weight (RM rule), i.e., if we pick the $2^n R$ vectors of largest weight, then the resulting code has the largest possible minimum distance. The following lemma gives a bound on the best possible minimum distance for any non-zero rate.

Lemma 6.3 (Bound on the Minimum Distance)**.** *For any rate $R > 0$ and any choice of information bits, the minimum distance of a polar code of length 2^n is bounded by*

$$d_{min} \leq 2^{\frac{n}{2} + c\sqrt{n}},$$

for $n > n_o(R)$ and a constant $c = c(R)$.

Proof. Lemma 6.2 implies that $d_{\min}$ is maximized by choosing the frozen bits according to the RM rule. Order the rows according to their weights and choose the largest weight $2^n R$ rows as the rows of the generator matrix. The matrix $G_2^{\otimes n}$ has $\binom{n}{i}$ rows of weight 2^i. Therefore,

$$d_{\min} \leq 2^k : \sum_{i=k+1}^{n} \binom{n}{i} < 2^n R \leq \sum_{i=k}^{n} \binom{n}{i}.$$

For $R > \frac{1}{2}$, more than half of the rows are in the generator matrix. Therefore, there is at least one row with weight less than or equal to $2^{\lceil \frac{n}{2} \rceil}$. This proves the claim for $R > \frac{1}{2}$. Consider therefore an R in the range $(0, 1/2]$. It is well known that for $\Delta = o(\sqrt{n \log n})$

$$\binom{n}{\lceil \frac{n}{2} \rceil + \Delta} = \frac{2^n}{\sqrt{\pi \frac{n}{2}}} e^{-\frac{2\Delta^2}{n}} (1 + o(1)).$$

Let us compute the number of rows with weight in the range $[2^{\lceil \frac{n}{2} \rceil - c\sqrt{n}}, 2^{\lceil \frac{n}{2} \rceil + c\sqrt{n}}]$, where c is a strictly positive constant. Neglecting the error term in the above approximation, we get

$$\begin{aligned} \sum_{i=\lceil \frac{n}{2} \rceil - c\sqrt{n}}^{\lceil \frac{n}{2} \rceil + c\sqrt{n}} \binom{n}{i} &= 2^n \sum_{i=-c\sqrt{n}}^{c\sqrt{n}} \frac{1}{\sqrt{\pi \frac{n}{2}}} e^{-\frac{2i^2}{n}} \\ &\approx 2^n \int_{-c}^{c} \frac{1}{\sqrt{\frac{\pi}{2}}} e^{-2x^2} dx = 2^n 2 \left(1 - Q\Big(\frac{c}{\sqrt{2}}\Big)\right). \end{aligned}$$

For $R > 0$, choose c sufficiently large so that $2(1 - Q(\frac{c}{\sqrt{2}})) > 1 - R$. Therefore, if we want to choose $2^n R$ information bits, $R > 0$, at least one row of the generator matrix has weight in the range $[2^{\lceil \frac{n}{2} \rceil - c\sqrt{n}}, 2^{\lceil \frac{n}{2} \rceil + c\sqrt{n}}]$. The claim now follows from Lemma 6.2. □

Theorem 6.4 (Bound on MAP Block Error Probability)**.** *Let $R > 0$ and $\beta > \frac{1}{2}$ be fixed. For any symmetric B-DMC W, and $N = 2^n$, $n \geq n(\beta, R, W)$, the probability of error for polar coding under MAP decoding at block length N and rate R satisfies*

$$P_e(N, R) > 2^{-N^\beta}.$$

Proof. For a code with minimum distance $d_{\min}$, the block error probability is lower bounded by $2^{-K d_{\min}}$ for some positive constant K, which only depends on the channel. This is easily seen by considering a genie decoder; the genie provides the correct value of all bits except those which differ between the actually transmitted codeword and its minimum distance cousin. Lemma 6.3 implies that for any $R > 0$, for n large enough, $d_{\min} < \frac{1}{K} N^\beta$ for any $\beta > \frac{1}{2}$. Therefore $P_e(N, R) > 2^{-N^\beta}$ for any $\beta > \frac{1}{2}$. □

This combined with Theorem 2.12, implies that the SC decoder achieves performance comparable to the MAP decoder in terms of the order of the exponent.

6.1.2 Dumer's Recursive Decoding

A recursive decoding algorithm was considered by Dumer [81, 82] for decoding RM codes. On close inspection, it turns out that Dumer's recursive algorithm is similar to the SC decoder of Arıkan. Dumer applied the algorithm to RM codes and showed that decoding beyond the minimum distance is possible with high probability.

Recall from the previous section that the RM(n, r) code is a polar code $\mathtt{C}_N(F, \bar{0})$ with $F = \{i : \mathtt{wt}(i) < n - r\}$. Such a rule indeed maximizes the minimum distance of the code. Empirically we noticed that (Figure 6.5) the RM codes perform well under MAP decoding. But their performance under SC decoding is significantly worse than that of polar codes.

As noted by Dumer, the reason for the bad performance of RM codes under recursive decoding is due to a few indices that are very weakly protected. The performance can be significantly improved with only a small loss in rate by freezing the weak indices. He suggests as an open problem to find a procedure that identifies the weak indices.

Polar codes can be seen as settling this problem conclusively. The rule used for choosing the frozen set F is based on the Bhattacharyya parameters (discussed in Chapter 2). Hence, unlike the RM codes, this rule is channel dependent.

6.2 Performance under Belief Propagation

In Chapter 2 we have seen that polar codes achieve the capacity of symmetric channels under SC decoding with an error probability decaying exponentially in the square root of the blocklength (for sufficiently large blocklengths). But Theorem 2.12 does not state what lengths are needed in order to achieve the promised rapid decay in the error probability nor does it specify the involved constants. Indeed, for moderate lengths polar codes under SC decoding are not record breaking. In this section we show various ways to improve the performance of polar codes by considering belief propagation (BP) decoding. BP was already used in [83] to compare the performance of polar codes based on Arıkan's rule and RM rule. For all the simulation points in the plots the 95% confidence intervals are shown. In most cases these confidence intervals are smaller than the point size and are therefore not visible.

6.2.1 Successive Decoding as a Particular Instance of BP

For communication over a binary erasure channel (BEC) one can easily show the following.

Lemma 6.5 (SC Decoder as instance of BP Decoder for the BEC)**.** *Decoding the bit U_i with the SC decoder is equivalent to applying BP with the knowledge of $U_0, \ldots, U_{i-1}$ and all other bits unknown (and a uniform prior on them).*

We conclude that if we use a standard BP algorithm (such a decoder has access also to the information provided by the frozen bits belonging to $U_{i+1}, \ldots, U_{N-1}$) then its performance is in general superior to that of the SC decoder. Indeed, it is not hard to construct explicit examples of codewords and erasure patterns where the BP decoder succeeds but the SC decoder does not. The inclusion is hence strict. Figure 6.4 shows the simulation results for the SC, the BP and the MAP decoders when transmission takes place over the BEC. As we can see from these simulation results, the performance of the BP decoder lies roughly half way between that of the SC decoder and that of the MAP decoder.

For the BEC the *scheduling* of the individual messages is irrelevant to the performance as long as each edge is updated repeatedly until a fixed point has been reached. For general B-DMCs the performance relies heavily on the specific schedule. We found empirically that a good performance can be achieved by the following schedule. Update the messages of each of the n sections of the trellis from right to left and then from left to right and so on. Each section consists of a collection of Z shaped subgraphs. We first update the

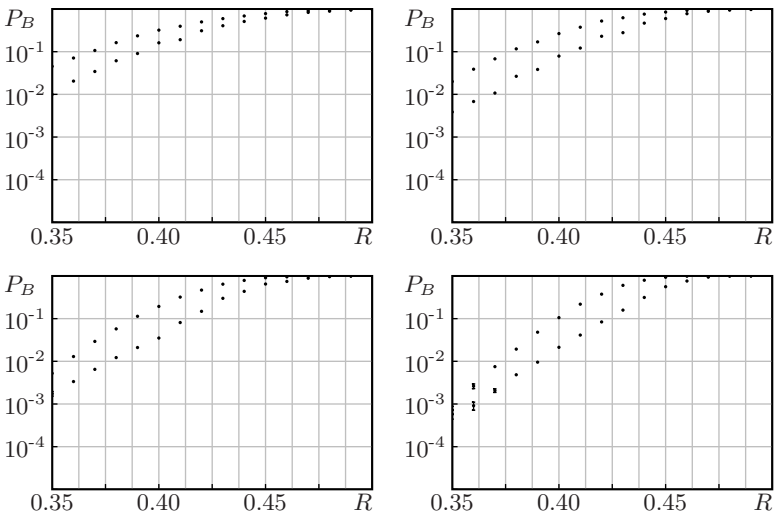

Figure 6.1: Comparison of (i) SC and (ii) BP decoder in terms of block error probability, when transmission takes place over the BAWGNC($\sigma = 0.97865$). The performance curves are shown for $n = 10$ (top left), 11 (top right), 12 (bottom left), and 13 (bottom right).

lower horizontal edge, then the diagonal edge, and, finally, the upper horizontal edge of each of these Z sections. In this schedule the information is spread from the variables belonging to one level to its neighboring level. Figure 6.1 shows the simulation results for the SC decoder and the BP decoder over the binary input additive white Gaussian noise channel (BAWGNC) of capacity $\frac{1}{2}$. Again, we can see a marked improvement of the BP decoder over the SC decoder.

6.2.2 Overcomplete Representation: Redundant Trellises

For the polar code of length 2^3 the trellis shown in Figure 2.6 is one of many possible representations. One can check that the trellises shown in Figure 6.2 also represent the same code. In fact, for a code of block length 2^n, there

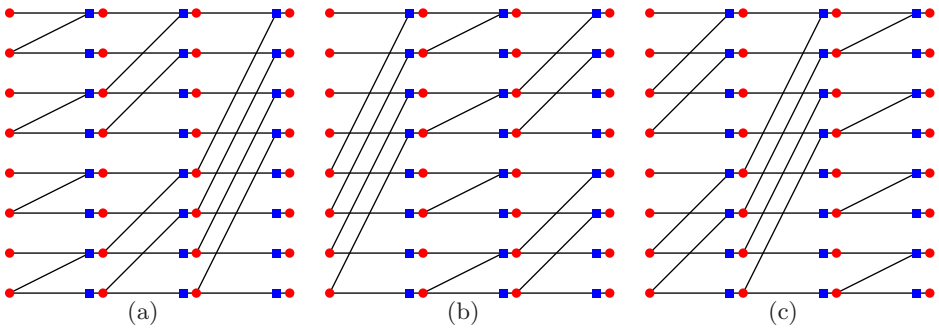

Figure 6.2: The factor graph (a) is the original factor graph shown in Figure 2.6. The factor graphs (b) and (c) are obtained by cyclic shifts of the 3 sections of the factor graph (a).

exist $n!$ different representations obtained by different permutations of the n layers of connections. Therefore, we can connect the vectors $\bar{x}$ and $\bar{u}$ with any number of these representations and this results in an *overcomplete* representation (similar to the concept used when computing the stopping redundancy of a code [84]). For example, the three trellises shown in Figure 6.2 can be connected as shown in the following figure.

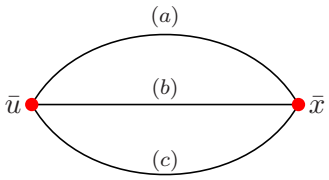

Figure 6.3: The three lines represent the three trellises on Figure 6.2.

For the BEC any such overcomplete representation only improves the performance of the BP decoder [84]. Further, the decoding complexity scales linearly with the number of different representations used. Keeping the complexity in mind, instead of considering all the $n!$ factorial trellises, we use only the n trellises obtained by cyclic shifts (e.g., see Figure 6.2). The complexity of this algorithm is $O(N(\log N)^2)$ as compared to $O(N \log N)$ if we use only one trellis. The performance of the BP decoder is improved significantly by using this overcomplete representation as shown in Figure 6.4. We leave a systematic investigation of good schedules and choices of overcomplete representations for general symmetric channels as an interesting open problem.

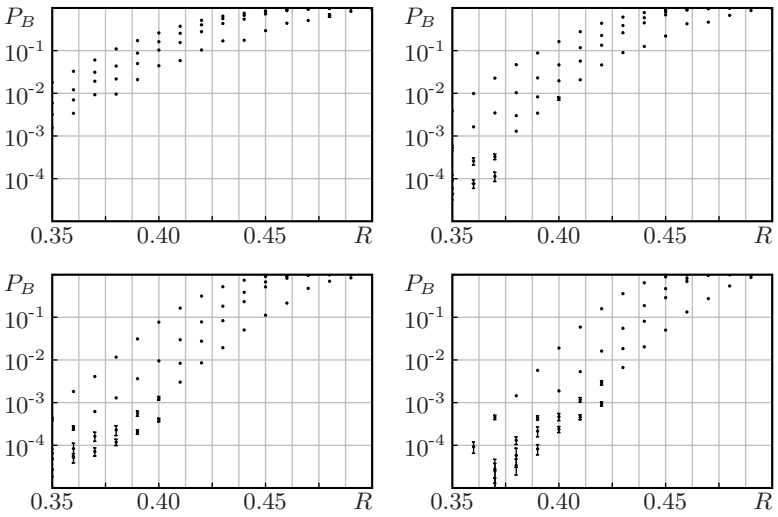

Figure 6.4: Comparison of (i) SC, (ii) BP, (iii) BP with multiple trellises, and (iv) MAP in terms of block error probability, when transmission takes place over the BEC($\frac{1}{2}$). The performance curves are shown for $n = 10$ (top left), 11 (top right), 12 (bottom left), 13 (bottom right).

6.2.3 Choice of Frozen Bits

For the BP or MAP decoding algorithm the choice of frozen bits as given by Arıkan is not necessarily optimal. In the case of MAP decoding we observe (see Figure 6.5) that the performance is significantly improved by picking the frozen bits according to the RM rule. This is not a coincidence; $d_{\min}$ is maximized for this choice. This suggests that there might be a rule for picking the frozen indices which is optimal for BP decoding. It is an interesting open question to find such a rule.

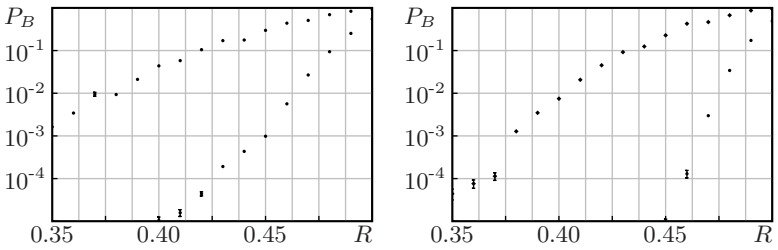

Figure 6.5: Comparison of block error probability curves under MAP decoding between codes picked according to Arıkan's rule and the RM rule. The performance curves are shown for $n = 10$ (left) and 11 (right). The RM rule performs much better than Arıkan's rule.

6.3 Compound Channel

Consider a communication scenario where the transmitter and the receiver do not know the channel. The only knowledge they have is the set of channels to which the channel belongs. Let $\mathcal{W}$ be a set of channels. The compound capacity of $\mathcal{W}$ is defined as the rate that can be reliably transmitted irrespective of the channel that is used. The compound capacity is given by [85]

$$C(\mathcal{W}) = \max_{P} \inf_{W \in \mathcal{W}} I_P(W),$$

where $I_P(W)$ denotes the mutual information between the input and out of W with the input distribution being P. Note that the compound capacity of $\mathcal{W}$ can be smaller than the infimum of the capacity of the individual channels in $\mathcal{W}$, because the capacity achieving distribution for the individual channels might be different. If the capacity achieving distribution is the same for all channels in $\mathcal{W}$, then the compound capacity is equal to the infimum of the individual capacities.

The question we are interested in is whether it is possible to achieve the compound capacity using polar codes and SC decoding. We assume that the receiver is aware of the channel. In practice this assumption is true most of the times, because the receiver can learn the channel law by using some pilot symbols at the beginning of the communication. For simplicity let $\mathcal{W}$ be a set of symmetric B-DMCs. Therefore the compound capacity of $\mathcal{W}$ is equal to the infimum of the capacities of the individual channels.

In the following we provide simple bounds which shows that polar codes cannot achieve the compound capacity using SC decoding. For further details please refer to [86].

Theorem 6.6 (Compound Capacity of Polar Codes)**.** *Let $\mathcal{W}$ denote a set of B-DMCs. Let $W_N^{(i)}$ be the channel defined in Chapter 2. Let $C_{P,SC}(\mathcal{W})$ denote the compound capacity achieved by polar codes using SC decoding. For any $N = 2^n$,*

$$C_{P,SC}(\mathcal{W}) \leq \sum_{i=0}^{N-1} \inf_{W \in \mathcal{W}} I(W_N^{(i)}),$$

$$C_{P,SC}(\mathcal{W}) \geq \sum_{i=0}^{N-1} \inf_{W \in \mathcal{W}} (1 - Z(W_N^{(i)})).$$

It can be shown that the bounds are monotonic in N. The polarization phenomenon implies that both bounds converge to the same value as $N \to \infty$. Note that if the channels in the set $\mathcal{W}$ are degraded, then the infimum is always obtained by the worst channel and hence the compound capacity is equal to the capacity of the worst channel.

Example 6.7 (Compound Capacity for BEC(0.5) and BSC(0.11002))**.** *Let $\mathcal{W} = \{BEC(0.5), BSC(0.11002)\}$. The compound capacity of $\mathcal{W}$ is $\frac{1}{2}$. Using*

Theorem 6.6, we obtain the following bounds on the compound capacity of polar codes using SC decoding.

$n=0$	1	2	3	4	5	6
0.5000	0.4818	0.4818	0.4818	0.4818	0.4817	0.4816
0.3742	0.4073	0.4266	0.4402	0.4491	0.4558	0.4609

These results suggest that the numerical value of $C_{P,SC}(BSC(0.11002), BEC(0.5))$ *is somewhere close to* 0.4816.

Similar bounds can be derived for source coding. Let us state some interesting open questions. In Chapter 5, we have considered polar codes based on general matrices G in order to improve the exponent. But perhaps this generalization is also useful in order to increase the compound capacity of polar codes. It is also interesting to inquire why polar codes are sub-optimal with respect to the compound capacity. Is this due to the codes themselves or is it a result of the sub-optimality of the decoding algorithm.

6.4 Non-Binary Polar Codes

One of the most important generalizations is the construction of non-binary polar codes. Polar codes for q-ary channels were constructed when q is either a prime or a power of prime in [70]. These codes achieve the symmetric mutual information (4.10) for the q-ary channel. Therefore, the open problem here is the construction of polar codes for any q-ary (not necessarily prime or power of prime) B-DMC.

Non-binary codes play an important role in source coding too. Here, the test channel is a q-DMC which implies that the reconstruction alphabet is q-ary. Let $W : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \{0, \ldots, q-1\}$ is the reconstruction alphabet and $\mathcal{Y}$ is the source alphabet. Let $\mathtt{d} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$ be the distortion function. The symmetric rate-distortion trade-off for q-ary sources is given by

$$R_s(D) = \min_{p(y,x):\mathbb{E}_p[\mathtt{d}(y,x)] \leq D, p(y)=P_Y(y), p(x)=\frac{1}{q}} I(Y;X). \tag{6.1}$$

The other open problem is the construction of non-binary codes for q-ary sources. For binary sources we have seen that we can achieve the symmetric rate-distortion bound and there is good reason to believe that an equivalent result holds for q-ary sources.

We have seen in Section 4.5 that achieving capacity for asymmetric channels, degraded broadcast channels as well as multiple access channels can all be mapped to achieving symmetric mutual information for q-ary channels. Using similar reasoning we can show that achieving Shannon's rate-distortion bound for asymmetric sources can be mapped to achieving the symmetric rate-distortion trade-off for non-binary sources.

6.5 Matrices with Large Exponent

One approach to improve the performance of polar codes was the generalization of Chapter 5. There we constructed polar codes using larger matrices, instead of $G_2 = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$. The performance measure that we considered was the rate of decay of the probability of error, which we referred to as the exponent. We showed that one has to consider large matrices, $\ell > 15$, to improve the exponent. We provided an explicit family of matrices based on BCH codes which achieves the best exponent for $\ell = 16$. An interesting theoretical question is to find matrices that achieve the best possible exponent for any ℓ.

Even though it is discouraging that the exponent cannot be improved by using small, lets say $\ell = 3, 4$ matrices, we should take these results with a grain of salt. The exponent is concerned with the asymptotic performance and it may not imply that a code with a larger exponent has a better performance for practical lengths. Hence, it is interesting to explore whether the exponent provides a qualitative measure at smaller lengths.

6.6 Complexity Versus Gap

We have seen that the complexity of encoding and decoding algorithms for polar codes scale as $O(N \log(N))$. How does the complexity grow as a function of the gap to the capacity or the rate-distortion bound? This is a much more subtle question.

To see what is involved in being able to answer this question, consider the Bhattacharyya constants $Z_N^{(i)} \triangleq Z(W_N^{(i)})$. Let $\tilde{Z}_N^{(i)}$ denote a re-ordering of these values in an increasing order, i.e., $\tilde{Z}_N^{(i)} \leq \tilde{Z}_N^{(i+1)}$, $i = 0, \ldots, N-2$. Define

$$m_N^{(i)} = \sum_{j=0}^{i-1} \tilde{Z}_N^{(i)}, \quad M_N^{(i)} = \sum_{j=N-i}^{N-1} \sqrt{2(1 - \tilde{Z}_N^{(i)})}.$$

For the channel coding problem we then get an upper bound on the block error probability P_N as a function the rate R of the form

$$(P_N, R) = (m_N^{(i)}, \frac{i}{N}).$$

On the other hand, for the source coding problem, we get an upper bound on the distortion D_N as a function of the rate of the form

$$(D_N, R) = (D + M_N^{(i)}, \frac{i}{N}).$$

Now, if we knew the distribution of $Z_N^{(i)}$s it would allow us to determine the rate-error probability and the rate-distortion performance achievable for this coding scheme for any given length. The complexity per bit is always $\Theta(\log N)$. Unfortunately, the computation of the quantities $m_N^{(i)}$ and $M_N^{(i)}$ is likely to be a very challenging problem.

7 Exchange of Limits

The preceding part of this thesis was concerned with polar codes. We now move on to a different topic. This is the first of two standalone chapters dealing with graphical models in communications. In this chapter, we discuss an important problem in the analysis of message passing decoders for low-density parity-check (LDPC) codes. We discuss only the main ideas behind the proofs but skip the details. For the complete proofs please refer to [87].

7.1 Introduction

Consider transmission over a binary-input memoryless output-symmetric (BMS) channel using a low-density parity-check (LDPC) code and decoding via a message-passing (MP) algorithm. We refer the reader to [51] for an introduction to the standard notation and an overview of the known results. It is well known that, for good choices of the degree distribution and the MP decoder, one can achieve rates close to the capacity of the channel with low decoding complexity [31].

The standard analysis of iterative decoding systems assumes that the block-length n is large (tending to infinity) and that a fixed number of iterations is performed. As a consequence, when decoding a given bit, the output of the decoder only depends on a fixed-size local neighborhood of this bit and this local neighborhood is tree-like. This local tree property implies that the messages arriving at nodes are conditionally independent, significantly simplifying the analysis. To determine the performance in this setting, we track the evolution of the message-densities as a function of the iteration. This process is called *density evolution* (DE). Denote the bit error probability of a code $\mathtt{G}$ after ℓ iterations by $P_b(\mathtt{G}, \epsilon, \ell)$, where ϵ is the channel parameter. Then DE computes $\lim_{N\to\infty} \mathbb{E}[P_b(\mathtt{G}, \epsilon, \ell)]$. If we now perform more and more iterations then we

get a limiting performance corresponding to

$$\lim_{\ell\to\infty}\lim_{N\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]. \tag{7.1}$$

It is known that not all computation graphs of depth ℓ can be trees unless $\ell \leq c\log(N)$, where c is a constant that only depends on the degree distribution. For a $(\mathtt{d_l},\mathtt{d_r})$-regular degree distribution pair a valid choice of c is $c(\mathtt{d_l},\mathtt{d_r}) = \frac{2}{\log(\mathtt{d_l}-1)(\mathtt{d_r}-1)}$, [88]. In practice, this condition is rarely fulfilled; standard blocklengths measure only in the hundreds or thousands but the number of iterations that have been observed to be useful in practice can easily exceed one hundred.

Consider therefore the situation where we fix the blocklength but let the number of iterations tend to infinity, i.e., we consider the limit $\lim_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$. Now take the blocklength to infinity, i.e., consider

$$\lim_{N\to\infty}\lim_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]. \tag{7.2}$$

What can we say about (7.2) and its relationship to (7.1)?

Consider the belief propagation (BP) algorithm. It was shown by McEliece, Rodemich, and Cheng [89] that one can construct specific graphs and noise realizations so that the messages on a specific edge either show a chaotic behavior or converge to limit cycles. In particular, this means that the messages do not converge as a function of the iteration. For a fixed length and a discrete channel, the number of graphs and noise realizations is finite. Therefore, if for a single graph and noise realization the messages do not converge as a function of ℓ, then it is likely that also $\lim_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$ does not exist (unless by some miracle the various non-converging parts cancel). Let us therefore consider $\limsup_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$ and $\liminf_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$. What happens if we increase the blocklength and consider $\lim_{N\to\infty}\limsup_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$ and $\lim_{N\to\infty}\liminf_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)]$?

The empirical observations strongly suggest that the exchange of limits is valid for *all* channel parameters ϵ, i.e.,

$$\lim_{N\to\infty}\limsup_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)] = \lim_{N\to\infty}\liminf_{\ell\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)] = \lim_{\ell\to\infty}\lim_{N\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell)].$$

However, in the following we limit our discussion to channel parameters ϵ for which the DE limit (7.1) goes to zero. In this case DE promises bit error probabilities that tend to zero.

Instead of considering the simple exchange of limits one can consider joint limits where the iteration is an arbitrary but increasing function of the blocklength, i.e., one can consider $\lim_{N\to\infty}\mathbb{E}[P_b(\mathtt{G},\epsilon,\ell(N))]$. Although our arguments extend to this case, for the sake of simplicity we restrict ourselves to the standard exchange of limits discussed above. In the same spirit, although some of the techniques and statements we discuss extend directly to the irregular case, in order to keep the exposition simple we restrict our discussion to

the regular ensemble LDPC($N, \mathtt{d_l}, \mathtt{d_r}$). All the difficulties encountered in the analysis are already contained in this case.

Let us proceed with a few definitions. Consider a MP algorithm with message alphabet $\mathcal{M}$. Assume that the algorithm is symmetric in the sense of [51][Definition 4.81, p. 210], so that for the purpose of analysis it is sufficient to restrict our attention to the all-one codeword assumption.

The tools we develop can be applied to a variety of MP decoders. To be concrete, we discuss below a few examples. In the following, by reliability of a message μ we mean its absolute value $|\mu|$. This means that the message $-\mu$ and μ have the same reliability.

Definition 7.1 (Bounded MS and BP Decoders)**.** *The* bounded *min-sum (MS(M)) decoder and the* bounded *belief propagation (BP(M)) decoder, both with parameter $M \in \mathbb{R}^+$, are identical to the standard min-sum and belief propagation decoder except that the reliability of the messages emitted by the check nodes is bounded to M before the messages are forwarded to the variable nodes.*

As another standard example we consider transmission over the BSC(ϵ) and decoding via the so-called *Gallager Algorithm B* (GalB).

Definition 7.2 (Gallager Algorithm B)**.** *Messages are elements of $\{\pm 1\}$. The initial messages from the variable nodes to the check nodes are the values received via the channel. The decoding process proceeds in iterations with the following processing rules:*

> *Check-Node Processing: At a check node the outgoing message along a particular edge is the product of the incoming messages along all the remaining edges.*
>
> *Variable-Node Processing: At a variable node the outgoing message along a particular edge is equal to the majority vote on the set of other incoming messages and the received value. Ties are resolved randomly.*

Before moving on to the main statement we review some results about expansion properties of random bipartite graphs that we need for our analysis. The analysis itself is split into two categories based on the degrees of the variable nodes. Section 7.3 deals with large variable degrees and Section 7.4 deals with small ones. We end the chapter by discussing some extensions (Section 7.5).

7.2 Expansion

Burshtein and Miller realized that one can use expansion arguments to show that MP algorithms have a fixed error correcting radius [90]. Although their results can be applied directly to our problem, we get stronger statements by using the expansion in a different manner.

The advantage of using expansion is that the argument applies to a wide variety of decoders and ensembles. On the negative side, the argument can only be applied to ensembles with large left degree. Why do we need large left degrees to prove the result? There are two reasons why a message emitted by a variable node can be bad (where bad in the present context means incorrect). This can be due to the received value, or it can be due to a large number of bad incoming messages. If the degree of the variable node is large then the received value plays only a minor role (think of a node of degree 1000; in this case the received value has only a limited influence on the outgoing message and this message is mostly determined by the 999 incoming messages). Suppose that the left degree is large and ignore therefore for a moment the received message. In this case large expansion helps for the following reason.

Consider a fixed iteration ℓ. Let $\mathcal{B}_\ell$ denote the set of bad variable nodes in iteration ℓ (the set of variable nodes that emit bad messages in iteration ℓ). Perform one further round of MP. In the next iteration only check nodes that are connected to $\mathcal{B}_\ell$ can send bad messages. Therefore, for a variable to belong to $\mathcal{B}_{\ell+1}$, it must be connected to a large number of bad check nodes, and hence must share many check-node neighbors with variables in $\mathcal{B}_\ell$. Suppose that $\mathcal{B}_\ell$ and $\mathcal{B}_{\ell+1}$ are sufficiently small for expansion to be valid on their union and that the graph has large expansion. Then the number of common check-node neighbors of $\mathcal{B}_\ell$ and $\mathcal{B}_{\ell+1}$ cannot be too large (since otherwise the expansion would be violated). This limits the maximum relative size of $\mathcal{B}_{\ell+1}$ with respect to $\mathcal{B}_\ell$. In other words, once $\mathcal{B}_\ell$ has reached a sufficiently small size (so that the expansion arguments can be applied), the number of errors quickly converges to zero with further iterations. In order to achieve good bounds the above argument has to be refined, but it does contain the basic idea of why large expansion helps.

On the other hand, if variable nodes have small degrees, then the received values play a dominant role and can no longer be ignored. As a consequence, for small degrees expansion arguments no longer suffice by themselves.

Definition 7.3 (Expansion)**.** *Let* $\mathtt{G}$ *be an element from LDPC*$(N, \mathtt{d_l}, \mathtt{d_r})$*.*
1) Left Expander: The graph $\mathtt{G}$ *is a* $(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$ left expander *if for every subset* $\mathcal{V}$ *of at most* αN *variable nodes, the size of the set of check nodes that are connected to* $\mathcal{V}$ *is at least* $\gamma|\mathcal{V}|\mathtt{d_l}$*.*
2) Right Expander: Let $M = N\frac{\mathtt{d_l}}{\mathtt{d_r}}$*, denote the number of check nodes. The graph* $\mathtt{G}$ *is a* $(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$ right expander *if for every subset* $\mathcal{C}$ *of at most* αM *check nodes, the size of the set of variable nodes that are connected to* $\mathcal{C}$ *is at least* $\gamma|\mathcal{C}|\mathtt{d_r}$*.*

Why are we using expansion arguments if we are interested in standard LDPC ensembles? It is well known that such codes are good expanders with high probability [90].

Theorem 7.4 (Expansion of Random Graphs [90])**.** *Let* $\mathtt{G}$ *be chosen uniformly at random from LDPC*$(n, \mathtt{d_l}, \mathtt{d_r})$*. Let* $\alpha_{\max}$ *be the positive solution of the equa-*

tion

$$\frac{\mathtt{d_l}-1}{\mathtt{d_l}}h_2(\alpha)-\frac{\mathtt{d_l}}{\mathtt{d_r}}h_2(\alpha\gamma\mathtt{d_r})-\alpha\gamma\mathtt{d_r}h_2(1/\gamma\mathtt{d_r})=0.$$

Let $\mathcal{X}(\mathtt{d_l},\mathtt{d_r},\alpha,\gamma)=\{\mathtt{G}\in \mathit{LDPC}(n,\mathtt{d_l},\mathtt{d_r}):\ \mathtt{G}\in(\mathtt{d_l},\mathtt{d_r},\alpha,\gamma)$ *left expander*$\}$. *If* $\gamma<1-\frac{1}{\mathtt{d_l}}$ *then* $\alpha_{\max}$ *is strictly positive and for* $0<\alpha<\alpha_{\max}$

$$\mathbb{P}\{\mathtt{G}\in\mathcal{X}(\mathtt{d_l},\mathtt{d_r},\alpha,\gamma)\}\geq 1-O(N^{-(\mathtt{d_l}(1-\gamma)-1)}). \tag{7.3}$$

Let $M=N\frac{\mathtt{d_l}}{\mathtt{d_r}}$. *We get the equivalent result for right expanders by exchanging the roles of* $\mathtt{d_l}$ *and* $\mathtt{d_r}$ *as well as* N *and* M.

In the proofs we also use the following concentration theorem.

Theorem 7.5 (Concentration Theorem [51, p. 222])**.** *Let* $\mathtt{G}$*, chosen uniformly at random from* $\mathit{LDPC}(N,\lambda,\rho)$*, be used for transmission over a* $\mathit{BMS}(\epsilon)$ *channel. Assume that the decoder performs* ℓ *rounds of message-passing decoding and let* $P_b^{MP}(\mathtt{G},\epsilon,\ell)$ *denote the resulting bit error probability. Then, for any given* $\delta>0$*, there exists an* $\alpha>0$*,* $\alpha=\alpha(\lambda,\rho,\delta)$*, such that*

$$\mathbb{P}\{|P_b^{MP}(\mathtt{G},\epsilon,\ell)-\mathbb{E}_{\mathit{LDPC}(N,\lambda,\rho)}\left[P_b^{MP}(\mathtt{G},\epsilon,\ell)\right]|>\delta\}\leq e^{-\alpha N}.$$

7.3 Case of Large Variable Degree

Let us now show that for codes with sufficient expansion the exchange of limits is indeed valid below the DE decoding threshold. The key to what follows is to find a proper definition of a "good" pair of message subsets.

Definition 7.6 (Good Message Subsets)**.** *For a fixed* $(\mathtt{d_l},\mathtt{d_r})$*-regular ensemble and a fixed MP decoder, let* β*,* $0<\beta\leq 1$*, be such that* $\beta(\mathtt{d_l}-1)\in\mathbb{N}$*. A "good" pair of subsets of* $\mathcal{M}$ *of "strength"* β *is a pair of subsets* $(G_{\mathtt{v}},G_{\mathtt{c}})$ *so that*

- *if at least* $\beta(\mathtt{d_l}-1)$ *of the* $(\mathtt{d_l}-1)$ *incoming messages at a variable node belong to* $G_{\mathtt{v}}$ *then the outgoing message on the remaining* edge *is in* $G_{\mathtt{c}}$
- *if all the* $(\mathtt{d_r}-1)$ *incoming messages at a check node belong to* $G_{\mathtt{c}}$ *then the outgoing message on the remaining* edge *is in* $G_{\mathtt{v}}$
- *if at least* $\beta(\mathtt{d_l}-1)+1$ *of all* $\mathtt{d_l}$ *incoming messages belong to* $G_{\mathtt{v}}$*, then the* variable *is decoded correctly*

We denote the probability of the bad message set $\mathcal{M}\backslash G_{\mathtt{v}}$ *after* ℓ *iterations of DE by* $p_{bad}^{(\ell)}$.

Theorem 7.7 (Expansion and Bit Error Probability)**.** *Consider an LDPC* $(N, \mathtt{d_l}, \mathtt{d_r})$ *ensemble, transmission over a family of BMS*(ϵ) *channels, and a symmetric MP decoder. Assume that this combination has a threshold under DE, call it* ϵ^{MP}*. Let* β *be the strength of the good message subset. If* $\beta < 1$ *and if for some* $\epsilon < \epsilon^{MP}$ *we have* $\limsup_{\ell\to\infty} p_{bad}^{(\ell)} = 0$ *then*

$$\lim_{N\to\infty} \limsup_{\ell\to\infty} \mathbb{E}_{LDPC(N,\mathtt{d_l},\mathtt{d_r})}[P_b^{MP}(\mathtt{G}, \epsilon, \ell)] = 0. \tag{7.4}$$

Proof. Here is the idea of the proof: we first run the MP algorithm for a fixed number of iterations such that the bit error probability is sufficiently small, say p. If the length N is sufficiently large then we can use DE to gauge the number of required iterations. Then, using the expansion properties of the graph, we show that the probability of error stays close to p for any number of further iterations. In particular, we show that the error probability never exceeds cp, where c is a constant, which only depends on the degree distribution and β. Since p can be chosen arbitrarily small, the claim follows.

Here is the fine print. Define

$$\gamma = \Big(1 - \frac{1}{\mathtt{d_l}}\Big)\frac{1+\beta}{2} \overset{\beta<1}{<} \Big(1 - \frac{1}{\mathtt{d_l}}\Big). \tag{7.5}$$

Let $0 < \alpha < \alpha_{\max}(\gamma)$, where $\alpha_{\max}(\gamma)$ is the function defined in Theorem 7.4. Let $p = \frac{\alpha(1-\beta)(\mathtt{d_l}-1)}{4}$ and let $\ell(p)$ be the number of iterations such that $p_{\text{bad}}^{(\ell(p))} \leq p$. Since $p_{\text{bad}}^{(\infty)} = 0$ and $p > 0$ this is possible. Let $P_{\mathtt{e}}(\mathtt{G}, \mathtt{E}, \ell)$ denote the fraction of messages belonging to the bad set for a given code $\mathtt{G}$ and noise realization $\mathtt{E}$ after ℓ iterations. Let Ω denote the space of code and noise realizations. Let $A \subseteq \Omega$ denote the subset

$$A = \{(\mathtt{G}, \mathtt{E}) \subseteq \Omega \,|\, P_{\mathtt{e}}(\mathtt{G}, \mathtt{E}, \ell(p)) \leq 2p\}. \tag{7.6}$$

From (the Concentration) Theorem 7.5 we know that

$$\mathbb{P}\{(\mathtt{G}, \mathtt{E}) \notin A\} \leq 2e^{-KNp^2} \tag{7.7}$$

for some strictly positive constant $K = K(\mathtt{d_l}, \mathtt{d_r}, p)$. In words, for most (sufficiently large) graphs and noise realizations the error probability after a fixed number of iterations behaves close to the asymptotic ensemble. We now show that once the error probability is sufficiently small it never increases substantially thereafter if the graph is an expander, regardless of how many iterations we still perform.

Let $V_0 \subseteq [n]$ be the *initial* set of bad variable nodes. More precisely, V_0 is the set of all variable nodes that are bad in the $\ell(p)$-th iteration. Consider a variable node and a fixed edge $\mathtt{e}$ connected to it: the outgoing message along $\mathtt{e}$ is determined by the received value as well as by the $(\mathtt{d_l} - 1)$ incoming messages along the other $(\mathtt{d_l} - 1)$ edges. Recall that if $\beta(\mathtt{d_l} - 1)$ of those messages are good then the outgoing message along edge $\mathtt{e}$ is good. Therefore,

if a variable node has $\beta(\mathtt{d_l}-1)+1$ good incoming messages, then *all* outgoing messages are good. We conclude that for a variable node to be bad at least $\mathtt{d_l}-\beta(\mathtt{d_l}-1)$ incoming messages must be bad. Therefore, it should connect to at least $\mathtt{d_l}-\beta(\mathtt{d_l}-1)$ bad check nodes. Recall that $2pN$ is number of bad messages in the $\ell(p)$-th iteration from check to variable nodes. Therefore,

$$|V_0| \leq \frac{2p}{\mathtt{d_l}-\beta(\mathtt{d_l}-1)}N \tag{7.8}$$

As a worst case we assume that all its outgoing edges are bad. Let the set of check nodes connected to V_0 be C_0. These are the only check nodes that potentially can send bad messages in the next iteration. Therefore, we call C_0 the initial set of *bad* check nodes. Clearly,

$$|C_0| \leq \mathtt{d_l}|V_0|. \tag{7.9}$$

We want to count the number of bad variables that are created in any of the future iterations. For convenience, once a variable becomes bad we will consider it to be bad for all future iterations. This implies that the set of bad variables is non-decreasing.

Let us now bound the number of bad variable nodes by the following process. The process proceeds in discrete steps. Let V_t, C_t denote the set of bad variable and check nodes at the beginning of time t. At step t, consider the set of variables that are not contained in V_t but that are connected to at least $\mathtt{d_l}-\beta(\mathtt{d_l}-1)$ check nodes in C_t (the set of "bad" check nodes). If at time t no such variable exists stop the process. Otherwise, choose one such variable at random and add it to V_t. This gives us the set V_{t+1}. We also add all neighbors of this variable to C_t. This gives us the set C_{t+1}. By this we are adding the variable nodes that can potentially become bad and the check nodes that can potentially send bad messages to V_t and C_t respectively. As discussed above, for a good variable to become bad it must be connected to at least $\mathtt{d_l}-\beta(\mathtt{d_l}-1)$ check nodes that are connected to bad variable nodes. Therefore, at most $\beta(\mathtt{d_l}-1)$ new check nodes are added in each step. Hence, if the process continues then

$$|V_{t+1}| = |V_t| + 1, \tag{7.10}$$
$$|C_{t+1}| \leq |C_t| + \beta(\mathtt{d_l}-1). \tag{7.11}$$

By assumption, the graph is an element of $\mathcal{X}(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$. Initially we have $|V_0| \leq \frac{2p}{\mathtt{d_l}-\beta(\mathtt{d_l}-1)}N = \frac{\alpha(\mathtt{d_l}-1)(1-\beta)}{2(\mathtt{d_l}-\beta(\mathtt{d_l}-1))}N \leq \alpha N$. Therefore, as long as $|V_t| \leq \alpha N$,

$$\gamma\mathtt{d_l}|V_t| \leq |C_t|, \tag{7.12}$$

since C_t contains all neighbors of V_t. Let T denote the stopping time of the process, i.e., the smallest time at which no new variable can be added to V_t. We will now show that the stopping time is finite. We have

$$\gamma\mathtt{d_l}(|V_0|+t) \overset{(7.10)}{=} \gamma\mathtt{d_l}|V_t| \overset{(7.12)}{\leq} |C_t| \overset{(7.11)}{\leq} |C_0| + t\beta(\mathtt{d_l}-1) \overset{(7.9)}{\leq} \mathtt{d_l}|V_0| + t\beta(\mathtt{d_l}-1).$$

Solving for t this gives us $T \leq \frac{|V_0|\mathtt{d}_\mathtt{l}(1-\gamma)}{\gamma\mathtt{d}_\mathtt{l}-\beta(\mathtt{d}_\mathtt{l}-1)}$. Therefore,

$$|V_T| \leq \frac{|V_0|\mathtt{d}_\mathtt{l}(1-\gamma)}{\gamma\mathtt{d}_\mathtt{l}-\beta(\mathtt{d}_\mathtt{l}-1)} + |V_0| \overset{(7.8)}{\leq} \frac{2p}{\gamma\mathtt{d}_\mathtt{l}-\beta(\mathtt{d}_\mathtt{l}-1)}N = \alpha N, \tag{7.13}$$

The whole derivation so far was based on the assumption that $|V_t| \leq \alpha n$ for $0 \leq t \leq T$. But as we can see from the above equation, this condition is indeed satisfied ($|V_t|$ is non-decreasing and $|V_T| \leq \alpha N$).

Putting all these things together, we get

$$\begin{aligned}\mathbb{E}[P_b^{\mathrm{MP}}(\mathtt{G},\epsilon,\ell)] =& \mathbb{E}[P_b^{\mathrm{MP}}(\mathtt{G},\mathtt{E},\ell)(\mathbb{1}_{\{(\mathtt{G},\mathtt{E})\in A\}} + \mathbb{1}_{\{(\mathtt{G},\mathtt{E})\notin A\}})] \\ \leq& \mathbb{E}[P_b^{\mathrm{MP}}(\mathtt{G},\mathtt{E},\ell)\mathbb{1}_{\{(\mathtt{G},\mathtt{E})\in A\}}\mathbb{1}_{\{\mathtt{G}\in\mathcal{X}(\mathtt{d}_\mathtt{l},\mathtt{d}_\mathtt{r},\alpha,\gamma)\}}]+ \\ & \mathbb{P}\{\mathtt{G}\notin\mathcal{X}(\mathtt{d}_\mathtt{l},\mathtt{d}_\mathtt{r},\alpha,\gamma)\} + \mathbb{P}\{(\mathtt{G},\mathtt{E})\notin A\}.\end{aligned}$$

Apply $\limsup_{\ell\to\infty}$ on both sides of the inequality. According to (7.13) the first term on the RHS is bounded by α. For the second term, since $\gamma < 1 - \frac{1}{\mathtt{d}_\mathtt{l}}$, we know from Theorem 7.4 that it is upper bounded by $O(n^{-(\mathtt{d}_\mathtt{l}(1-\gamma)-1)})$. For the third term we know from (7.7) that it is bounded by $2e^{-KNp^2}$ for some strictly positive constant $K = K(\mathtt{d}_\mathtt{l},\mathtt{d}_\mathtt{r},p)$. Therefore, if we subsequently apply the limit $\limsup_{N\to\infty}$ then we get

$$\limsup_{N\to\infty}\limsup_{\ell\to\infty}\mathbb{E}[P_b^{\mathrm{MP}}(\mathtt{G},\epsilon,\ell)] \leq \alpha.$$

Since this conclusion is valid for any $0 < \alpha \leq \alpha_{\max}$, it follows that

$$\lim_{N\to\infty}\limsup_{\ell\to\infty}\mathbb{E}[P_b^{\mathrm{MP}}(\mathtt{G},\epsilon,\ell)] = 0.$$

□

The proof idea is somewhat different from the one used in [90]. We first perform a small number of iterations to bring the error probability down to a small value. But rather than asking that the error probability decreases to zero by performing a sufficient number of further iterations, we only require that it stays small. The payoff for this less stringent requirement is that the necessary conditions are less stringent as well. The following theorem is very much in the spirit of [90].

Theorem 7.8 (Expansion and Block Error Probability)**.** *Consider an LDPC $(N,\mathtt{d}_\mathtt{l},\mathtt{d}_\mathtt{r})$ ensemble, transmission over a family of BMS(ϵ) channels, and a symmetric MP decoder. Assume that this combination has a threshold under DE, call it ϵ^{MP}. Let β be the strength of the good message subset. If $\beta < \frac{\mathtt{d}_\mathtt{l}-2}{\mathtt{d}_\mathtt{l}-1}$ and if for some $\epsilon < \epsilon^{MP}$ we have $\limsup_{\ell\to\infty} p_{bad}^{(\ell)} = 0$ then*

$$\lim_{N\to\infty}\limsup_{\ell\to\infty}\mathbb{E}_{LDPC(N,\mathtt{d}_\mathtt{l},\mathtt{d}_\mathtt{r})}[P_B^{MP}(\mathtt{G},\epsilon,\ell)] = 0. \tag{7.14}$$

As in Theorem 7.7 we first perform a fixed number of iterations to bring down the bit error probability below a desired level. We then use Theorem 7.9, a modified version of a theorem by Burshtein and Miller [90], to show that for a graph with sufficient expansion the MP algorithm decodes the whole block correctly once the bit error probability is sufficiently small.

Theorem 7.9 ([90]). *Consider a* $(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$*-left expander. Assume that* $0 \leq \beta \leq 1$*,* $\beta(l-1) \in \mathbb{N}$*, and* $\beta\frac{\mathtt{d_l}-1}{\mathtt{d_l}} \leq 2\gamma - 1$*. Let* $N_0 \leq \frac{\alpha}{\mathtt{d_l}\mathtt{d_r}}N$*. If at some iteration* ℓ *the number of bad variable nodes is less than* N_0 *then the MP algorithm will decode successfully.*

Discussion: Theorem 7.8 has a stronger implication (the block error probability tends to zero as a function of the iteration, assuming the bit error probability has reached a sufficiently small value) than Theorem 7.7 (here we are only guaranteed that the bit error probability stays small once it has reached a sufficiently small value). But it also requires a stronger condition.

Let us now apply the previous theorems to some examples.

Example 7.10 (BSC and GalB Algorithm). *For this algorithm* $\mathcal{M} = \{-1, +1\}$*. Pick* $G_{\mathtt{v}} = G_{\mathtt{c}} = \{+1\}$*. Assume that the received (via the channel) value is incorrect. In this case at least* $\lceil(\mathtt{d_l} - 1)/2\rceil + 1$ *of the* $(\mathtt{d_l} - 1)$ *incoming messages should be good to ensure that the outgoing message is good and at least* $\lceil(\mathtt{d_l} - 1)/2\rceil + 2$ *of the* $\mathtt{d_l}$ *incoming messages should be good to ensure that the variable is decoded correctly. Therefore,* $\beta = \frac{\lceil(\mathtt{d_l}-1)/2\rceil+1}{\mathtt{d_l}-1}$*. If the probability of the bad message set goes to* 0 *in the DE limit, then from Theorem 7.7 the limits can be exchanged if* $\mathtt{d_l} - 1 > 1 + \lceil(\mathtt{d_l} - 1)/2\rceil$*, i.e., for* $\mathtt{d_l} \geq 5$ *and from Theorem 7.8, the block error probability goes to zero if* $\mathtt{d_l} - 2 > 1 + \lceil(\mathtt{d_l} - 1)/2\rceil$*, i.e., for* $\mathtt{d_l} \geq 7$*.*

The key to applying expansion arguments to decoders with a continuous alphabet is to ensure that the received values are no longer dominant once DE has reached small error probabilities. This can be achieved by ensuring that the input alphabet is smaller than the message alphabet. Let us give a few examples here.

Example 7.11 (MS(5) Decoder). *Consider the* $(\mathtt{d_l} \geq 5, \mathtt{d_r})$ *ensemble and fix* $M = 5$*. Let the channel log likelihoods belong to* $[-1, 1]$*. It is easy to check that in this case we can choose* $G_{\mathtt{v}} = G_{\mathtt{c}} = [4, 5]$ *and that it has strength* $\beta \leq \frac{3}{4}$*. Therefore, if the probability of the bad message set goes to* 0 *under DE, then according to Theorem 7.7 the limits can be exchanged. If instead we consider* $(\mathtt{d_l} \geq 7, \mathtt{d_r})$ *then* $\beta \leq \frac{1}{3}$*. Hence, according to Theorem 7.8 the block error probability tends to* 0*.*

Example 7.12 (BP(10) Decoder). *Let* $\mathtt{d_l} = 5$ *and* $\mathtt{d_r} = 6$ *and fix* $M = 10$*. Let the channel log likelihoods belong to* $[-1, 1]$*. We claim that in this case the message subset pair* $G_{\mathtt{v}} = [9, 10], G_{\mathtt{c}} = [16, 41]$ *is good with strength* $\beta = \frac{3}{4}$*. This can be seen as follows: If all the incoming messages to a check node belong*

to $G_{\mathtt{c}}$, then the outgoing message is at least 14.39*, which is mapped down to* 10*. Suppose that at a variable node at least* $3(=\beta(\mathtt{d_l}-1))$ *out of the* 4 *incoming messages belong to $G_{\mathtt{v}}$. In this case the reliability of the outgoing message is at least* $16 = 3\times 9 - 10 - 1$*. The maximum reliability is* 41*. Moreover, if all the incoming messages belong to $G_{\mathtt{v}}$ then the variable is decoded correctly. Therefore if the probability of outgoing messages from check nodes being in* $[9, 10]$ *goes to* 1 *in the DE limit then from Theorem 7.7, the limits can be exchanged.*

It is clear that Theorems 7.7 and 7.8 apply to an infinite variety of decoders. But in all these cases the required variable node degrees are rather large. In the next section we discuss an alternative method which can sometimes be applied to ensembles with small variable-node degrees.

7.4 Case of Small Variable Degree

As we have mentioned before, if the left degree is small then the received value retains a large influence on emitted messages regardless of the number of iterations. In this case expansion arguments no longer suffice to prove the desired result. Although the results below can be extended to more general scenarios, we limit the subsequent discussion to the Gallager decoding algorithm B (GalB). All the complications are already present for this case.

Lemma 7.13 (Exchange of Limits for GalB and $\ell \geq 3$)**.** *Consider transmission over the BSC(ϵ) using random elements from the $(\mathtt{d_l}, \mathtt{d_r})$-regular ensemble and decoding by the GalB algorithm. If $\epsilon < \epsilon^{LGalB}$ then*

$$\lim_{N\to\infty} \limsup_{\ell\to\infty} \mathbb{E}[P_b^{GalB}(\mathtt{G}, \epsilon, \ell)] = 0,$$

where ϵ^{LGalB} is the smallest parameter ϵ for which a solution to the following fixed point equation exists in $(0, \epsilon]$.

$$x = \epsilon \sum_{k=0}^{\lfloor \frac{\mathtt{d_l}-1}{2} \rfloor} \binom{\mathtt{d_l}-1}{k} y^k (1-y)^{\mathtt{d_l}-1-k} + \bar{\epsilon} \sum_{k=\lfloor \frac{\mathtt{d_l}}{2} \rfloor + 1}^{\mathtt{d_l}-1} \binom{\mathtt{d_l}-1}{k} (1-y)^k y^{\mathtt{d_l}-1-k}$$
$$+ \frac{\mathbb{1}_{\{\frac{\mathtt{d_l}}{2}\in\mathbb{N}\}}}{2} \binom{\mathtt{d_l}-1}{\frac{\mathtt{d_l}}{2}} \left(\epsilon y^{\frac{\mathtt{d_l}}{2}} (1-y)^{\frac{\mathtt{d_l}}{2}-1} + \bar{\epsilon}(1-y)^{\frac{\mathtt{d_l}}{2}} (y)^{\frac{\mathtt{d_l}}{2}-1} \right), \tag{7.15}$$

where $y = (1-x)^{\mathtt{d_r}-1}$.

Discussion: Note that the threshold $\epsilon^{\mathrm{LGalB}}$ introduced in the preceding lemma is in general smaller than the DE threshold ϵ^{GalB}, although not much smaller for large values of $\mathtt{d_r}$ (see Tables 7.1 and 7.2). The extension of the result to channel values all the way up to the DE threshold ϵ^{GalB} is a challenging open problem.

$\mathtt{d_r}$	rate	ϵ^{Sha}	ϵ^{GalB}	$\epsilon^{\mathrm{LGalB}}$
4	0.25	≈ 0.2145	≈ 0.1068	≈ 0.0847
5	0.4	≈ 0.1461	≈ 0.06119	≈ 0.0506
6	0.5	≈ 0.11002	≈ 0.0394	≈ 0.0336
7	0.5714	≈ 0.08766	≈ 0.02751	≈ 0.02398
8	0.625	≈ 0.07245	≈ 0.02027	≈ 0.01795
9	0.667	≈ 0.06141	≈ 0.01554	≈ 0.01395
10	0.7	≈ 0.05324	≈ 0.01229	≈ 0.01115

Table 7.1: Threshold values for some degree distributions with $\mathtt{d_l} = 3$.

$\mathtt{d_r}$	rate	ϵ^{Sha}	ϵ^{GalB}	$\epsilon^{\mathrm{LGalB}}$
5	0.2	≈ 0.1461	≈ 0.0464	≈ 0.0399
6	0.333	≈ 0.11002	≈ 0.0292	≈ 0.0258
7	0.4286	≈ 0.08766	≈ 0.0200	≈ 0.018
8	0.5	≈ 0.07245	≈ 0.0146	≈ 0.0133
9	0.556	≈ 0.06141	≈ 0.0111	≈ 0.0102
10	0.6	≈ 0.05324	≈ 0.0087	≈ 0.0081

Table 7.2: Threshold values for some degree distributions with $\mathtt{d_l} = 4$.

In what follows we mainly discuss the case of the GalB algorithm and $\mathtt{d_l} = 3$. Fix $0 \leq \epsilon < \epsilon^{\mathrm{LGalB}}$. We prove that for every $\alpha > 0$ there exists an $N(\alpha, \epsilon)$ so that $\limsup_{\ell \to \infty} \mathbb{E}[P_b^{\mathrm{GalB}}(\mathtt{G}, \epsilon, \ell)] < \alpha$ for $N \geq N(\alpha, \epsilon)$. The proof proceeds in two phases akin to the proof for large degrees. However, the ideas are quite different because as mentioned before, expansion arguments by themselves do not work for small degrees.

We proceed by a sequence of simplifications for the algorithm, ensuring in each step that the modified algorithm is an upper bound on the original algorithm.

Linearized Gal B

Without loss of generality we can assume that the all-one codeword was sent. Therefore, the message 1 signifies in the sequel a *correct* message, whereas -1 implies that the message is *incorrect*. The analysis is simplified considerably by *linearizing* the decoding algorithm in the following way.

Definition 7.14 (Linearized GalB)**.** *The* linearized *GalB decoder, denoted by LGalB, is defined as follows: at the variable node the computation rule is same as that of the GalB decoder. At the check node the outgoing message is the minimum of the incoming messages.*

Discussion: The LGalB is not a practical decoding algorithm but rather a convenient device for analysis; it is understood that we assume that the all-

one codeword was transmitted and that quantities like the error probability refer to the variables decoded as -1. With some abuse of terminology, we nevertheless refer to it as a decoder.

The LGalB decoder is monotone also with respect to the incoming messages at check nodes. Moreover, it satisfies the following property.

Lemma 7.15 (LGalB is Upper Bound on GalB)**.** *For any graph* $\mathtt{G}$*, any noise realization* $\mathtt{E}$*, any starting set of "bad" edges, and any* ℓ*, we have* $P_{\mathsf{e}}^{GalB}(\mathtt{G},\mathtt{E},\ell) \leq P_{\mathsf{e}}^{LGalB}(\mathtt{G},\mathtt{E},\ell)$*, where* $P_{\mathsf{e}}(\mathtt{G},\mathtt{E},\ell)$ *denotes the fraction of erroneous messages after* ℓ *iterations of decoding.*

From the above lemma it suffices to prove the exchange of limits for the linearized algorithm. Note that $\epsilon^{\mathrm{LGalB}}$ as defined in Lemma 7.13 is the threshold of the LGalB algorithm. We will prove that for every $0 \leq \epsilon < \epsilon^{\mathrm{LGalB}}$ and every $\alpha > 0$ there exists an $N(\alpha,\epsilon)$ so that $\limsup_{\ell\to\infty} \mathbb{E}[P_b^{\mathrm{LGalB}}(\mathtt{G},\epsilon,\ell)] < \alpha$ for $N \geq N(\alpha,\epsilon)$. As we will see later, the monotonicity property of LGalB considerably simplifies the analysis. But the price paid for the simplification is that the technique works only for $\epsilon < \epsilon^{\mathrm{LGalB}}$, which is slightly smaller than the DE threshold.

Marking Process

The *marking* process allows us (i) to consider an *asynchronous* version of LGalB (i.e., the schedule of the computation is no longer important) and (ii) ensures that we are dealing with a monotone increasing function.

More precisely, we split the process into two phases: we start with LGalB for $\ell(p)$ iterations to get the error probability below p; we then continue the marking process associated with an infinite number of further iterations of LGalB. This means that we mark any variable that is bad in at least one iteration $\ell \geq \ell(p)$. Clearly, the union of all variables that are bad at at least one point in time $\ell \geq \ell(p)$ is an upper bound on the maximum number of variables that are bad at any specific instance in time.

The standard *schedule* of the LGalB is parallel, i.e., all incoming messages (at either variable or check nodes) are processed at the same time. This is the natural schedule for an actual implementation. For the purpose of analysis it is convenient to consider an *asynchronous* schedule.

Here is how the general asynchronous marking process proceeds. We are given a graph $\mathtt{G}$ and a noise realization $\mathtt{E}$. We are also given a set of *marked* edges. These marked edges are directed, from variable node to check node. At the start of the process mark the variable nodes that are connected to the marked edges. Declare all other variables and edges as *unmarked*. Unmarked edges do not have a direction. The process proceeds in discrete steps. At each step we pick a marked edge and we perform the processing described below. We continue until no more marked edges are left. Here are the processing rules:
If the marked edge $\mathtt{e}$ goes from variable to check:

- Let c be the check node connected to e. Declare e to be *unmarked* but *mark* all other edges connected to c; orient these marked edges from check to variable;

If the marked edge e goes from check to variable:

- Let v be the connected variable node. If v has a *good* associated channel realization and v is unmarked then mark v and declare e to be unmarked.
- Let v be the connected variable node. If v has an associated *bad* channel realization or if v has an associated *good* channel realization but is *marked*: (i) mark v and all its outgoing edges; (ii) orient the edges from variable to check; (iii) unmark e.

Let $\mathcal{M}(\mathtt{G}, \mathtt{E}, \mathcal{S})$ denote the set of marked variables assuming that we start with the set of marked edges $\mathcal{S}$ and that we run the asynchronous marking process. Let $M(\mathtt{G}, \mathtt{E}, \mathcal{S}) = |\mathcal{M}(\mathtt{G}, \mathtt{E}, \mathcal{S})|$. As a special case, let $\mathcal{M}(\mathtt{G}, \mathtt{E}, \ell)$ denote the set of marked variables at the end of the process assuming that the initial set of marked edges is the set of bad edges after ℓ rounds of LGalB. As before, $M(\mathtt{G}, \mathtt{E}, \ell) = |\mathcal{M}(\mathtt{G}, \mathtt{E}, \ell)|$.

It is not hard to see that for any $\ell \geq \ell'$, $P_b^{\mathrm{LGalB}}(\mathtt{G}, \epsilon, \ell) \leq M(\mathtt{G}, \mathtt{E}, \ell')/N$: for $\ell = \ell'$ both processes start with the same set of bad edges and both are operating on the same graph and noise realization. At the check-node side the processing rules are identical. At the variable-node side both processes also behave in the same way if they encounter a variable node with a bad channel realization. The difference lies in the behavior when they encounter a variable node with a good channel realization. In such a case the outgoing message for the LGalB is bad only if there are two bad messages entering *at the same time instance*. The asynchronous marking process algorithm declares the outgoing message to be bad if there are two incoming bad messages, even if the two messages might correspond to different time instants as measured by the parallel schedule. We conclude that for $\ell' \in \mathbb{N}$

$$\limsup_{\ell \to \infty} \mathbb{E}[P_b^{\mathrm{LGalB}}(\mathtt{G}, \epsilon, \ell)] \leq \frac{1}{N} \mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell')]. \tag{7.16}$$

Witness

It remains to bound $\mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell)]$. The difficulty in analyzing the marking process lies in the fact that after $\ell(p)$ iterations the set of starting edges for the marking process depends on the noise realization as well as the graph. Our aim therefore is to reduce this correlated case to the uncorrelated case by a sequence of transformations. As a first step we show how to get rid of the correlation with respect to the noise realization.

Consider a fixed graph G. Assume that we have performed ℓ iterations of LGalB. For each edge e that is bad in the ℓ-th iteration we construct a "witness." A witness for e is a subset of the computation tree of height ℓ

for $\mathtt{e}$ consisting of paths that carry bad messages. We construct the witness recursively starting with $\mathtt{e}$. Orient $\mathtt{e}$ from check node to variable node. At any point in time while constructing the witness associated to $\mathtt{e}$ we have a partial witness that is a tree with oriented edges. The initial such partial witness is $\mathtt{e}$. One step in the construction consists of taking a leaf edge of the partial witness and to "grow it out" according to the following rules.

If an edge enters a variable node that has an incorrect received value then add the *smallest* (according to some fixed but arbitrary order on the set of edges) edge that carries an incorrect incoming message to the witness and continue the process along this edge. The added edge is directed from variable node to check node. If an edge enters a variable node that has a correct received value then add both incoming edges to the witness and follow the process along both edges. (Note that in this case both of these edges must have carried bad messages.) Again, both of these edges are directed from variable to check node. If an edge enters a check node then choose the smallest incoming edge that carries an incorrect message and add it to the witness. Continue the process along this edge. The added edge is directed from check to variable node. Continue the process until depth ℓ. Fig. 7.1 shows an example for $\mathtt{d}_\mathtt{l} = 3$, $\mathtt{d}_\mathtt{r} = 4$, and $\ell = 2$. Denote the union of all witnesses for all edges that are bad

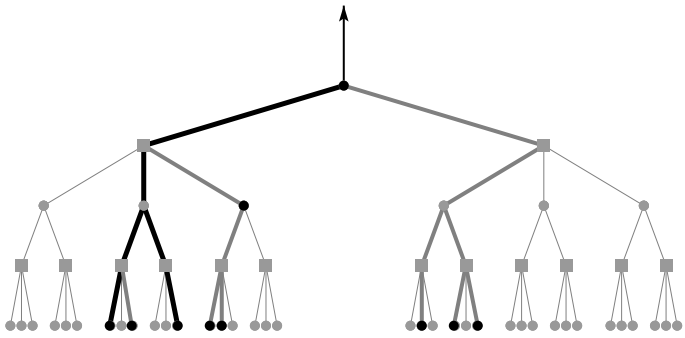

Figure 7.1: Construction of the witness for a bad edge. The *dark* variables represent channel errors. The part of the tree with *dark* edges represents the witness. The *thick* edges, including both dark and gray, represent the bad messages in the past iterations.

in the ℓ-th iteration by $\mathcal{W}(\mathtt{G}, \mathtt{E}, \ell)$. We simply call it *the witness*. The witness is a part of the graph that on its own explains why the set of bad edges after ℓ iterations is bad.

How large is $\mathcal{W}$? The larger ℓ, the fewer bad edges we expect to see in iteration ℓ. On the other hand, the size of the witness for each bad edge grows as a function of ℓ. Fortunately one can show that the first effect dominates and that the size of the witness vanishes as a function of the iteration number.

Lemma 7.16 (Size of Witness)**.** *Consider the* $(3, \mathtt{d}_\mathtt{r})$*-regular ensemble. For*

$0 \leq \epsilon < \epsilon^{LGalB}$,

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}[|\mathcal{W}(\mathtt{G},\mathtt{E},\ell)|] = o_\ell(1).$$

Why do we construct a witness? It is intuitive that if we keep the witness fixed but randomize the structure as well as the received values on the remainder of the graph then the situation should only get worse: already the witness itself explains all the bad messages and hence any further bad channel values can only create more bad messages. In the next two sections we show that under some suitable technical conditions this intuition is indeed correct.

Randomization

A witness $\mathcal{W}$ consists of two parts, (i) the graph structure of $\mathcal{W}$ and (ii) the channel realizations of the variables in $\mathcal{W}$. By some abuse of notation we write $\mathcal{W}$ also if we refer only to the graph structure or only to the channel realizations.

Fix a graph $\mathtt{G}$ and a witness $\mathcal{W}$, $\mathcal{W} \subseteq \mathtt{G}$. Let $\mathcal{E}_{\mathtt{G},\mathcal{W}}$ denote the set of all error realizations $\mathtt{E}$ that give rise to $\mathcal{W}$, i.e., $\mathcal{W}(\mathtt{G},\mathtt{E},\ell) = \mathcal{W}$. Clearly, for all $\mathtt{E} \in \mathcal{E}_{\mathtt{G},\mathcal{W}}$ we must have $\mathcal{W} \subseteq \mathtt{E}$. In words, on the set of variables fixed by the witness the errors are fixed by the witness itself. Therefore, the various $\mathtt{E}$ that create this witness differ only on $\mathtt{G}\backslash\mathcal{W}$. As a convention, we define $\mathcal{E}_{\mathtt{G},\mathcal{W}} = \emptyset$ if $\mathcal{W} \not\subseteq \mathtt{G}$.

Let $\mathcal{E}'_{\mathtt{G},\mathcal{W}}$ denote the set of projections of $\mathcal{E}_{\mathtt{G},\mathcal{W}}$ onto the variables in $\mathtt{G}\backslash\mathcal{W}$. Let $\mathtt{E}' \in \mathcal{E}'_{\mathtt{G},\mathcal{W}}$. Think of $\mathtt{E}'$ as an element of $\{0,1\}^{|\mathtt{G}\backslash\mathcal{W}|}$, where 0 denotes a correct received value and 1 denotes an incorrect received value. In this way, $\mathcal{E}'_{\mathtt{G},\mathcal{W}}$ is a subset of $\{0,1\}^{|\mathtt{G}\backslash\mathcal{W}|}$.

This is important: $\mathcal{E}'_{\mathtt{G},\mathcal{W}}$ has structure. We claim that, if $\mathtt{E}' \in \mathcal{E}'_{\mathtt{G},\mathcal{W}}$ then $\mathcal{E}'_{\mathtt{G},\mathcal{W}}$ also contains $\mathtt{E}'_{\leq}$, i.e., it contains all elements of $\{0,1\}$ that are smaller than $\mathtt{E}'$ with respect to the natural partial order on $\{0,1\}^{|\mathtt{G}\backslash\mathcal{W}|}$. More precisely, if the noise realization $\mathtt{E}' \in \mathcal{E}'_{\mathtt{G},\mathcal{W}}$ gives rise to the witness $\mathcal{W}$ then converting any incorrect received value in $\mathtt{E}'$ to a correct one will also give rise to $\mathcal{W}$. The proof of the following lemma relies heavily on this property. By some abuse of notation, let $\mathcal{M}(\mathtt{G},\mathtt{E},\mathcal{W})$, be the marking process with the edges in $\mathcal{W}$ as the initial set of bad edges.

Lemma 7.17 (Channel Randomization)**.** *Fix* $\mathtt{G}$ *and let* $\mathcal{W} \subseteq \mathtt{G}$*. Let* $\mathbb{E}_{\mathtt{E}'}[\cdot]$ *denote the expectation with respect to the channel realizations* $\mathtt{E}'$ *in* $\mathtt{G}\backslash\mathcal{W}$*. Then*

$$\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G},(\mathcal{W},\mathtt{E}'),\mathcal{W})\mathbb{1}_{\{\mathtt{E}'\in\mathcal{E}'_{\mathtt{G},\mathcal{W}}\}}] \leq \mathbb{E}_{\mathtt{E}'}[M(\mathtt{G},(\mathcal{W},\mathtt{E}'),\mathcal{W})]\mathbb{E}_{\mathtt{E}'}[\mathbb{1}_{\{\mathtt{E}'\in\mathcal{E}'_{\mathtt{G},\mathcal{W}}\}}].$$

Discussion: Lemma 7.17 has the following important operational significance. If we divide both sides by $\mathbb{E}_{\mathtt{E}'}[\mathbb{1}_{\{\mathtt{E}'\in\mathcal{E}'_{\mathtt{G},\mathcal{W}}\}}]$, the left-hand side is the expectation of marked variables, where the expectation is computed over all those channel realizations that give rise to the given witness $\mathcal{W}$, whereas the

right-hand side gives the expectation over all channel realizations (outside the witness) regardless whether they give rise to $\mathcal{W}$ or not. Clearly, the right-hand side is much easier to compute, since the channel is now independent of $\mathcal{W}$. The lemma states that, if we assume that the channel outside $\mathcal{W}$ is independently chosen then we get an upper bound on the size of the marked variables.

We can now upper bound the right-hand side of (7.16) as follows.

Lemma 7.18 (Markov Inequality)**.** *Consider the* $(\mathtt{d_l} = 3, \mathtt{d_r})$*-regular ensemble and transmission over the BSC*(ϵ)*. Let* $(\mathtt{G}, \mathtt{E})$ *be chosen uniformly at random. Let* $\ell \in \mathbb{N}$ *and* $\theta > 0$ *so that* $\mathbb{E}[|\mathcal{W}(\mathtt{G}, \mathtt{E}, \ell)|] \leq \theta^2 N$*. Then*

$$\mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell)] \leq \sum_{\mathcal{W}:|\mathcal{W}|\leq\theta N} \sum_{\mathtt{G}} \mathbb{P}\{\mathtt{G}\}\mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\}\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G}, (\mathcal{W}, \mathtt{E}'), \mathcal{W})] + \theta N.$$

Back to Expansion

Now where we have randomized the channel values we can use expansion arguments to deal with the dependence on the graph. The basic idea is simple. Assume that the neighborhood of initially bad edges (at the start of the marking process) is perfectly tree-like. This means that two bad edges never converge on the same variable node in their future. In this case the only bad messages emitted by a variable node are due to bad received values, but these received values can be thought of being chosen independently from the rest of the process. It follows that the whole marking process can be modeled as a birth and death process. When we grow out an edge then with probability ϵ we encounter a variable with a bad received value. In this case, the variable emits bad messages along its two outgoing edges and those in return each create $\mathtt{d_r} - 1$ bad outgoing messages at the output of their connected check nodes. In other words, with probability ϵ one bad edge is transformed to $2(\mathtt{d_r} - 1)$ bad edges. With probability $1 - \epsilon$ the process along the particular edge dies. By the stability condition of the LGalB decoder $2(\mathtt{d_r} - 1)\epsilon^{\mathrm{LGalB}} \leq 1$. We conclude that the expected number of newly generated children is strictly less than 1 for $\epsilon < \epsilon^{\mathrm{LGalB}}$. Therefore the corresponding birth and death process dies with probability 1.

Since in general the expansion of the local neighborhood is not perfectly tree-like, using the expansion on the check node side, the above argument has to be extended to account for this. But the gist of the argument remains the same.

Lemma 7.19 (Upper Bound)**.** *Fix* $\mathtt{G}$ *and* $\mathcal{W}$ *such that* $\mathcal{W} \subseteq \mathtt{G}$ *and* $\mathtt{G}$ *is a* $(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$*-right expander, with* $\gamma > 1 - \frac{2\mathtt{d_r}-1}{2\mathtt{d_r}(\mathtt{d_r}-1)}$*. If* $|\mathcal{W}| \leq c\alpha N$ *for some* $c = c(\mathtt{d_l}, \mathtt{d_r}, \epsilon, \gamma)$ *then*

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G}, (\mathcal{W}, \mathtt{E}'), \mathcal{W})] \leq \alpha\frac{\mathtt{d_l}}{\mathtt{d_r}}.$$

Putting It All Together

We are now ready to prove Lemma 7.13 using the results developed in the previous sections.

Proof of Lemma 7.13. Recall that we consider a $(\mathtt{d_l} = 3, \mathtt{d_r})$-regular ensemble and that $0 \leq \epsilon < \epsilon^{\text{LGalB}}$. Let $\alpha_{\max}(\gamma)$ be the constant defined in Theorem 7.4. Note that $\alpha_{\max}(\gamma)$ is strictly positive since δ is strictly positive. Choose $0 < \alpha < \alpha_{\max}(\gamma)$. Let $\mathcal{X}(\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma)$ denote the set of graphs $\{\mathtt{G} \in \text{LDPC}(N, \mathtt{d_l}, \mathtt{d_r}) : \mathtt{G} \in (\mathtt{d_l}, \mathtt{d_r}, \alpha, \gamma) \text{ right expander}\}$. From Theorem 7.4 we know that

$$\mathbb{P}\{\mathtt{G} \notin \mathcal{X}\} = o_N(1). \tag{7.17}$$

Let $c = c(\mathtt{d_l}, \mathtt{d_r}, \epsilon, \delta)$ be the coefficient appearing in Lemma 7.19 and define $\theta = c\alpha$. From Lemma 7.16 we know that there exists an iteration ℓ such that

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}[|\mathcal{W}(\mathtt{G}, \mathtt{E}, \ell)|] \leq \frac{1}{2}\theta^2. \tag{7.18}$$

Let $N(\theta)$ be such that for $N \geq N(\theta)$, $\mathbb{E}[|\mathcal{W}(\mathtt{G}, \mathtt{E}, \ell)|] \leq \theta^2 N$.

Using Lemma 7.18, and splitting the expectation over $\mathcal{X}$ and its complement, we get

$$\begin{aligned}\mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell)] \leq &\sum_{\mathcal{W}:|\mathcal{W}|\leq\theta N} \sum_{\mathtt{G}:\mathtt{G}\in\mathcal{X}} \mathbb{P}\{\mathtt{G}\}\mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\}\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G}, (\mathcal{W}, \mathtt{E}'), \mathcal{W})] + \\ &\sum_{\mathcal{W}:|\mathcal{W}|\leq\theta N} \sum_{\mathtt{G}:\mathtt{G}\notin\mathcal{X}} \mathbb{P}\{\mathtt{G}\}\mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\}\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G}, (\mathcal{W}, \mathtt{E}'), \mathcal{W})] + \theta N.\end{aligned}$$

Consider the first term. From Lemma 7.19 we know that

$$\mathbb{E}_{\mathtt{E}'}[M(\mathtt{G}, (\mathcal{W}, \mathtt{E}'), \mathcal{W})] \leq \alpha\frac{\mathtt{d_l}}{\mathtt{d_r}}N + o(N). \tag{7.19}$$

Consider the second term. Bound the expectation by N and remove the restriction on the size of the witness. This gives the bound

$$\sum_{\mathcal{W}} \sum_{\mathtt{G}:\mathtt{G}\notin\mathcal{X}} \mathbb{P}\{\mathtt{G}\}\mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\}N.$$

Switch the two summations and use the fact that, for a given $\mathtt{G}$, each $\mathtt{E}$ realization maps to only one $\mathcal{W}$. We get

$$\sum_{\mathtt{G}:\mathtt{G}\notin\mathcal{X}} \mathbb{P}\{\mathtt{G}\} \sum_{\mathcal{W}:\mathcal{W}\subseteq\mathtt{G}} \mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\} = \sum_{\mathtt{G}:\mathtt{G}\notin\mathcal{X}} \mathbb{P}\{\mathtt{G}\} = \mathbb{P}\{\mathtt{G} \notin \mathcal{X}\} \overset{(7.17)}{=} o_N(1). \tag{7.20}$$

From (7.19) and (7.20) we conclude that for $N \geq N(\theta)$,

$$\frac{1}{N}\mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell)] \leq \sum_{\mathcal{W}:|\mathcal{W}|\leq\theta N} \sum_{\mathtt{G}:\mathtt{G}\in\mathcal{X}} \mathbb{P}\{\mathtt{G}\}\mathbb{P}\{\mathcal{E}_{\mathtt{G},\mathcal{W}}\}\Big(\alpha\frac{\mathtt{d_l}}{\mathtt{d_r}} + o_N(1)\Big) + c\alpha$$

$$\leq \left(\frac{\mathtt{d_l}}{\mathtt{d_r}} + c\right) \alpha + o_N(1).$$

If we now let N tend to infinity then we get

$$\limsup_{N\to\infty} \limsup_{\ell\to\infty} \mathbb{E}[P_b^{\mathrm{LGalB}}(\mathtt{G}, \epsilon, \ell)] \leq \limsup_{N\to\infty} \frac{1}{N}\mathbb{E}[M(\mathtt{G}, \mathtt{E}, \ell)] \leq \left(\frac{\mathtt{d_l}}{\mathtt{d_r}} + c\right) \alpha.$$

Since this conclusion is valid for any $0 < \alpha \leq \alpha_{\max}(\gamma)$ it follows that

$$\lim_{N\to\infty} \limsup_{\ell\to\infty} \mathbb{E}[P_b^{\mathrm{LGalB}}(\mathtt{G}, \epsilon, \ell)] = 0. \qquad \square$$

7.5 Extensions

The proofs can be extended to other decoders. For a given MP decoder, the idea is to define an appropriate *linearized* version of the decoder (LMP) and go through the whole machinery as done for the GalB.

For example, consider the MS(M) decoder and transmission over the BSC(ϵ). The channel realizations are mapped to $\{\pm 1\}$. Let $M \in \mathbb{N}$, the message alphabet is $\mathcal{M} = \{-M, \ldots, M\}$. For transmission of the all-one codeword, the *linearized* version of the decoder (LMS(M)) is defined as in Definition 7.14: i.e., at the check node the outgoing message is the minimum of the incoming messages and the variable node rule is unchanged.

One can check that the LMS algorithm defined above is monotonic with respect to the input log-likelihoods at both the variable and check nodes and the number of errors in the MS decoder can be upper bounded by the errors of the LMS decoder.

Lemma 7.20 (MS(M) Decoder, BSC and $\mathtt{d_l} \geq 3$)**.** *Consider an LDPC* $(N, \mathtt{d_l}, \mathtt{d_r})$ *ensemble and transmission over the BSC(ϵ). Let* ϵ^{LMS} *be the channel parameter below which* $p_{\{M\}}^{(\infty)} = 1$*. If* $\epsilon < \epsilon^{LMS}$*, then*

$$\lim_{N\to\infty} \limsup_{\ell\to\infty} \mathbb{E}[P_b^{MS}(\mathtt{G}, \epsilon, \ell)] = 0.$$

Example 7.21 (MS(2) and BSC)**.** *Consider communication using an LDPC* $(N, 3, 6)$ *code over the BSC(ϵ) and decoding using the MS(2) algorithm. For this setup, the DE threshold is* 0.063*. The linearized decoder of this algorithm has* $p_{\{2\}}^{(\infty)} = 1$ *for* $\epsilon < 0.031$*. Therefore, from Lemma 7.20 the limits can be exchanged below* ϵ*.*

Similar results can be obtained for the BP(M) decoder, and channels with continuous outputs. But the analysis of these decoders is more complicated because we have to deal with densities of messages.

Here, we only considered channel parameters below the DE threshold. But the regime above this threshold is equally interesting. One important application of proving the exchange of limits in this regime is the finite-length analysis via a scaling approach [91] since the computation of the scaling parameters heavily depends on the fact that this exchange is permissible.

8 Capacity of CDMA

We now get to the second topic in graphical models. In this chapter we consider communication using a code division multiple access (CDMA) system with binary inputs. As in the previous chapter, we discuss only the main ideas of the proofs. The details can be found in [92].

8.1 Introduction

Code Division Multiple Access (CDMA) has been a successful scheme for reliable communication between multiple users and a common receiver. The scheme consists of K users modulating their information symbols by a signature sequence (spreading sequence) of length N and transmitting the resulting signal. The number N is sometimes referred to as the spreading gain or the number of chips per sequence. The receiver obtains the sum of all transmitted signals and the noise which is assumed to be white and Gaussian (AWGN).

The capacity region for continuous inputs with input power constraints and optimal decoding has been characterized in [49]. There it is shown that the achievable rates depend only on the correlation matrix of the spreading coefficients. If the spreading sequences are not orthogonal then the complexity of optimum detectors scales exponentially with the number of users. Therefore, in practice low-complexity linear detectors [93] are considered.

Random spreading sequences were considered in the literature, because it allows to obtain nice analytic formulas for various quantities of interest in the *large system limit* ($K \to \infty, N \to \infty, \frac{K}{N} = \beta$), see [94, 95, 96]. It also provides qualitative insights to the problem. In practice too it is reasonable to assume random spreading. It models the scenario where the spreading sequences are pseudo-noise sequences having length much larger than the symbol intervals. Then the spreading sequence corresponding to each symbol interval behaves

as a randomly chosen sequence. Random spreading also models the scenario where the signal is distorted by channel fading. In [95] and [96], the authors considered random spreading and analyzed the spectral efficiency, defined as the bits per chip that can be reliably transmitted, for these detectors. In the large-system limit they obtained analytical formulas for the spectral efficiency and showed that it concentrates with respect to the randomness in the spreading sequences. These results follow from the known spectrum of large random covariance matrices. We can say that the system is reasonably well understood for Gaussian inputs.

In practice however, the input of the user is restricted to a constellation, say PAM or QAM. Not much is known in this case. A notable exception is the spectral efficiency for binary input in the case of high SNR [97]. The main reason for this disparity in understanding is that the random matrix tools which played a central role in the analysis for Gaussian inputs are not applicable here.

The work of Tanaka [98] is a breakthrough in the analysis of binary input CDMA system. Using the non-rigorous "replica method", developed for analyzing random spin systems in statistical mechanics, Tanaka computed the spectral efficiency for both optimal and sub-optimal detectors and also computed the bit error rate (BER) for uncoded transmission. The analysis is extended in [99] to include the case of unequal powers and other constellations. The replica method, though non-rigorous, is believed to yield exact results for some models in statistical mechanics [100], known an mean-field models. The CDMA system can be viewed as one such model and hence the conjectures are believed to be true. Some evidence to this belief is given by Montanari and Tse [101]. Using tools developed for sparse graph codes, they compute the spectral efficiency for some range of β and the formulas obtained match with the conjectured formula of Tanaka.

In this chapter we prove various results for the binary input CDMA systems. Our approach is based on rigorous mathematical tools developed in statistical mechanics. Our main result is an upper bound on the spectral efficiency which matches with the conjecture of Tanaka and hence believed to be tight. In the process we show that the spectral efficiency concentrates around its average. We also show that the spectral efficiency is independent of the spreading sequence distribution. Although these methods work for other constellations including Gaussian inputs, we restrict our discussion to binary inputs for the sake of clarity.

In Section 8.2 we introduce the statistical mechanics tool that we use along with an example. We discuss the communication set-up and its formulation as a statistical mechanics problem in Section 8.3. In Section 8.4 we discuss Tanaka's conjecture on the spectral efficiency. Our results, along with the proof ideas, follow in the later sections. In Section 8.5 we provide concentration results for the spectral efficiency. We show that the spectral efficiency is independent of the spreading sequence distribution in Section 8.6. Section 8.7 provides an upper bound on the spectral efficiency for all values of β. We end

this chapter with some extensions in Section 8.9.

8.2 Statistical Mechanics Approach

There is a natural connection between various communication systems and statistical mechanics of random spin systems, stemming from the fact that often in both systems there is a large number of degrees of freedom (bits or spins), interacting in a random environment. So far, there have been applications of two important but somewhat complementary approaches of statistical mechanics of random systems.

The first one is the very important but mathematically non-rigorous replica method. The merit of this approach is that it allows one to obtain explicit formulas for quantities of interest such as, conditional entropy, or error probability. The replica method has been applied to many scenarios in communication including channel and source coding using sparse graph codes, multiuser settings like broadcast channel (see for example [102], [103], [104]) and the case of interest here [98]: randomly spread CDMA with binary inputs.

The second type of approach aims at a rigorous understanding of the replica formulas and has its origins in methods stemming from mathematical physics (see [105, 106, 107]). For systems whose underlying degrees of freedom have Gaussian distribution (Gaussian input symbols or Gaussian spins in continuous spin systems) random matrix methods can successfully be employed. However when the degrees of freedom are binary (binary information symbols or Ising spins) these seem to fail. Fortunately in the later case, the recently developed interpolation method by Guerra has had a lot of success.[1]

As a pedagogical example, let us consider the well studied Sherrington-Kirkpatrick (SK) model in statistical mechanics. Guerra's interpolation method was originally developed for this model. The system consists of N spins denoted by x_i taking values in $\{\pm 1\}$. The system can therefore lie in 2^N possible states. To every state $\bar{x}$ we associate a Hamiltonian or energy function, denoted by $\mathtt{H}(\bar{x})$. One of the main postulates of statistical mechanics is that, when the system is in equilibrium, the probability that it lies in a particular state is proportional to $e^{\mathtt{H}(\bar{x})}$.[2] Therefore the probability is given by

$$\Pr(\text{state of the system is } \bar{x}) = \frac{e^{\mathtt{H}(\bar{x})}}{Z},$$

where Z is the normalization factor $Z = \sum_{\bar{x}\in\{\pm 1\}^N} e^{\mathtt{H}(\bar{x})}$. The quantity Z is called the partition function which is in turn used to define the free energy of

[1] Let us point out that, as it will be shown later in this chapter, the interpolation method can also serve as an alternative to random matrix theory for Gaussian inputs.

[2] In statistical mechanics it is more natural to consider $e^{-\beta \mathtt{H}(\bar{x})}$, where β plays the role of inverse temperature. Here, with some abuse of notation we include the negative sign in our definition of Hamiltonian. Moreover, since the precise value of β is not important for our discussion, we set $\beta = 1$.

the system as follows. The free energy denoted by $\mathcal{F}$ is defined as

$$\mathcal{F} = \frac{1}{N} \log Z.$$

Many physical properties of a system like its internal energy, specific heat, and magnetization can be expressed in terms of free energy. Therefore, computing the free energy has been a central topic in statistical mechanics.

For the SK model the Hamiltonian is a random function given by

$$\mathtt{H}(\bar{x}) = -\frac{1}{\sqrt{N}} \sum_{i<j} J_{ij} x_i x_j,$$

where $J_{ij} \sim \mathcal{N}(0, J)$. The randomness is used to model the impurities in the system. The partition function depends on $\{J_{ij}\}$. Therefore let us denote it by $Z(\{J_{ij}\})$. The free energy is then given by

$$\mathcal{F}(\{J_{ij}\}) = \frac{1}{N} \log Z(\{J_{ij}\}).$$

The free energy defined above is a random quantity due to the randomness in $\{J_{ij}\}$. Let $\mathcal{F}(N)$ denote the free energy averaged over the randomness in $\{J_{ij}\}$, i.e.,

$$\mathcal{F}(N) = \mathbb{E}_{\{J_{ij}\}}[\mathcal{F}(\{J_{ij}\})].$$

The regime of interest is the limit of large number of spins $N \to \infty$, known in the literature as thermodynamic limit.

A common assumption in statistical mechanics is that the free energy of the random system is self-averaging. In other words, the free energy concentrates around its expectation. Therefore, much of the work in statistical mechanics is concentrated on computing the average free energy in the thermodynamic limit. Computing $\lim_{N\to\infty} \mathcal{F}(N)$ is not an easy task. The difficulty arises from the fact that we need to compute the expectation of the logarithm of a sum of coupled random variables ($e^{\mathtt{H}(\bar{x})}$).

Parisi [108] developed a sequence of approximations for the free energy using the replica method. According to the replica prediction, the free energy for $J \leq 1$ is given by[3]

$$\lim_{N\to\infty} \mathcal{F}(N) = \min_{m\in[0,1]} \left\{ \frac{J^2}{4}(1-m)^2 + \int_{-\infty}^{+\infty} \log(2\cosh(Jmz)) Dz \right\}, \qquad (8.1)$$

where $Dz \equiv \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$. We refer the reader to the book by Mézard, Parisi, and Virasoro [109] for the details of the remarkable replica method.

The rigorous justification of Parisi's free energy remained an open problem for at least two decades until the invention of the interpolation method. The

[3]The solution for $J > 1$ is much more subtle and complicated. We do not need discuss this result here.

basic idea of the interpolation method is to study a system which interpolates between the original system and a simpler one. The simpler system is usually guessed from the solution of the replica method.

For the SK model, the Hamiltonian of the interpolated system is given as follows. For any $t \in [0,1]$ and $m \in [0,1]$, let $\mathtt{H}_t(\bar{x})$ denote the "interpolated" Hamiltonian defined as

$$\mathtt{H}_t(\bar{x}) = -\sqrt{t}\frac{1}{\sqrt{N}}\sum_{i<j} J_{ij}x_i x_j - \sqrt{1-t}\sum_i J_i x_i, \tag{8.2}$$

where $J_i \sim \mathcal{N}(0, Jm)$. Let the corresponding free energy be denoted by $\mathcal{F}_t$. It is given by

$$\mathcal{F}_t = \frac{1}{N}\mathbb{E}\Big[\log \sum_{\bar{x}} e^{\mathtt{H}_t(\bar{x})}\Big],$$

where the expectation is over both $\{J_{ij}\}$ and $\{J_i\}$. Note that $\mathcal{F}_1 = \mathcal{F}(N)$, the free energy of the original system. As t is varied from 1 to 0 the Hamiltonian $\mathtt{H}_t(\bar{x})$ interpolates between the original Hamiltonian and a simpler one. The Hamiltonian $H_0(\bar{x})$ consists of only single spin terms and hence it can be easily shown to be

$$\mathcal{F}_0 = \frac{1}{N}\mathbb{E}\Big[\log \prod_i \sum_{x_i\in\{\pm 1\}} e^{-J_i x_i}\Big] = \int_{-\infty}^{+\infty} \log(2\cosh(Jmz))Dz.$$

The free energy $\mathcal{F}_0$ is equal to the integral term in (8.1). This is not a coincidence. On the other hand, the Hamiltonian $\mathtt{H}_t(\bar{x})$ is chosen intentionally to result in this free energy at $t = 0$. Using the fundamental theorem of calculus, we can express $\mathcal{F}_1$ as

$$\mathcal{F}_1 = \mathcal{F}_0 + \int_0^1 dt \frac{d\mathcal{F}_t}{dt}.$$

Therefore, to compute $\mathcal{F}_1$ it is sufficient to know the derivative of $\mathcal{F}_t$. The derivative can be expressed as

$$\frac{d\mathcal{F}_t}{dt} = \frac{J^2}{4}(1-m)^2 - R(t),$$

where $R(t)$ is a complicated quantity. Estimating $R(t)$ is as difficult as the original task. But the beauty of the interpolation method is that, after some clever manipulations one can show that $R(t)$ is the square of another real quantity and hence $R(t) > 0$. Therefore, the free energy can be bounded as

$$\mathcal{F} \leq \frac{\beta^2 J^2}{4}(1-m)^2 + \int_{-\infty}^{+\infty} \log(2\cosh(\beta m)z)Dz,$$

for any $m \in [0,1]$ which matches with the expression in (8.1). Later, Guerra extended this idea to obtain a sequence of bounds for the free energy that

match with the conjectured free energy of Parisi [108] for all values of β. Talagrand [110] extended the technique in an intelligent way to obtain lower bounds as well and thus proving the long standing conjecture of Parisi. The interpolation method is very powerful tool. In [106, 107] it has been used to show the concentration of the free energy and also to show the existence of the thermodynamic limit of the free energy.

So far in communications, the interpolation method has been developed only for sparse graph codes (LDPC and LDGM) over binary input symmetric channels [111, 112, 113]. We develop the interpolation method for the random CDMA system with binary inputs. The situation is qualitatively different than the ones mentioned above in that the "underlying graph" is complete and the structure of the interaction between the degrees of freedom are more complicated.

8.3 Communication Setup

The system consists of K users sending binary information symbols $x_k \in \{\pm 1\}$, to a common receiver. The user k has a random signature sequence $\bar{s}_k = (s_{1k}, ..., s_{Nk})^\top$, where the components s_{ik} are i.i.d. realizations of a random variable S. The random variable S is assumed to be symmetric, i.e., $p_S(s) = p_S(-s)$. For each time division (or chip interval) $i = 1, ..., N$ the received signal y_i is given by

$$y_i = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} s_{ik} x_k + \sigma n_i,$$

where n_i are i.i.d. realizations of $\mathcal{N}(0, 1)$. Therefore, the noise power is σ^2. The variance of S is assumed to be 1 and the scaling factor $1/\sqrt{N}$ is introduced so that the power (per symbol) of each user is normalized to 1.

In the sequel we write $\mathbf{s}$ for the $N \times K$ matrix (s_{ik}), $\mathbf{S}$ for the corresponding random matrix. We use $\bar{x}$ to denote the vector $(x_1, \ldots, x_K)^\top$ and $\bar{X}$ to denote the vector of random variables $(X_1, \ldots, X_K)^\top$. Similarly, $\bar{y}$ and $\bar{Y}$ denote the N dimensional vectors $(y_1, \ldots, y_N)^\top$ and $(Y_1, \ldots, Y_N)^\top$ respectively. Let $p_{\bar{X}}(\bar{x}) = \prod_k p_k(x_k)$ denote the input distribution. The quantity of interest is

$$C_K = \frac{1}{K} \max_{\prod_{k=1}^{K} p_k(x_k)} I(\bar{X}; \bar{Y} \,|\, \mathbf{S}) \tag{8.3}$$

in the large-system limit, i.e., $K \to +\infty$ with $\frac{K}{N} = \beta$ fixed. It is easy to show that the maximum is obtained for $X_k \sim \mathrm{Ber}(\frac{1}{2})$ [92]. We refer to C_K as the capacity of the CDMA system. The spectral efficiency is related to the capacity as $\frac{1}{\beta} C_K$.

Let us give a couple of reasons that justify the expectation over the spreading sequences. In the case where the spreading sequences are much longer than the symbol interval, i.e., they span many symbols, then the average over the spreading sequences can be interpreted as the average over time. Hence, the

expectation gives the ergodic capacity. In the case where the same spreading sequence is used for every symbol, the average is still justified due to the concentration of the capacity around its average (see Section 8.5).

Let us now collect a few formulas that will be useful in the sequel. Fix $\mathbf{s}$ and consider the probability distribution

$$p(\bar{x} \mid \bar{y}, \mathbf{s}) = \frac{p_{\bar{X}}(\bar{x})}{Z(\bar{y}, \mathbf{s})} \exp\big(-\frac{1}{2\sigma^2}\|\bar{y} - N^{-\frac{1}{2}}\mathbf{s}\bar{x}\|^2\big) \tag{8.4}$$

with the normalization factor

$$Z(\bar{y}, \mathbf{s}) = \sum_{\bar{x}} p_{\bar{X}}(\bar{x}) e^{-\frac{1}{2\sigma^2}\|\bar{y} - N^{-\frac{1}{2}}\mathbf{s}\bar{x}\|^2}. \tag{8.5}$$

In the language of statistical mechanics, the bits x_i play the role of spins. The Hamiltonian for the state $\bar{x}$ is given by $\mathtt{H}(\bar{x}) = -\frac{1}{2\sigma^2}\|\bar{y} - N^{-\frac{1}{2}}\mathbf{s}\bar{x}\|^2$. The random variables $\mathbf{S}$ and the Gaussian noise play the role of $\{J_{ij}\}$ for the SK-model. The normalization factor (8.5) can be interpreted as the partition function. In view of this it is not surprising that the free energy

$$f(\bar{y}, \mathbf{s}) = \frac{1}{K} \ln Z(\bar{y}, \mathbf{s}) \tag{8.6}$$

plays a crucial role. One can easily show [92] that the free energy is related to the mutual information as

$$\frac{1}{K} I(\bar{X}; \bar{Y} \mid \mathbf{S} = \mathbf{s}) = -\frac{1}{2\beta} - \mathbb{E}_{\bar{Y} \mid \mathbf{s}}[f(\bar{Y}, \mathbf{s})]. \tag{8.7}$$

Therefore

$$C_K = -\frac{1}{2\beta} - \mathbb{E}_{\bar{Y}, \mathbf{S}}[f(\bar{Y}, \mathbf{S})], \tag{8.8}$$

where $p_{\bar{X}}(\bar{x}) = \frac{1}{2^K}$.

8.4 Tanaka's Conjecture

In this section let us restrict the input distribution to be uniform, i.e., $p_{\bar{X}}(\bar{x}) = \frac{1}{2^K}$. Using the replica method, Tanaka conjectured that the capacity of the CDMA system is given by

$$\lim_{K \to \infty} C_K = \min_{m \in [0,1]} c_{RS}(m), \tag{8.9}$$

where the "replica symmetric capacity functional" is given by

$$c_{RS}(m) = \frac{\lambda}{2}(1 + m) - \frac{1}{2\beta} \ln \lambda\sigma^2 - \int_{-\infty}^{+\infty} Dz \ln(\cosh(\sqrt{\lambda} z + \lambda)). \tag{8.10}$$

The parameter λ is defined by

$$\lambda = \frac{1}{\sigma^2 + \beta(1-m)}, \tag{8.11}$$

and Dz is the standard Gaussian measure $Dz \equiv \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz$. For a Gaussian random variable Z and any differentiable function f, using integration by parts, one can show that

$$\int_{-\infty}^{+\infty} zf(z)Dz = \int_{-\infty}^{+\infty} f'(z)Dz.$$

Using this formula it is easy to show that the minimizer in (8.9) must satisfy the fixed point condition

$$m = \int_{-\infty}^{+\infty} Dz \tanh(\sqrt{\lambda}z + \lambda). \tag{8.12}$$

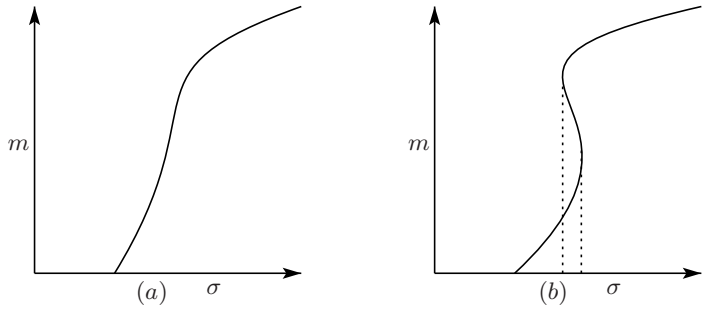

Figure 8.1: An illustration of the behavior of solutions of (8.12). The figure (a) shows the behavior for $\beta < \beta_s$ (no phase transition), where (8.12) has only one solution for every σ. The figure (b) corresponds to $\beta > \beta_s$ (phase transition), where (8.12) has multiple solutions between the two dashed lines.

The work of Montanari and Tse [101] provides strong support to the conjecture at least in a regime of β without phase transitions. More precisely, the proof is valid for $\beta \leq \beta_s$ where β_s is the maximal value of β such that the solution of (8.12) remains unique for all $\sigma \in [0, \infty)$ (see Figure 8.4). The authors first solve the case of sparse signature sequence in the limit $K \to \infty$. Then the dense signature sequence (which is of interest here) is recovered by exchanging the $K \to \infty$ and $sparse \to dense$ limits.

Tanaka also conjectured that the capacity in the large system limit is independent of the spreading sequence distribution, as long as the distributions are symmetric with equal second moment and finite fourth moment.

8.5 Concentration of Capacity

In the case of a Gaussian input signal, the concentration can be deduced from general theorems on the concentration of the spectral density for random

matrices. But this approach breaks down for binary inputs. Using other techniques we obtain two concentration results.

For compactness let us first introduce the following notation. Let $\imath(\bar{X};\bar{Y}\,|\,\mathbf{s}) = I(\bar{X},\bar{Y}\,|\,\mathbf{S}=\mathbf{s})$. We will treat $\imath(\bar{X};\bar{Y}\,|\,\mathbf{s})$ as a function of $\mathbf{s}$. Then $\imath(\bar{X};\bar{Y}\,|\,\mathbf{S})$ is a random variable depending on $\mathbf{S}$ whose expectation is given by $\mathbb{E}_{\mathbf{S}}[\imath(\bar{X};\bar{Y}\,|\,\mathbf{S})] = I(\bar{X};\bar{Y}\,|\,\mathbf{S})$.

Theorem 8.1 (Gaussian Spreading Sequences)**.** *Consider a CDMA system with binary inputs and let the spreading distribution be the standard Gaussian distribution. Given $\epsilon > 0$, there exists an integer $K_1 = O(|\ln\epsilon|)$ independent of $p_{\bar{X}}$, such that for all $K > K_1$,*

$$\mathbb{P}[|\imath(\bar{X};\bar{Y}\,|\,\boldsymbol{S}) - \mathbb{E}_{\boldsymbol{S}}[\imath(\bar{X};\bar{Y}\,|\,\boldsymbol{S})]| \geq \epsilon K] \leq 3e^{-\alpha K},$$

where $\alpha(\beta,\sigma,\epsilon) > 0$ and is independent of K.

To prove this theorem we use powerful probabilistic tools developed by Ledoux and Talagrand for Lipschitz functions of many Gaussian random variables [100].

Unfortunately the same tools do not apply directly to the case of other spreading sequences. However in this case the following weaker result can be obtained.

Theorem 8.2 (General Spreading Sequence)**.** *Consider a CDMA system with binary inputs and let the spreading sequence distribution be symmetric with finite fourth moment. There exists an integer K_1 independent of $p_{\bar{X}}$, such that for all $K > K_1$*

$$\mathbb{P}[|\imath(\bar{X};\bar{Y}\,|\,\boldsymbol{S}) - \mathbb{E}_{\boldsymbol{S}}[\imath(\bar{X};\bar{Y}\,|\,\boldsymbol{S})]| \geq \epsilon K] \leq \frac{\alpha}{K\epsilon^2}$$

where $\alpha(\beta,\sigma) > 0$ and is independent of K.

To prove such estimates it is enough (by Chebycheff) to control second moments. For the proof we use another interpolation technique developed by Pastur, Shcherbina and Tirozzi [114, 115]. Since the concentration proofs are mainly technical we direct the reader to [92] for their proofs.

8.6 Independence of Capacity with respect to the Spreading Sequence Distribution

The replica method leads to the same formula for the capacity for all symmetric spreading sequence distributions with equal second moment and finite fourth moment. Here we rigorously show a result of that flavor for the following class of distributions.

Class A. *The distribution $p_S(s)$ is symmetric*

$$p_S(s) = p_S(-s)$$

and has a rapidly decaying tail. More precisely, there exist positive constants s_0 *and* A *such that* $\forall s \geq s_0$

$$\Pr(S \geq s) \leq e^{-As^2}.$$

In particular, the Gaussian and binary cases are included in this class, and also any compactly supported distribution. We believe that a better approximation of some of the error terms in our proofs would widen the class of distributions to the one predicted by replica method.

Theorem 8.3 (Independence of Capacity with respect to Spreading Sequence Distribution)**.** *Consider a CDMA system with binary inputs. Let* C_K *denote the capacity for a spreading sequence distribution belonging to Class A. Let* C_K^g *denote the capacity for Gaussian spreading sequence distribution having the same second moment. Then*

$$\lim_{K \to +\infty} |C_K - C_K^g| = 0.$$

The proof is based on an interpolation between appropriately chosen systems. Let $\{r_{ik}\}$ denote the spreading sequence realizations of distribution belonging to class A and let $\{s_{ik}\}$ denote the spreading sequences generated from Gaussian distribution. The idea is to interpolate between the two systems using spreading sequences of the form

$$v_{ik}(t) = \sqrt{t} r_{ik} + \sqrt{1-t} s_{ik}, \qquad 0 \leq t \leq 1.$$

Let $\mathbf{v}(t)$ denote the matrix with entries $v_{ik}(t)$. By the fundamental theorem of calculus the capacities are related by

$$C_K - C_K^g = \mathbb{E}_{\mathbf{R}}[C(\mathbf{R})] - \mathbb{E}_{\mathbf{S}}[C(\mathbf{S})] = \int_0^1 dt \frac{d}{dt} \mathbb{E}_{\mathbf{V}(t)}[C(\mathbf{V}(t))].$$

The proof follows by showing that the term in the integral vanishes in the large user limit. For the details of the proof please refer to [92].

8.7 Tight Upper Bound on the Capacity

Our main result is that Tanaka's formula (8.10) is an upper bound to the capacity for all values of β.

Theorem 8.4 (Upper Bound on Capacity)**.** *Consider a CDMA system with binary inputs and let the spreading sequence distribution belong to Class A. Let* C_K *denote its capacity. Then*

$$\lim_{K \to \infty} C_K \leq \min_{m \in [0,1]} c_{RS}(m), \tag{8.13}$$

where $c_{RS}(m)$ *is given by (8.10).*

To be precise the left hand side must be lim sup instead of lim. However, in the next section, we show that this limit exists which allows us to replace the lim sup with lim.

Combining the above theorem with an inequality in Montanari and Tse [101], one can deduce that the equality holds for some regime of noise smaller than a critical value for all β. This value corresponds to the threshold for belief propagation decoding. Note that this equality is valid even if β is such that there is a phase transition (the fixed point equation (8.12) has many solutions), whereas in [101] the equality holds for values of β for which the phase transition does not occur.

Since the proof is rather complicated we give the main ideas in an informal way. The integral term in (8.10) suggests that we can replace the original system with a simpler system where the user bits are sent through K independent Gaussian channels given by

$$y'_k = x_k + \frac{1}{\sqrt{\lambda}} w_k, \tag{8.14}$$

where $W_k \sim \mathcal{N}(0,1)$ and λ is an effective SNR. Of course this argument is a bit naive because this effective system does not account for the extra terms in (8.10), but it has the merit of identifying the correct interpolation.

We introduce an interpolating parameter $t \in [0,1]$ such that the independent Gaussian channels correspond to $t = 0$ and the original CDMA system corresponds to $t = 1$ (see Figure 8.7) It is convenient to denote the SNR of the original Gaussian channel as B (that is $B = \sigma^{-2}$). Then (8.11) becomes

$$\lambda = \frac{B}{1 + \beta B(1-m)}.$$

We introduce two interpolating SNR functions $\lambda(t)$ and $B(t)$ such that

$$\lambda(0) = \lambda,\ B(0) = 0, \qquad \text{and} \qquad \lambda(1) = 0,\ B(1) = B, \tag{8.15}$$

and

$$\frac{B(t)}{1 + \beta B(t)(1-m)} + \lambda(t) = \frac{B}{1 + \beta B(1-m)}. \tag{8.16}$$

The meaning of (8.16) is the following. In the interpolating t-system the effective SNR seen by each user has an effective t-CDMA part and an independent channel part $\lambda(t)$ chosen such that the total SNR is fixed to the effective SNR of the CDMA system. There is a whole class of interpolating functions satisfying the above conditions but it turns out that we do not need to specify them more precisely except for the fact that $B(t)$ is increasing, $\lambda(t)$ is decreasing and with continuous first derivatives. Subsequent calculations are independent of the particular choices of functions.

The parameter m is to be considered as fixed to any arbitrary value in $[0,1]$. All the subsequent calculations are independent of its value, which is to be optimized to tighten the final bound.

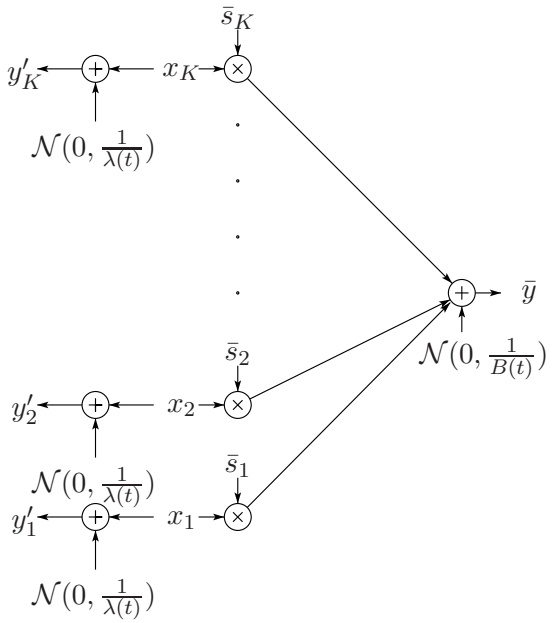

Figure 8.2: The information bits x_k are transmitted through the normal CDMA channel with variance $\frac{1}{B(t)}$ and through individual Gaussian channels with noise $\frac{1}{\lambda(t)}$.

We now have two sets of channel outputs $\bar{y}$ (from the CDMA with noise variance $B(t)^{-1}$) and $\bar{y}'$ (from the independent channels with noise variance $\lambda(t)^{-1}$) and the interpolating communication system has a posterior distribution

$$p_t(\bar{x}|\bar{y},\bar{y}',\mathbf{s}) = \frac{2^{-K}}{Z_t(\bar{y},\bar{y}',\mathbf{s})} \exp\left(-\frac{B(t)}{2}\|\bar{y}-N^{-\frac{1}{2}}\mathbf{s}\bar{x}\|^2 - \frac{\lambda(t)}{2}\|\bar{y}'-\bar{x}\|^2\right), \quad (8.17)$$

where Z_t is the normalizing constant. The mutual information $I_t(\bar{X};\bar{Y},\bar{Y}' \,|\, \mathbf{S})$ corresponds to the distribution $p_{\bar{X}}(\bar{x}^0)p_{\mathbf{S}}(\mathbf{s})p_t(\bar{y},\bar{y}' \,|\, \bar{x}^0,\mathbf{s})$, where

$$p_t(\bar{y},\bar{y}' \,|\, \mathbf{s},\bar{x}^0) = \frac{1}{(\sqrt{2\pi B(t)^{-1}})^N(\sqrt{2\pi\lambda(t)^{-1}})^K} e^{-\frac{B(t)}{2}\|\bar{y}-N^{-\frac{1}{2}}\mathbf{s}\bar{x}^0\|^2-\frac{\lambda(t)}{2}\|\bar{y}'-\bar{x}^0\|^2}, \quad (8.18)$$

and $p_{\bar{X}}(\bar{x}^0) = 2^{-K}$ by assumption.

Let $\langle - \rangle_t$ denote the expectation with respect to (8.18), i.e., for any function $g(\bar{x}^0,\bar{x})$

$$\langle g(\bar{x}^0,\bar{x})\rangle_t = \sum_{\bar{x}^0,\bar{x}} g(\bar{x}^0,\bar{x})p_t(\bar{x},\bar{x}^0 \,|\, \bar{y},\bar{y}',\mathbf{s}).$$

Let $f_t(\bar{y},\bar{y}',\mathbf{s}) = \frac{1}{K}\ln Z_t(\bar{y},\bar{y}',\mathbf{s})$, be the free energy of the interpolating system. The free energy of the original CDMA system $\mathbb{E}[f(\bar{Y},\mathbf{S})]$ can be

expressed as

$$\mathbb{E}[f(\bar{Y}, \mathbf{S})] = \frac{1}{2} + \mathbb{E}[f_1(\bar{Y}, \bar{Y}', \mathbf{S})].$$

The interpolation method suggests to compute $\mathbb{E}[f_1(\bar{Y}, \bar{Y}', \mathbf{S})]$ through

$$\mathbb{E}[f_1(\bar{Y}, \bar{Y}', \mathbf{S})] = \mathbb{E}[f_0(\bar{Y}, \bar{Y}', \mathbf{S})] + \int_0^1 \frac{d}{dt}\mathbb{E}[f_t(\bar{Y}, \bar{Y}', \mathbf{S})]dt.$$

If we proceed in this manner, it turns out that we need a concentration result on empirical average of the "magnetization", $m_1 = \frac{1}{K}\sum_{k=1}^K x_k^0 x_k$. More precisely, we require a self-averaging result of the form

$$\lim_{N\to\infty} \mathbb{E}\langle |m_1 - \mathbb{E}\langle m_1\rangle_t|\rangle_t = 0.$$

Proving such a result is not easy. To overcome this difficulty we consider a slightly perturbed system.

Consider a slightly more general interpolation system where the perturbation term

$$h_u(\bar{x}) = \sqrt{u}\sum_{k=1}^K h_k x_k + u\sum_{k=1}^K x_k^0 x_k - \sqrt{u}\sum_{k=1}^K |h_k| \tag{8.19}$$

is added in the exponent of the measure (8.17). Here h_k are realizations of $H_k \sim \mathcal{N}(0,1)$. For the moment $u \geq 0$ is arbitrary but in the sequel we will take $u \to 0$.

$$p_{t,u}(\bar{x}|\bar{y}, \bar{y}', \bar{h}, \mathbf{s}) = \frac{2^{-K}}{Z_{t,u}(\bar{y}, \bar{y}', \bar{h}, \mathbf{s})} \exp\left(-\frac{B(t)}{2}\|\bar{y} - N^{-\frac{1}{2}}\mathbf{s}\bar{x}\|^2 - \frac{\lambda(t)}{2}\|\bar{y}' - \bar{x}\|^2 + h_u(\bar{x})\right) \tag{8.20}$$

with the obvious normalization factor $Z_{t,u}(\bar{y}, \bar{y}', \bar{h}, \mathbf{s})$. We define the new interpolated free energy

$$f_{t,u}(\bar{y}, \bar{y}', \bar{h}, \mathbf{s}) = \frac{1}{K}\ln Z_{t,u}(\bar{y}, \bar{y}', \bar{h}, \mathbf{s}). \tag{8.21}$$

For this perturbed system we can show the following concentration result. Let $\langle - \rangle_{t,u}$ denote the expectation with respect to the interpolating distribution $p_{t,u}(\bar{x}, \bar{x}^0|\bar{y}, \bar{y}', \bar{h}, \mathbf{s})$.

Theorem 8.5 (Concentration of Magnetization)**.** *Fix any $\epsilon > 0$. For Lebesgue almost every $u > \epsilon$,*

$$\lim_{N\to\infty} \int_0^1 dt\mathbb{E}\langle |m_1 - \mathbb{E}\langle m_1\rangle_{t,u}|\rangle_{t,u} = 0.$$

The proof of this theorem is quite involved and we the direct interested reader to [92]. Observe that the theorem is not proved for every u and especially not for $u = 0$ which is the original CDMA system.

From now on, for pedagogic reasons we drop the arguments of the function f whenever it is clear from the subscripts. For $t = 1$ and $u = 0$ we recover the original free energy,

$$\mathbb{E}[f(\bar{Y}, \mathbf{S})] = \frac{1}{2} + \mathbb{E}[f_{1,0}]$$

while for $t = 0$ and $u = 0$, the statistical sums decouple and we have the explicit result[4]

$$\frac{1}{2} + \mathbb{E}[f_{0,0}] = -\frac{1}{2\beta} - \lambda + \int_{-\infty}^{+\infty} Dz \ln(2\cosh(\sqrt{\lambda}z + \lambda)) \tag{8.22}$$

where $\mathbb{E}$ denotes the appropriate collective expectation over random objects. Using $|h_u(\bar{x})| \leq 2\sqrt{u}\sum_k |h_k| + Ku$ it easily follows that

$$|\mathbb{E}[f_{t,u}] - \mathbb{E}[f_{t,0}]| \leq 2\sqrt{u}\mathbb{E}[|H_k|] + u. \tag{8.23}$$

The continuity with respect to u is uniform in K and this implies that we can compute $\mathbb{E}[f_{1,0}] = \lim_{K\to+\infty} \lim_{u\to 0} \mathbb{E}[f_{1,u}]$ by taking the limit

$$\lim_{u\to 0} \lim_{K\to+\infty} \mathbb{E}[f_{1,u}].$$

By the fundamental theorem of calculus,

$$\mathbb{E}[f_{1,u}] = \mathbb{E}[f_{0,u}] + \int_0^1 dt \frac{d}{dt}\mathbb{E}[f_{t,u}]. \tag{8.24}$$

Our task is now reduced to estimating

$$\lim_{u\to 0} \lim_{K\to+\infty} \int_0^1 dt \frac{d}{dt}\mathbb{E}[f_{t,u}].$$

After lengthy calculations and using the concentration result of Theorem 8.5, we arrive at the expression

$$\frac{d}{dt}\mathbb{E}[f_{t,u}] = \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(1 + \beta(1-m)B(t))^2} - \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(1 + \beta B(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u})} + o_N(1). \tag{8.25}$$

We add and subtract the term $\frac{1}{2\beta}\ln(1 + \beta B(1-m))$ from (8.24) and use the integral representation

$$\frac{1}{2\beta}\ln(1 + \beta B(1-m)) = \frac{1}{2\beta}\int_0^1 dt \frac{\beta B'(t)(1-m)}{1 + \beta B(t)(1-m)}$$

[4] it is also straightforward to compute the full u dependence and see that it is $O(\sqrt{u})$, uniformly in K

to obtain

$$\mathbb{E}[f_{1,u}] = \mathbb{E}[f_{0,u}] - \frac{1}{2\beta}\ln(1+\beta B(1-m)) + \int_0^1 dt\left(\frac{d}{dt}\mathbb{E}[f_{t,u}] + \frac{B'(t)(1-m)}{2(1+\beta B(t)(1-m))}\right).$$

We now substitute the expression (8.25) for the derivative. After careful manipulations, a remarkable algebra occurs and the integrand becomes

$$R(t) + \frac{B'(t)(1-m)}{2(1+\beta B(t)(1-m))^2},$$

with

$$R(t) = \frac{\beta B'(t)B(t)(\mathbb{E}\langle m_1 - m\rangle_{t,u})^2}{2(1+\beta B(t)(1-m))^2(1+\beta B(t)\mathbb{E}\langle 1-m_1\rangle_{t,u})} \geq 0.$$

So the integral has a positive contribution $\int_0^1 dtR(t) \geq 0$ plus a computable contribution equal to $\frac{B(1-m)}{2(1+\beta B(1-m))} = \frac{\lambda}{2}(1-m)$. Finally, thanks to (8.22), we have

$$\frac{1}{2} + \mathbb{E}[f_{1,u}] = \int_{-\infty}^{+\infty} Dz\ln(2\cosh(\sqrt{\lambda}z+\lambda)) - \frac{1}{2\beta} - \frac{1}{2\beta}\ln(1+\beta B(1-m)) - \frac{\lambda}{2}(1+m) + \int_0^1 R(t)dt + o_N(1) + O(\sqrt{u}) \tag{8.26}$$

where for a.e $u > \epsilon$, $\lim_{N\to\infty} o_N(1) = 0$. We take first the limit $N \to \infty$, then $u \to \epsilon$ (along some appropriate sequence) and then $\epsilon \to 0$ to obtain a formula for the free energy where the only non-explicit contribution is $\int_0^1 dtR(t)$. Since this is positive for all m, we obtain a lower bound on the free energy which is equivalent to the announced upper bound on the capacity.

8.8 Existence of the Limit

Using the interpolation method we can also show that the large system limit of C_K exists.

Theorem 8.6 (Existence of the Limit)**.** *Consider a CDMA with binary inputs and let the spreading sequence distribution belong to Class A. Let C_K denote its capacity. Then*

$$\lim_{K\to\infty} C_K \qquad \textit{exists.} \tag{8.27}$$

Note that it is sufficient to prove the above theorem for Gaussian spreading sequences. The general case then follows from Theorem 8.3.

The relation between the free energy and the capacity in (8.8) implies that it is sufficient to show the existence of limit for the average free energy $\mathcal{F}_K = \mathbb{E}[f(\bar{Y}, \mathbf{S})]$. The idea is to use Fekete's lemma which states that if a sequence $\{a_n\}$ is super additive, i.e., $a_{m+n} \geq a_m + a_n$ then $\lim_{n\to\infty} \frac{a_n}{n}$ exists. The relevant sequence for us is $\{K\mathcal{F}_K\}$. The aim is therefore to show that $K\mathcal{F}_K \geq K_1\mathcal{F}_{K_1} + K_2\mathcal{F}_{K_2}$ for $K = K_1 + K_2$.

As in the previous sections, working directly with the CDMA system is difficult and hence we perturb the Hamiltonian with $h_u(\bar{x})$ as defined in (8.19). Let us express the Hamiltonian as

$$\mathtt{H}_u(\bar{x}) = -\frac{1}{2\sigma^2}\|\bar{n} + \frac{1}{\sqrt{N}}\mathbf{s}(\bar{x}^0 - \bar{x})\|^2 + h_u(\bar{x}). \tag{8.28}$$

The expression follows by replacing $\bar{y} = \bar{n} + \mathbf{s}\bar{x}^0$, where $\bar{x}^0$ is the transmitted vector. Let us define the corresponding partition function as Z_u and the free energy as $\mathcal{F}_K(u) = \frac{1}{K}\mathbb{E}[\ln Z_u]$. The original free energy is obtained by substituting $u = 0$, i.e., $\mathcal{F}_K = \mathcal{F}_K(0)$. From the uniform continuity of $\mathcal{F}_K(u)$, it is sufficient to show the convergence of $\mathcal{F}_K(u)$ for some u close to zero. Even this turns out to be difficult and what we can show is the existence of the limit $\int_{u=\epsilon}^{a} \mathcal{F}_K(u)du$ for any $a > \epsilon > 0$. However this is sufficient for us due to the following: from the continuity of the free energy with u (8.23) we have

$$\int_{\epsilon}^{2\epsilon} (\mathcal{F}_K(u) - |O(1)|\sqrt{u})du \leq \epsilon\mathcal{F}_K \leq \int_{\epsilon}^{2\epsilon} (\mathcal{F}_K(u) + |O(1)|\sqrt{u})du.$$

Since the limit of the integral exists, we have

$$|\limsup_{K\to\infty} \mathcal{F}_K - \liminf_{K\to\infty} \mathcal{F}_K| \leq |O(1)|\sqrt{\epsilon}.$$

This ϵ can be made as small as desired and hence the theorem follows.

Let $K = K_1 + K_2$ and let $N_1 = \lfloor \frac{K_1}{\beta} \rfloor, N_2 = \lfloor \frac{K_2}{\beta} \rfloor$. The last assumption can be removed by considering their integer parts. For simplicity, let us assume that $\frac{K_1}{\beta}, \frac{K_2}{\beta} \in \mathbb{N}$. We split the $N \times K$ dimensional spreading matrix $\mathbf{s}$ in to two parts of dimension $N_1 \times K$ and $N_2 \times K$ and denote these matrices by $\mathbf{s}_1, \mathbf{s}_2$ respectively. Let $\mathbf{t}_1, \mathbf{t}_2$ be two spreading matrices with dimensions $N_1 \times K_1$ and $N_2 \times K_2$. All the entries of these matrices are distributed as $\mathcal{N}(0, 1)$ and the noise is Gaussian with variance σ^2. Similarly split the noise vector $\bar{n} = (\bar{n}_1, \bar{n}_2)$ where $\bar{n}_i$ is of length N_i and $\bar{x} = (\bar{x}_1, \bar{x}_2)$ where $\bar{x}_i$ is of length K_i. Similarly split the transmitted vector $\bar{x}^0 = (\bar{x}_1^0, \bar{x}_2^0)$ Let us consider the following Hamiltonian:

$$\begin{aligned}\mathtt{H}_{t,u}(\bar{x}) = &-\frac{1}{2\sigma^2}\|\bar{n}_1 + \frac{\sqrt{t}}{\sqrt{N}}\mathbf{s}_1(\bar{x}_1^0 - \bar{x}) + \frac{\sqrt{1-t}}{\sqrt{N_1}}\mathbf{t}_1(\bar{x}_1^0 - \bar{x}_1)\|^2 \\ &-\frac{1}{2\sigma^2}\|\bar{n}_2 + \frac{\sqrt{t}}{\sqrt{N}}\mathbf{s}_2(\bar{x}_2^0 - \bar{x}) + \frac{\sqrt{1-t}}{\sqrt{N_2}}\mathbf{t}_2(\bar{x}_2^0 - \bar{x}_2)\|^2 + h_u(\bar{x}).\end{aligned}$$

For a moment neglect the $h_u(\bar{x})$ part of the Hamiltonian and consider the remaining part. At $t = 1$, we get the Hamiltonian corresponding to an $N \times K$ CDMA system with spreading matrix $[\begin{smallmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{smallmatrix}]$. At $t = 0$ we get the Hamiltonian corresponding to two independent CDMA systems with spreading matrices $\mathbf{t}_i$ of dimensions $N_i \times K_i$. As before we perturb the Hamiltonian with $h_u(\bar{x})$ so that we can use the concentration results for the magnetization.

Let $Z_{t,u}$ be the partition function with this Hamiltonian and the corresponding average free energy is given by $g_{t,u} = \frac{1}{K}\mathbb{E}[\ln Z_{t,u}]$. Note that $g_{1,u} = \mathcal{F}_K(u)$ and $g_{0,u} = \frac{K_1}{K}\mathcal{F}_{K_1}(u) + \frac{K_2}{K}\mathcal{F}_{K_2}(u)$. From the fundamental theorem of calculus,

$$g_{1,u} = g_{0,u} + \int_0^1 \frac{d}{dt} g_{t,u} dt. \tag{8.29}$$

After some calculus we can show that for $a > \epsilon > 0$,

$$\int_\epsilon^a \int_0^1 \frac{d}{dt} g_{t,u} dt du + o_K(1) \leq 0 \tag{8.30}$$

Therefore,

$$\int_\epsilon^a g_{1,u} du + o_K(1) \leq \int_\epsilon^a g_{0,u} du$$

which implies

$$\int_\epsilon^a \mathcal{F}_K(u) du + o_K(1) \leq \frac{K_1}{K} \int_\epsilon^a \mathcal{F}_{K_1}(u) du + \frac{K_2}{K} \int_\epsilon^a \mathcal{F}_{K_2}(u) du.$$

This in turn implies that $\lim_{K\to\infty} \int_\epsilon^a \mathcal{F}_K(u) du$ exists.

8.9 Extensions

The interpolation method that we have developed can be extended to many other cases in a straightforward manner to obtain upper bounds on the respective capacities. These include the case of users with unequal powers, colored noise as well as users with Gaussian input distribution. Other cases which we do not discuss here, but the methods are valid include non-binary constellations and complex channels. Our methods combined with the interpolation of Montanari [111] would result in bounds for CDMA with LDPC coded communication.

8.9.1 Unequal Powers

Let the power of the k-th user be P_k, i.e.,

$$y_i = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} s_{ik} \sqrt{P_k} x_k + \sigma n_i,$$

with normalized average power $\frac{1}{K}\sum P_k = 1$. We assume that the empirical distribution of the P_k tends to a distribution and denote the corresponding expectation by $\mathbb{E}_P[-]$.

The interpolation method can be applied as before. We interpolate between the original system and a decoupled one where

$$y'_k = \sqrt{P}_k x_k + \frac{1}{\sqrt{\lambda}} w_k.$$

The SNRs for the t-CDMA, namely $\lambda(t)$ and $B(t)$ are related as in (8.15). The perturbation term in the Hamiltonian is given by

$$h_u(\bar{x}) = \sqrt{u}\sum_{k=1}^{K} h_k\sqrt{P_k}x_k + u\sum_{k=1}^{K} P_k x_k^0 x_k - \sqrt{u}\sum_{k=1}^{K}|h_k|\sqrt{P_k}.$$

The whole analysis can again be performed in exactly the same manner with the proviso that the "magnetization" is defined as $m_1 = \frac{1}{N}\sum P_k x_k^0 x_k$. Then we can deduce the upper bound (8.13) on the capacity with

$$c_{RS}(m) = -\mathbb{E}_P\left[\int_{-\infty}^{+\infty} Dz \ln(\cosh(\sqrt{P\lambda}z + P\lambda)\right] + \frac{\lambda}{2}(1+m) - \frac{1}{2\beta}\ln\lambda\sigma^2.$$

8.9.2 Colored Noise

Now consider the scenario where

$$y_i = \frac{1}{\sqrt{N}}\sum_{k=1}^{K} s_{ik}x_k + n_i,$$

with colored noise of finite memory. More precisely we assume that the the covariance matrix $\mathbb{E}[N_i N_j] = C(i,j)$ (depends on $|i-j|$) is circulant as $N \to +\infty$ and has well defined (real) Fourier transform (the noise spectrum) $\hat{C}(\omega)$. The covariance matrix is real symmetric and thus can be diagonalized by an orthogonal matrix: $\Gamma = OCO^T$ with $OO^T = O^TO = I$. As $N \to +\infty$ the eigenvalues are well approximated by $\gamma_n \equiv \hat{C}(2\pi\frac{n}{N})$. Multiplying the received signal by $\Gamma^{-1/2}O$ the input-output relation becomes

$$y'_i = \frac{1}{\sqrt{N}}\sum_{k=1}^{K} t_{ik}x_k + n'_i,$$

where

$$y'_i = (\Gamma^{-1/2}O\bar{y})_i, \qquad t_{ik} = (\Gamma^{-1/2}O\mathbf{s})_{ik}, \qquad n'_i = (\Gamma^{-1/2}O\bar{n})_i.$$

The new noise vector $\bar{N}'$ is white with unit variance, but the spreading matrix is now correlated as

$$\mathbb{E}[T_{ik}T_{jl}] = \delta_{ij}\delta_{kl}\gamma_i^{-1}.$$

One may guess that this time the interpolation is done between the true system and the decoupled channels

$$y'_k = x_k + \frac{1}{\sqrt{\lambda_{col}}} w_k,$$

where this time

$$\lambda_{col} = \int_0^{2\pi} \frac{d\omega}{2\pi} \frac{B}{\hat{C}(\omega) + \beta B(1-m)}.$$

Note that $\hat{C}(\omega) = 1$ when the noise is white and we get back the λ defined in (8.11). The interpolating system has the same posterior as in (8.20) but with $\lambda_{col}(t)$ and $B(t)$ related by

$$\int_0^{2\pi} \frac{d\omega}{2\pi} \frac{B(t)}{\hat{C}(\omega) + \beta B(t)(1-m)} + \lambda_{col}(t) = \int_0^{2\pi} \frac{d\omega}{2\pi} \frac{B}{\hat{C}(\omega) + \beta B(1-m)}.$$

Continuing with the interpolation method for this system, in place of (8.25) we get

$$\frac{1}{N}\sum_{n=1}^{N} \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(\gamma_n + \beta(1-m)B(t))^2} - \frac{1}{N}\sum_{n=1}^{N} \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(\gamma_n + \beta B(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u})} + o_N(1)$$

$$\int_0^{2\pi} \frac{d\omega}{2\pi} \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(S(\omega) + \beta B(t)(1-m))^2} - \int_0^{2\pi} \frac{d\omega}{2\pi} \frac{B'(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u}}{2(S(\omega) + \beta B(t)\mathbb{E}\langle 1 - m_1\rangle_{t,u})}.$$

This finally leads to the bound on capacity with,

$$c_{RS}(m) = -\int_{-\infty}^{+\infty} Dz \ln(\cosh(\sqrt{\lambda_{col}} z + \lambda_{col})) + \frac{\lambda_{col}}{2}(1+m) + \frac{1}{2\beta}\int_0^{2\pi} \frac{d\omega}{2\pi} \ln \frac{\hat{C}(\omega)}{\hat{C}(\omega) + \beta(1-m)}.$$

8.9.3 Gaussian Input

In the case of continuous inputs ($x_k \in \mathbb{R}$), the $\sum_{\bar{x}}$ in (8.5) is replaced by $\int d\bar{x}$. The capacity is maximized by a Gaussian prior,

$$p_{\bar{X}}(\bar{x}) = \frac{e^{-\frac{||\bar{x}||^2}{2}}}{(2\pi)^{N/2}} \tag{8.31}$$

and one can express it in terms of a determinant involving the correlation matrix of the spreading sequences. Using the exact spectral measure given by random matrix theory Shamai and Verdú [95] obtained the rigorous result

$$\lim_{K\to\infty} C_K = \frac{1}{2}\log(1 + \sigma^{-2} - \frac{1}{4}Q(\sigma^{-2}, \beta))$$

$$+\frac{1}{2\beta}\log(1+\sigma^{-2}\beta-\frac{1}{4}Q(\sigma^{-2},\beta))-\frac{Q(\sigma^{-2},\beta)}{8\beta\sigma^{-2}} \tag{8.32}$$

where

$$Q(x,z)=\left(\sqrt{x(1+\sqrt{z})^2+1}-\sqrt{x(1-\sqrt{z})^2+1}\right)^2$$

On the other hand Tanaka applied the formal replica method to this case and obtained (8.9) with

$$c_{RS}(m)=\frac{1}{2}\log(1+\lambda)-\frac{1}{2\beta}\log\lambda\sigma^2-\frac{\lambda}{2}(1-m) \tag{8.33}$$

where $\lambda=(\sigma^2+\beta(1-m))^{-1}$. The maximizer satisfies

$$m=\frac{\lambda}{1+\lambda} \tag{8.34}$$

Solving (8.34) we obtain $m=\frac{\sigma^2}{4\beta}Q(\sigma^{-2},\beta)$ and substituting this in (8.33) gives the equality between (8.32) and (8.33). So at least for the case of Gaussian inputs we are already assured that the replica method finds the correct solution.

The interpolation is done as explained in section 8.7 except that (8.17) is multiplied by the Gaussian distribution (8.31). In (8.18) we also have to include this Gaussian factor and the sum over $\bar{x}_0$ is replaced by an integral. The main difference is that now the expectation $\mathbb{E}$ is also with respect to the Gaussian vector $\bar{x}_0$. The interpolation method gives an upper bound on the capacity

$$\lim_{K\to\infty} C_K \leq \min_{m\in[0,1]} c_{RS}(m),$$

where $c_{RS}(m)$ is the same as in (8.33). This provides an evidence that the bounds obtained by our method are tight.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, July and October 1948.

[2] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec., pt. 4*, vol. 27, pp. 142–163, 1959.

[3] P. Elias, "Coding for noisy channels," in *IRE International Convention Record*, Mar. 1955, pp. 37–46.

[4] R. L. Dobrushin, "Asymptotic optimality of group and systematic codes for certain channels," *Teor. Veroyat. i primenen.*, vol. 8, pp. 52–66, 1963.

[5] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[6] T. J. Goblick, Jr., "Coding for discrete information source with a distortion measure," Ph.D. dissertation, MIT, 1962.

[7] T. Berger, *Rate Distortion Theory.* London: Prentice Hall, 1971.

[8] D. J. Costello Jr. and G. D. Forney Jr., "Channel coding: The road to channel capacity," *Proceedings of the IEEE*, vol. 95, no. 6, June 2007.

[9] R. C. Bose and D. K. Ray-Chaudhuri, "On a class of error correcting binary group codes," *Info. and Control*, vol. 3, no. 1, pp. 68–79, Mar. 1960.

[10] A. Hocquenghem, "Codes correcteurs d'erreurs," *Chiffres*, vol. 2, pp. 147–156, 1959.

[11] I. S. Reed, "A class of multipl-error-correcting codes and the decoding scheme," *IRE Transactions on Inform. Theory*, vol. 4, pp. 38–49, 1954.

[12] D. E. Muller, "Application of Boolean algebra to switching circuit design," *IRE Transactions on Electron. Comput.*, vol. 3, pp. 6–12, 1954.

[13] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. SIAM*, vol. 8, no. 2, pp. 300–304, June 1960.

[14] P. Elias, "Error-free coding," *IEEE Trans. Inform. Theory*, vol. 4, pp. 29–37, Sept. 1954.

[15] G. D. Forney, Jr., *Concatenated Codes*. MIT Press, 1966.

[16] A. J. Viterbi, "Error bounds of convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[17] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[18] R. M. Fano, "A heuristic discussion of probabilistic decoding," *IEEE Trans. Inform. Theory*, vol. 9, no. 2, pp. 64–74, Apr. 1963.

[19] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA, USA: M.I.T. Press, 1963.

[20] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding," in *Proc. of ICC*, Geneve, Switzerland, May 1993, pp. 1064–1070.

[21] D. J. C. MacKay and R. M. Neal, "Good codes based on very sparse matrices," in *Cryptography and Coding. 5th IMA Conference*, ser. Lecture Notes in Computer Science, C. Boyd, Ed. Berlin: Springer, 1995, no. 1025, pp. 100–111.

[22] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 1710–1722, Nov. 1996.

[23] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *European Transactions on Telecommuncations*, vol. 6, pp. 513–526, Sept. 1995.

[24] N. Wiberg, "Codes and decoding on general graphs," Ph.D. dissertation, Linköping University, S-581 83, Linköping, Sweden, 1996.

[25] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. A. Spielman, and V. Stemann, "Practical loss-resilient codes," in *Proc. of the 29th annual ACM Symposium on Theory of Computing*, 1997, pp. 150–159.

[26] M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. A. Spielman, "Analysis of low density codes and improved designs using irregular graphs," in *Proc. of the 30th Annual ACM Symposium on Theory of Computing*, 1998, pp. 249–258.

[27] ——, "Efficient erasure correcting codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 569–584, Feb. 2001.

[28] ——, "Improved low-density parity-check codes using irregular graphs," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 585–598, Feb. 2001.

[29] M. G. Luby, M. Mitzenmacher, and M. A. Shokrollahi, "Analysis of random processes via and-or tree evaluation," in *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, San Francisco, California, United States, 1998, pp. 364–373.

[30] T. Richardson and R. Urbanke, "The capacity of low-density parity check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[31] S.-Y. Chung, G. D. Forney, Jr., T. Richardson, and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit," *IEEE Communications Letters*, vol. 5, no. 2, pp. 58–60, Feb. 2001.

[32] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *accepted for publication in IEEE Trans. Inform. Theory*, 2009.

[33] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, 1952.

[34] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[35] ——, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[36] A. J. Viterbi and J. K. Omura, "Trellis encoding of memoryless disctre-time sources with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 325–332, 1974.

[37] Y. Matsunaga and H. Yamamoto, "A coding theorem for lossy data compression by LDPC codes," *IEEE Trans. Inform. Theory*, vol. 49, no. 9, pp. 2225–2229, 2003.

[38] M. J. Wainwright and E. Martinian, "Low-density graph codes that are optimal for source/channel coding and binning," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1061–1079, 2009.

[39] E. Martinian and J. Yedidia, "Iterative quantization using codes on graphs," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, 2003.

[40] T. Murayama, "Thouless-Anderson-Palmer approach for lossy compression," *J. Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.*, vol. 69, 2004.

[41] S. Ciliberti and M. Mézard, "The theoretical capacity of the parity source coder," *Journal of Statistical Mechanics:Theory and Experiment*, vol. 1, no. 10003, 2005.

[42] S. Ciliberti, M. Mézard, and R. Zecchina, "Lossy data compression with random gates," *Physical Rev. Lett.*, vol. 95, no. 038701, 2005.

[43] A. Braunstein, M. Mézard, and R. Zecchina, "Survey propagation: Algorithm for satisfiability," *Random Structures and Algorithms*, vol. 28, pp. 340–373, 2006.

[44] M. J. Wainwright and E. Maneva, "Lossy source coding via message-passing and decimation over generalized codewords of LDGM codes," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Adelaide, Australia, Sept. 2005, pp. 1493–1497.

[45] T. Filler and J. Fridrich, "Binary quantization using belief propagation with decimation over factor graphs of LDGM codes," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, 2007.

[46] E. Arıkan, "Channel combining and splitting for cutoff rate improvement," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 628–639, 2006.

[47] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, July 2009, pp. 1493–1495.

[48] S. B. Korada, E. Şaşoğlu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," *accpeted for publication in IEEE Trans. Inform. Theory*, 2009.

[49] S. Verdú, "Capacity region of gaussian CDMA channels: The symbol synchronous case," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, Oct. 1986.

[50] I. Land and J. B. Huber, *Information Combining*, ser. Foundations and Trends in Communications and Information Theory. Delft, the Netherlands: NOW, Nov. 2006, vol. 3, available online at http://www.ee.technion.ac.il/people/sason/monograph.html.

[51] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[52] R. Mori and T. Tanaka, "Performance and Construction of Polar Codes on Symmetric Binary-Input Memoryless Channels," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, July 2009, pp. 1496–1500.

[53] S. B. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *submitted to IEEE Trans. Inform. Theory*, 2009.

[54] A. Montanari, F. Ricci-Tersenghi, and G. Semerjian, "Solving constraint satisfaction problems through belief propagation-guided decimation," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, USA, Sep 26–Sep 28 2007.

[55] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[56] S. I. Gelfand and M. S. Pinsker, "Coding for channel with random parameters," *Problemy Peredachi Informatsii*, vol. 9(1), pp. 19–31, 1983.

[57] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1250–1216, 2002.

[58] J. Chou, S. S. Pradhan, and K. Ramachandran, "Turbo and trellis-based constructions for source coding with side information," in *Data Compression Conference*, Mar. 2003.

[59] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[60] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Nested convolutional/turbo codes for the binary Wyner-Ziv problem," in *Proceedings of the International Conference on Image Processing*, Sept. 2003, pp. 601–604.

[61] Y. Yang, V. Stankovic, Z. Xiong, and W. Zhao, "On multiterminal source code design," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2278–2302, 2008.

[62] J. Chou, S. S. Pradhan, and K. Ramachandran, "Turbo coded trellis-based constructions for data embedding: Channel coding with side information," in *Proceedings of the Asilomar Conference*, Nov. 2001, pp. 305–309.

[63] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3417–3432, 2005.

[64] Y. Sun, A. D. Liveris, V. Stankovic, and Z. Xiong, "Near-capacity dirty-paper code designs based on TCQ and IRA codes," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Sept. 2005, pp. 184–188.

[65] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[66] A. D. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Trans. Inform. Theory*, vol. 19, no. 6, pp. 772–777, Nov. 1973.

[67] D. Aldous and J. A. Fill, *Reversible Markov chains and random walks on graphs*. Available at www.stat.berkeley.edu/users/aldous/book.html.

[68] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1159–1180, 2003.

[69] G. Caire, S. Shamai, and S. Verdú, "Lossless data compression with error correcting code," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Yokohama, Japan, June 29–July 4 2003, conference, p. 22.

[70] E. Arıkan and E. Telatar, "Personal Communication," 2009.

[71] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.

[72] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[73] J.A. Bondy and U.S.R. Murty, *Graph Theory*. Springer, 2008.

[74] R. E. Blahut, *Theory and Practice of Error Control Codes*. Addison-Wesley, 1983.

[75] F. J. MacWilliams and N. J. Sloane, *The Theory of Error-Correcting Codes*. North-Holland, 1977.

[76] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1313 – 1325, Apr. 2005.

[77] I. Land, S. Huettinger, P. A. Hoeher, and J. B. Huber, "Bounds on information combining," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 612–619, 2005.

[78] A. Ashikhmin, G. Kramer, and S. ten Brink, "Extrinsic information transfer functions: model and erasure channel property," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2657–2673, Nov. 2004.

[79] N. Hussami, S. B. Korada, and R. Urbanke, "Performance of polar codes for channel and source coding," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, July 2009, pp. 1488–1492.

[80] G. D. Forney, Jr., "Codes on graphs: Normal realizations," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001.

[81] I. Dumer, "Recursive decoding and its performance for low-rate Reed-Muller codes," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 811–823, 2004.

[82] ——, "Soft-decision decoding of Reed-Muller codes: a simplified algorithm," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 954–963, 2006.

[83] E. Arıkan, "A performance comparision of Polar codes and Reed-Muller codes," *IEEE Communications Letters*, vol. 12, no. 6, 2008.

[84] M. Schwartz and A. Vardy, "On the stopping distance and the stopping redundancy of codes," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 922 – 932, Mar. 2006.

[85] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *The Annals of Mathematical Statistics*, vol. 3, no. 4, pp. 1229–1241, 1959.

[86] S. H. Hassani, S. B. Korada, and R. Urbanke, "The compound capacity of polar codes," *in preparation*, 2009.

[87] S. B. Korada and R. Urbanke, "Exchange of limits: Why iterative decoding works," *accepted for publication in IEEE Trans. Inform. Theory*, 2008.

[88] R. G. Gallager, "Low-density parity-check codes," *IRE Transactions on Inform. Theory*, vol. 8, pp. 21–28, Jan. 1962.

[89] R. J. McEliece, E. Rodemich, and J.-F. Cheng, "The turbo decision algorithm," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, 1995.

[90] D. Burshtein and G. Miller, "Expander graph arguments for message-passing algorithms," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 782–790, Feb. 2001.

[91] A. Amraoui, A. Montanari, T. Richardson, and R. Urbanke, "Finite-length scaling for iteratively decoded LDPC ensembles," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, Oct. 2003.

[92] S. B. Korada and N. Macris, "Tight bounds on the capacity of binary input random CDMA systems," *accepted for publication in IEEE Trans. Inform. Theory*, 2009.

[93] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.

[94] A. J. Grant and P. D. Alexander, "Randomly selected spreading sequences for coded CDMA," in *4th Int. Spread Spectrum Techniques and Applications*, Mainz, Germany, Sept. 1996, pp. 54–57.

[95] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 622–640, 1999.

[96] D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 641–657, 1999.

[97] D. N. C. Tse and S. Verdú, "Optimum asymptotic multiuser efficiency of randomly spread cdma," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2718–2722, 2000.

[98] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[99] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[100] M. Talagrand, "The generalized Parisi formula," *Comptes Rendus Mathematique*, vol. 337, pp. 111–114, 2003.

[101] A. Montanari and D. Tse, "Analysis of belief propagation for non-linear problems: The example of CDMA (or : How to prove Tanaka's formula)," in *Proc. of the IEEE Inform. Theory Workshop*, Punta del Este, Uruguay, Mar 13–Mar 17 2006.

[102] A. Montanari, "The glassy phase of Gallager codes," *Eur. Phys. J. B*, vol. 23, pp. 121–136, 2001.

[103] Y. Kabashima and T. Hosaka, "Statistical mechanics of source coding with a fidelity criterion," *Progress of theoretical physics. Supplement*, no. 157, pp. 197–204, 2005.

[104] K. Nakamura, Y. Kabashima, R. Morelos-Zaragoza, and D.Saad, "Statistical mechanics of broadcast channels using low-density parity-check codes," *Phys. Rev. E*, vol. 67, no. 036703 (1-9), 2003.

[105] F. Guerra, "Sum rules for the free energy in the mean field spin glass model," *Fields Institute Communications*, vol. 30, p. 161, 2001.

[106] F. Guerra and F. L. Toninelli, "Quadratic replica coupling in the Sherrington-Kirkpatrick mean field spin glass model," *J. Math. Phys.*, vol. 43, pp. 3704–3716, 2002.

[107] ——, "The infinite volume limit in generalized mean field disordered models," *Markov Proc. Rel. Fields.*, vol. 49, no. 2, pp. 195–207, 2003.

[108] G. Parisi, "A sequence of approximate solutions to the S-K model for spin glasses," *J. Phys*, vol. A13, pp. L–115, 1980.

[109] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond*. World Scientific Publishing Co. Pte. Ltd., 1987.

[110] M. Talagrand, "The Parisi formula," *Annals of Mathematics*, vol. 163, pp. 221–263, 2006.

[111] A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3221–3246, Sept. 2005.

[112] S. Kudekar and N. Macris, "Sharp bounds for MAP decoding of general irregular LDPC codes," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seattle, USA, Sept. 2006.

[113] S. B. Korada, S. Kudekar, and N. Macris, "Exact solution for the conditional entropy of poissonian LDPC codes over the binary erasure channel," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Nice, France, July 2007.

[114] L. A. Pastur and M. Shcherbina, "Absence of self-averaging of the order parameter in the Sherrington-Kirkpatrick model," *Journal of Statistical Physics*, vol. 62, no. 1-2, pp. 1–19, Jan. 1991.

[115] M. Shcherbina and B. Tirozzi, "The free energy of a class of Hopfield models," *Journal of Statistical Physics*, vol. 72, pp. 113–125, 1993.

Curriculum Vitae

Education:

- **Swiss Federal Institute of Technology at Lausanne (EPFL)**
  Ph.D. in Communication Systems
  Thesis Title: Polar Codes for Channel and Source Coding
  September 2005 - June 2009.
  Pre-doctoral school
  October 2004 - July 2005.

- **Indian Institute of Technology, Delhi**
  Bachelor of Technology in Electrical Engineering
  July 2000 - May 2004.

Journal Publications:

[J1] Satish Babu Korada, Rüdiger Urbanke, "Exchange of Limits: Why Iterative Decoding Works", accepted for publication in IEEE Transactions on Information Theory.

[J2] Satish Babu Korada, Nicolas Macris, "Tight Bounds on the Capacity of Binary Input Random CDMA Systems", accepted for publication in IEEE Transactions on Information Theory.

[J3] Satish Babu Korada, Rüdiger Urbanke, "Polar Codes are Optimal for Lossy Source Coding", submitted to IEEE Transactions on Information Theory.

[J4] Satish Babu Korada, Eren Şaşoğlu, Rüdiger Urbanke, "Polar Codes: Characterization of Exponent, Bounds, and Constructions", accepted for publication in IEEE Transactions on Information Theory.

[J5] Satish Babu Korada, Nicolas Macris, "Exact Solution of the Gauge Symmetric p-Spin Glass Model on a Complete Graph", accepted for publication in Journal of Statistical Mechanics.

[J6] Dinkar Vasudevan, Satish Babu Korada, "Polymatroidal Flows on Two Classes of Information Networks", submitted to IEEE Transactions on Information Theory.

[J7] Darryl Veitch, Julien Ridoux, Satish Babu Korada, "Robust Synchronization of Absolute and Difference Clocks over Networks", in IEEE/-ACM Transactions on Networking, April, 2009.

Conference Publications:

[C1] Satish Babu Korada, Eren Sasoglu, "A Class of Transformations that Polarize binary-input memoryless channels", Proc. IEEE International Symposium on Information Theory, Seoul, July 2009.

[C2] Satish Babu Korada, Eren Sasoglu, Rudiger Urbanke "Polar Codes: Characterization of Exponent, Bounds and Constructions", Proc. IEEE International Symposium on Information Theory, Seoul, July 2009 (**Best Student Paper**).

[C3] Nadine Hussami, Satish Babu Korada, Rudiger Urbanke, "Performance of Polar Codes for Channel and Source Coding", Proc. IEEE International Symposium on Information Theory, Seoul, July 2009.

[C4] Dinkar Vasudevan, Satish Babu Korada, "On Broadcast with a Common Message over Networks", Proc. IEEE International Symposium on Information Theory and Applications, Auckland, December 2008.

[C5] Satish Babu Korada, Rüdiger Urbanke, "Exchange of Limits: Why Iterative Decoding Works", Proc. IEEE International Symposium on Information Theory, Toronto, July 2008 (**Best Student Paper**).

[C6] Satish Babu Korada, Shrinivas Kudekar, Nicolas Macris, "Concentration of Magnetization for Linear Block Codes", Proc. IEEE International Symposium on Information Theory, Toronto, July 2008.

[C7] Satish Babu Korada, Dinkar Vasudevan, "Broadcast and Slepian Wolf Multicast for Aref Networks", Proc. IEEE International Symposium on Information Theory, Toronto, July 2008.

[C8] Satish Babu Korada, Nicolas Macris, "On the Capacity of a Code Division Multiple Access System", Proc. Allerton Conference on Communication, Control and Computing, Monticello, Sept. 2007.

[C9] Satish Babu Korada, Nicolas Macris, "On the Concentration of the Capacity for a Code Division Multiple Access System", Proc. IEEE International Symposium on Information Theory, Nice, June 2007.

[C10] Satish Babu Korada, Shrinivas Kudekar, Nicolas Macris, "Exact Solution for the Conditional Entropy of Poissonian LDPC Codes over the Binary Erasure Channel", Proc. IEEE International Symposium on Information Theory, Nice, June 2007.

[C11] Satish Babu Korada, Nicolas Macris, "Exact Solution of a p-Spin Model and its Relationship to Error Correcting Codes", Proc. IEEE International Symposium on Information Theory, Seattle, June 2006.

[C12] Darryl Veitch, Satish Babu, Attila Pasztor, "Robust Synchronization of Software Clocks Across the Internet", Proc. ACM SIGCOMM Internet Measurement Conference, 2004.